geographical analysis

Geographical Analysis (2024) **56**, 163–184

A Framework for Moving Beyond Computational Reproducibility: Lessons from Three Reproductions of Geographical Analyses of COVID-19

Peter Kedron[1,2], Sarah Bardin[1], Joseph Holler[3], Joshua Gilman[4], Bryant Grady[1], Megan Seeley[1], Xin Wang[5], and Wenxin Yang[1]

[1]School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, Arizona, USA, [2]Department of Geography, University of California Santa Barbara, Santa Barbara, California, USA, [3]Department of Geography, Middlebury College, Middlebury, Vermont, USA, [4]School of Life Sciences, Arizona State University, Tempe, Arizona, USA, [5]School of Sustainability, Arizona State University, Tempe, Arizona, USA

Despite recent calls to make geographical analyses more reproducible, formal attempts to reproduce or replicate published work remain largely absent from the geographic literature. The reproductions of geographic research that do exist typically focus on computational reproducibility – whether results can be recreated using data and code provided by the authors – rather than on evaluating the conclusion and internal validity and evidential value of the original analysis. However, knowing if a study is computationally reproducible is insufficient if the goal of a reproduction is to identify and correct errors in our knowledge. We argue that reproductions of geographic work should focus on assessing whether the findings and claims made in existing empirical studies are well supported by the evidence presented. We aim to facilitate this transition by introducing a model framework for conducting reproduction studies, demonstrating its use, and reporting the findings of three exemplar studies. We present three model reproductions of geographical analyses of COVID-19 based on a common, open access template. Each reproduction attempt is published as an open access repository, complete with pre-analysis plan, data, code, and final report. We find each study to be partially reproducible, but moving past computational reproducibility, our assessments reveal conceptual and methodological concerns that raise questions about the predictive value and the magnitude of the associations presented in each study. Collectively, these reproductions and our template materials offer a practical framework others can use to reproduce and replicate empirical spatial analyses and ultimately facilitate the identification and correction of errors in the geographic literature.

Correspondence: Peter Kedron, Department of Geography, University of California Santa Barbara, Santa Barbara, CA 93106
e-mail: peterkedron@ucsb.edu

Submitted: April 18, 2022. Revised version accepted: June 5, 2023.

doi: 10.1111/gean.12370

Introduction

The geographic literature is quickly becoming crowded with calls to make geographical research more reproducible (see Brunsdon 2016; Muenchow, Schäfer, and Krüger 2019; Yin et al. 2019; Brunsdon and Comber 2021; Goodchild et al. 2021; Kedron et al. 2021a, b). In principle, reproducible research publicly discloses the evidence base for claims from prior work not only to improve the transparency of scientific communication but also to facilitate the independent verification of those claims (Schmidt 2009; Nosek, Spies, and Motyl 2012; Earp and Trafimow 2015). Reproducibility is therefore tied to at least two questions about the results and claims of prior work (National Academies of Sciences, Engineering, and Medicine 2019). First, are the data and methods used in a prior study shared clearly enough to allow for the results to be recreated? Second, once an attempt to recreate the results of a prior study has been made, do the data, analysis, and results in fact support the claim(s) made by the study? Research that addresses either question can help make geographic research more reproducible and facilitate the verification and accumulation of geographic knowledge.

To date, geographers have largely focused their efforts on the first of these two questions by working to assess and address whether the data, code, and methods needed to reproduce research are available. Researchers have catalogued the availability of data and code in subsets of the geographic literature (Konkol, Kray, and Pfeiffer, 2019; Ostermann and Granell 2017), identified actions geographers can take to better share their data and methods (Tullis and Kar 2021; Kedron et al. 2021b), offered guidelines for how to do so (Hofer et al. 2019; Nüst and Pebesma 2021; Wilson et al. 2021), and created infrastructure to host researcher materials and recreate analyses (Wang 2016; Nüst and Hinz 2019; Yin et al. 2019). These activities set the stage for reproduction studies that assess the claims made in the existing geographic literature but do not themselves assess those claims.

Formal attempts to reproduce published studies and assess whether the claims presented in those studies are well-supported remain largely absent from the geographic science literature. The few recently published reproduction studies that exist in the field focus on assessing whether studies can be computational reproduced (i.e., whether the computational results of a prior study can be recreated using the same data and code). These studies are similar to traditional manuscript reviews, but additionally attempt to execute available code, numerically compare outputs to results reported in the manuscript, and report (and sometimes correct) errors in code compilation or execution. While these studies do attempt to reproduce prior results, they do not take the additional step of explicitly assessing whether the evidence presented does, in fact, support the claims being made. Narrowly focusing reproduction attempts on recreating the results and correcting the coding errors of prior studies reduces reproduction to a form of quality audit that provides limited information about the conclusion validity and internal validity of prior work. This approach is understandable, as the reproducibility crisis across the sciences is often linked to the ubiquitous use of expanding computing resources to perform complex analyses of complicated problems (Stodden, Leisch, and Peng 2014; Stodden et al. 2016; National Academies of Sciences, Engineering, and Medicine 2019). Unfortunately, ending the evaluation of a study at an assessment of its computational reproducibility may hinder scientific progress if the recreation of results is misconstrued as affirmation that questionable decisions leading to those results were valid.

We advocate that geographers move beyond checks of computational reproducibility and begin to develop a body of reproduction studies focused on the assessment of the claims of prior work. The objective of this article is to facilitate this transition by introducing a model

framework for conducting reproduction studies, demonstrating its use, and reporting the findings of three exemplar studies. First, we introduce a model workflow for conducting reproduction studies aimed at assessing the claims of published research. Second, to demonstrate the use of our approach and materials, we report the findings of our attempts to reproduce and assess the claims of three published geographical analyses of COVID-19 in the United States. Third, we review the reproduction process and use the information gathered during our attempts to identify how we might systematically use reproduction studies to assess and enhance future geographical research. Through these contributions, we position geographers to build on recent efforts to make reproducibility more achievable and shift their focus to the evaluation of research through rigorous recreation and reanalysis. Our work therefore reorients the field toward the second question posed by the NASEM, which has been under-discussed in the geographic literature.

The remainder of this article is organized into six sections. The following section provides background on reproduction studies in the geographical sciences. We highlight the current focus on computational reproduction and argue for a more comprehensive approach to reproduction in which the reproducing authors document, catalog, and evaluate research decisions and claims. In the third section, we present our approach to reproduction in the form of a model workflow and a set of open template materials, and we discuss how to implement our approach. In the fourth section, we introduce our three reproduction studies. We establish the need to reproduce studies of COVID-19 and outline our selection of candidate studies. We then describe how we conducted these reproductions in the fifth section. In the sixth section, we present results from each reproduction study, selected from our published reports and organized to illustrate how reproductions studies can be used to identify and address issues in the conceptualization, measurement, analysis, and communication of research. Those findings inform a concluding section that outlines how we might continue to use reproduction and replication to advance geographical analysis.

The reproduction of geographic research

Numerous geographers have made calls to strengthen geographical analysis by improving the reproducibility of geographic research and making reproduction studies part of normal disciplinary practice (Brunsdon 2016; Brunsdon and Comber 2021; Goodchild et al. 2021; Goodchild and Li 2021; Kedron et al. 2021a, b). In a reproduction study, independent researchers evaluate prior research by attempting to recreate the results of a study using the data and procedures of the original work (NASEM 2019). Researchers conducting a reproduction may focus on different goals. It is helpful to distinguish which of the two questions raised by NASEM (2019) a researcher wishes to answer: (1) whether it is possible to simply recreate the specific results of the original study or (2) whether the data, analysis, and results in fact support the conclusions and claims drawn from the original study.

When narrowly focused on simply identifying if results can be recreated, a reproduction study acts as a check of how a study was executed and shared. The NASEM (2019) categorizes this type of reproduction study as an enriched form of literature review. Simply recreating the result of a study does not establish the validity of the claims made by the researchers that conducted the original study. It merely guarantees that information about the data and methods required to assess those claims is shared with sufficient openness and detail for someone else to recreate the results. Such reproduction studies are therefore simply audits of prior research for the quality of reproducibility. In the era of sophisticated methods and reproducibility crises,

such quality audits may restore some degree of trustworthiness to research but contribute limited information about the quality of the research design or validity of the claims being made.

When a researcher attempts to reproduce a study, they either must have access to or must attempt to identify the decisions and materials that were used to create the prior result. As the reproducing researcher gathers this information and uses it to recreate the earlier work, they also have the opportunity to evaluate the claims of the original researchers in light of their decisions and to evaluate and test each decision against alternative options (Clemens 2017; Christensen, Freese, and Miguel 2019). If the reproducing researcher possesses the requisite knowledge and chooses to take these opportunities, they may gain information about how the prior study was conceptualized, designed, and executed. They can then use this information to make qualified statements about whether the conclusions are reasonable (conclusion validity) and whether those relationships may be attributable to other factors (internal validity). Statements about the conclusion or internal validity of a study must be qualified, because any assessment remains contingent upon numerous additional factors such as the design of the original study and the expertise of the reproducing researchers. While reproductions never provide conclusive evidence for or against a finding, they can provide insight into whether a study has a flawed research design or whether errors may have been made during its execution (Earp and Trafimow 2015; Nichols et al. 2021). Building from the insights of a reproduction attempt, studies can be redesigned and errors can be corrected. In this way, reproduction studies help progressively improve our understanding of phenomena by reducing the number of errors made and decreasing uncertainty.

A flurry of recent activity has begun creating environments to support reproduction studies in the the geographical sciences. Workshops and conference sessions (see Nüst et al. 2018; SPARC 2019; Kmoch, Nust, and Uuemma 2020) have formed a research community around the subject, while review articles (Brunsdon 2016; Kedron et al. 2021b) and a special forum in the *Annals of the American Association of Geographers* (Goodchild et al. 2021) have raised awareness. Several publications have also laid crucial foundations by connecting reproduction to the discipline's traditions (Wainwright 2021; Wolf et al. 2021), methodological approaches (Brunsdon and Singleton 2015; Singleton, Spielman, and Brunsdon 2016; Kedron et al. 2021a), educational priorities (Muenchow, Schäfer, and Krüger 2019; Kedron et al., 2021c), and theoretical debates (Goodchild and Li 2021; Sui and Kedron 2021; Kedron and Holler 2022a). The development of computational and institutional infrastructure (see Wang 2016; Nüst and Hinz 2019; Konkol, Nüst, and Goulier 2020; Nüst and Pebesma 2021; Wilson et al. 2021) has also reduced the barriers to conducting reproductions. Despite these developments though, few formal reproductions have been published in the geographic literature.

The reproductions that do exist typically focus on establishing whether it is possible to recreate the outcomes of a prior study by cataloging study components that can affect reproducibility or by verifying specific computational results. For example, Ostermann and Granell (2017) use a literature review of volunteered geographic information research publications to assess computational reproducibility based on availability of original data, metadata, source code, or pseudocode. Researchers taking part in an ongoing reproducible research initiative of the Association of Geographic Information Laboratories (AGILE) in Europe have reviewed the computational reproducibility of 31 research papers submitted to the annual conference for the past three years (Nüst et al. 2020, 2021, 2022) and 75 papers from the GIScience conference series (Ostermann et al. 2021). In addition to assessing the availability of data, methods (code), and results, the researchers also attempted to independently re-execute the analyses and share their findings as short reproducibility reports. Konkol, Kray, and Pfeiffer (2019) similarly attempted

computational reproductions of the coded analyses of 41 open-access research papers applying spatial statistical methods and found that most were difficult to computationally reproduce. While this research usefully summarizes technical barriers to computational reproducibility, all of these studies limit their discussion to coding errors and differences in figures and maps. Their central focus is on determining whether an independent researcher can re-execute a study's analytical code and create identical outputs, and their broader impact is to derive guidelines for publishing computationally reproducible research.

In contrast, if the primary goal of a reproduction study is to assess whether the data, analysis, and results of a study in fact support the claims made by a researcher, then it is insufficient to only determine whether the code can succeed at exactly recreating the original results and figures. In the geographical sciences, it is particularly critical for a researcher seeking to evaluate a work by attempting to reproduce it to attend to threats to validity involving geographic space (Schmitt 1978). Reproductions lend themselves to evaluations of the conclusion or internal validity of a study. A study with a flawed research design may still be computationally reproducible. Even when a study is well-designed and properly executed, reproducing the results without critically reflecting on the design and execution of the study will do little to advance knowledge. To understand whether a result is credible or reliable, a researcher conducting a reproduction study must also examine how the original researchers conceptualized, designed, and implemented their study (Kedron et al. 2021a). If research findings depend on decisions that are not justified, then the findings themselves are not justified (Christensen, Freese, and Miguel 2019).

When an independent researcher makes an argument that there is a better way to analyze the original data than was reported in a study, reproduction can be a platform for introducing procedural differences that may affect the result of the original study. By introducing those changes, it is possible to begin to determine whether the approach adopted by the original researchers was somehow inadequate or erroneous. Davies (1968) provides an early example of this approach to reproduction in geography. By examining the predictions of central place theory, Davies reanalyzes the data of two studies using slightly different techniques to draw conclusions about the validity of the original analysis and offer possible extensions for future work. A few reproductions by Kedron et al. (2022a, b) have brought this approach into the present, but formal, published reproductions and replications that systematically examine the entire research process remain rare in the geographic literature.

A practical approach to the reproduction of geographic research

The present dearth of reproductions evaluating the entire research process is likely due, at least in part, to the current absence of a model approach that researchers can use to guide their reproduction attempts. Here we introduce such an approach with a workflow and template materials to facilitate implementation by others. Building on prior workflow models of the computational reproduction process, we developed a three-stage workflow (Fig. 1) to guide the reproduction of geographic research. Our workflow model presents a high-level organization of key tasks common across reproduction attempts. Almost every component within the model could be further expanded into a significant submodel and customized for different subfields in geography. However, we restrict our presentation here to the higher level because our goal is to instigate a shift in how we pursue reproduction across a variety of research areas. Below, we outline the Planning, Implementation, and Evaluation steps of our approach.

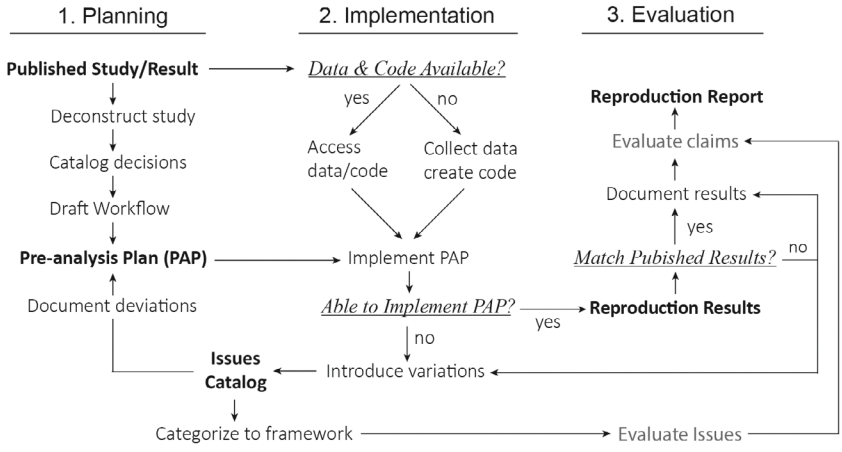

Figure 1. A researcher attempting to reproduce a prior study will first deconstruct the existing work and plan their attempt. When implementing that plan the research may encounter unexpected issues that should be cataloged, leading to plan amendments. Once the attempt is completed, the researcher will then complete and share a report of the reproduction attempt.

To facilitate adoption, we have paired our model with a template repository designed to help organize the reproduction process. The repository contains document templates and suggestions on how to use and modify the repository structure. The template repository is available online as a Git repository under a BSD 3Clause License through (Kedron and Holler 2022b). We used these materials to conduct the reproductions presented in this article.

Planning

Before beginning any data analysis, researchers attempting a reproduction should first carefully deconstruct the design and implementation of the prior analysis and create a workflow model for their own analyses. It is essential for researchers to clearly articulate the aspects of the prior study they intend to reproduce. For example, in the case of hypotheses-driven research, the reproducing researcher should communicate which research questions and hypotheses will be the focus of their reproduction and how they intend to gather data, execute their analyses, and compare their results. While this step may appear trivial, many studies do not formally state the hypotheses and provide only a partial description of the analytical plans. Researchers also often test a large number of hypotheses during the course of their study but highlight only a handful of those results. This situation leaves the reproducing researcher with a choice as to which hypotheses to retest and the need to explain why some hypotheses were omitted.

In our approach, researchers are prompted to formally record and present their reproduction workflow as part of a pre-analysis plan that details the data collection, processing, and analysis that will be undertaken as part of their reproduction attempt. The workflow should be built on the most complete and precise understanding of procedures including supplemental information and code. If complete procedural detail cannot be reconstructed, the plan should include the reproducing researchers' best approximation of the procedures for data processing and analysis. The pre-analysis plan should also include the criteria the reproducing researcher will use to compare their results to those of the original study. Ideally, this plan is publicly registered before

any reproduction attempt begins. Pre-analysis plans can be shared via platforms such as GitHub, or more formally registered through services like the Open Science Framework.[1]

In the exemplar reproductions presented below, we first created initial workflow models and drafted pre-analysis plans. Iterative revisions of the plans were vital to identifying ambiguity and error in the original research design. Researchers should catalog such issues as they arise during any reproduction attempt.

Implementation

Once pre-analysis planning is completed, reproduction attempts move to data collection, preparation, and analysis. A common practical barrier is whether or not the data, as well as the procedures and protocols used to gather those data, are available. If data are not included with the publication, they can sometimes be accessed from an original source (e.g., U.S. Census). In instances where the data need to be processed prior to analysis, it is recommended to use the original code. When the original code is unavailable, processing steps should be recorded, and any deviations from the original procedural plan should be documented.

Pre-analysis plans are dynamic communication tools that are designed to track unanticipated changes that may occur throughout the reproduction process. As these plans are implemented, any ambiguities encountered should be cataloged and evaluated (Kedron et al. 2021a). Any decisions made to resolve these ambiguities should be included as amendments to the pre-analysis plan. Researchers can also introduce differences into their reproduction attempts to test the sensitivity of the original analysis to alternative conceptualizations or research designs. However, as these variations are introduced, they should be tracked along with a justification for each change.

Evaluation

The results of a reproduction attempt should be compared to those of the original study as they are created. If discrepancies arise early in the analysis (e.g., differences in descriptive statistics), the procedures should be revised and these unplanned deviations from the original workflow should be documented in the pre-analysis plan before proceeding to subsequent analyses. There is no universally agreed upon criteria to assess whether the results of an original study have been reproduced, and most literature focuses on replication (Verhagen and Wagenmakers 2014; Open Science Collaboration 2015; Simonsohn 2015; Lakens 2017). In prior reproductions of geographic research, evaluation has been based on exactly matching numerical results or producing similar figures and maps. At a minimum, the direction, magnitude, and uncertainty associated with both sets of results should be compared. Observed differences between results will often motivate the introduction of further variations in research design and analysis. When new variations are introduced, they should be tracked.

The documentation and comparison of the results of a reproduction attempt and an original study is only one part of the evaluation process. To evaluate the claims made in a study, the reproducing authors should also evaluate the complete set of decisions made by the authors of the original study. Careful documentation of the decisions and changes a reproducing researcher makes during a reproduction attempt provides the foundation for the evaluation of research claims.

Empirical context, and the selection of studies for reproduction

Empirical context

To demonstrate the approach outlined above and move beyond computational reproductions, we attempted to reproduce three geographical analyses of COVID-19. The COVID-19 pandemic

has highlighted the importance of making reproductions and replications a standard part of the geographic research process. The rate of research publication during the pandemic has led to concerns over the quality of peer review and the rate of retractions (Yeo-Teh and Tang 2020). Well into the COVID-19 pandemic, researchers continue to produce studies intended to advance our understanding of the spatial patterns of this disease (e.g., Sun et al. 2020; Chakraborty 2021; Sugg et al. 2021) and the spatial processes that may be responsible for the spread of the SAR-CoV-2 virus (e.g., Andersen et al. 2021; Lee and Ramírez 2022). Many of these geographical analyses have been undertaken by researchers with expertise outside of geography and published at an accelerated pace due to the urgency and scale of the pandemic. Medical professionals, government officials, and policymakers continue to use this stream of research to inform their pandemic response. To ensure those groups have access to the best possible research for decision making, the reliability and the credibility those results must be known so that findings can be appropriately assessed. Understanding the validity of these studies is also important because they are already becoming the foundation for future research.

Many authors (e.g. Gustot 2020; Sumner et al. 2020; Collins and Alexander 2022) have emphasized the importance of reproducing COVID-19 research and have begun to catalog the availability of code and data within the literature. Geographers have produced similar catalogs of geographical analyses of COVID-19 but have limited their reviews to listing and categorizing the literature by topical focus and methodological approach (Agbehadji et al. 2020; Ahasan et al. 2020; Franch-Pardo et al. 2020, 2021). Very few formal reproductions of geographical analyses of COVID-19 are available in the literature (Paez 2022; Kedron et al. 2022a). Conducting reproductions of COVID-19 research facilitates the assessment of the internal validity of selected studies and demonstrates the value of reproductions.

Selection of candidate studies for reproduction

To identify candidate studies for reproduction, we conducted an electronic search for peer-reviewed spatial analyses of COVID-19 published in English language journals between January 1, 2020 and March 15, 2021. To enhance the impact of our work, we sought to identify studies that relied on the most commonly used sources of COVID-19 data and were based on methods frequently used in spatial epidemiology. Candidate studies were identified by searching Elsevier's Scopus database using the search query:

> ("COVID-19" OR "SarS-CoV-2" OR "2019-nCoV" OR "2019 coronavirus" OR "2019 novel coronavirus" OR "novel coronavirus") AND ("GIS" OR "Spatial Analysis" OR "Geospatial Analysis" OR "ArcGIS" OR "Geographic Information System" OR "Geographic Mapping")

We designed this query to mirror the search criteria of Ahasan et al. (2020)'s review of geographical analyses of COVID-19. We also independently searched the Google Scholar database using the same search terms to identify additional studies. The first Scopus search was run February 9, 2021, and the Google Scholar search was conducted February 18, 2021. A limited updated literature search was performed between March 15, 2021 and March 30, 2021. These searches yielded 540 unique articles. We collected abstracts and full texts for each of these articles.

Article abstracts were further filtered according to whether the study occurred in the United States, which was done in order to better ensure access to data. This criteria narrowed the set to 60 articles. We then reviewed the full articles to determine if the statistical method used was

Table 1. Characteristics of the geographical analyses of COVID-19 selected for reproduction

	Mollalo et al. (2020)	Saffary et al. (2020)	Vijayan et al. (2020)
Data available	Yes	No	No
Code available	No	No	No
Processing environment	Not specified	Not specified	Not specified
Spatial extent	United States	United States	LA County
Spatial support	County	County	10-km Hexagons
Temporal extent	January–April 2020	February–May 2020	February-June 2020
Hypothesis tests	1000s	1000s	1000s
Methods	SEM, SLM, GWR, MGWR	Moran's I, Bivariate Moran's I	Moran's I, SLM

common in spatial epidemiology (e.g., spatial regression, cluster analysis). This review narrowed our list to 15 candidate articles. Based on the completeness of their publication details, study objectives, data sources, data and code availability, and spatial methodology, we selected three articles – Mollalo, Vahedi, and Rivera (2020), Saffary et al. (2020), and Vijayan et al. (2020). These articles use spatial statistical methods common in both spatial epidemiology and the broader geographic literature and appeared feasible to reproduce Table 1.

Implementation of the reproduction attempts

We followed the three-stage process of planning, implementation, and evaluation outlined by the model approach (Fig. 1). The entire reproduction process for each study is documented in a research compendium that includes our reproduction plans, reports, data, and code. Each compendium is available online as a Git repository under a BSD 3-Clause License to allow other researchers to examine our approach and use our work as a model for future reproductions. The details of each reproduction can be accessed through Kedron, Bardin, and Holler (2023) – https://osf.io/wxyf5/, and the template repository used to create each reproduction is available through Kedron and Holler (2022b) – https://osf.io/w29mq.

During the planning stage of each reproduction attempt, we focused on developing a model workflow and pre-analysis plan. We used an iterative process to create the reproduction workflow. Each author developed their own model workflow, which they then presented to the other authors. We then collectively identified the chain of researcher decisions and points of uncertainty that needed to be addressed, resulting in a single common workflow. The workflow for each reproduction attempt then became the foundation of the pre-analysis plans, which also identified the key hypotheses we sought to retest and any deviation we anticipated due to a lack of information provided in the original study.

Whenever possible, we used the data provided by the authors and followed the procedures described in the original article when attempting each reproduction. When data were not available, we acquired the data from public sources or contacted the corresponding author to request inaccessible data. When we encountered missing data sources or procedures, we attempted the reproduction using alternative data sources and procedural decisions. We attempted to use the processing environment described in the original study but also translated the workflow into R code.

Following the approach described above, our evaluation of each reproduction attempt consisted of two parts: (1) an assessment of result similarity and computational reproducibility, and (2) an evaluation of the execution and claims of the original study. To assess reproducibility, we used the simple criteria of whether the results of the reproduction and original analysis were numerically or graphically identical. However, for analyses that relied on conditional permutation of the data to estimate parameters and make statistical inferences, we relaxed the criteria of identical reproduction and instead focused our evaluation on the comparison of parameter estimates, related uncertainty estimates, and statistical (P-values)[2]. These criteria mirror those presented by the NASEM (2019) and those used by Open Science Collaboration (2015). We were able to partially verify the original findings of each of the three studies, albeit not without challenges (section 6.1).

Our meticulous deconstruction of each study and efforts to achieve computational reproduction revealed points of concerns in each study. Those questions motivated our further investigation through a reanalysis of each study. Moving beyond an assessment of computational reproducibility, we evaluated each study using the framework presented by Kedron et al. (2021a) and linked points of concern that arose during each reproduction attempt to different stages of the research cycle. We also compared those issues across the studies to identify common points of strength or weakness (section 6.2).

After completing the reproduction attempts, we created reproduction reports by updating the original pre-analysis plans to include a record of any unplanned deviations, the results of the reproduction including comparison to the original study, and a discussion of the results. Final reports were posted within each reproduction compendium.

Lessons from the three reproductions

The three studies selected for reproduction use spatial regression techniques and local spatial statistics to make associational inferences about COVID-19 (Table 1). Mollalo, Vahedi, and Rivera (2020) specified a series of spatial regression models to evaluate variation in county-level COVID-19 incidence using a set of socioeconomic and demographic characteristics as predictor variables. The authors present five regression models including an ordinary least squares (OLS) model, spatial lag model (SLM), spatial error model (SEM), geographically weighted regression (GWR), and a multiscale GWR (MGWR). Neither the data nor the code used for the original analysis was made available by the authors. Saffary et al. (2020) use bivariate Moran's I to examine whether socio-demographics and health care resources are correlated in space with COVID-19 cases and deaths across the contiguous United States. The authors do share the county-scale data used in their analyses. Vijayan et al. (2020) examine whether spatial patterns existed in SARS-CoV-2 age-adjusted testing rates, age-adjusted diagnosis rates, and crude positivity rates in Los Angeles County (LAC), and use a spatial regression model to explore associations between crude positivity rates and a series of predictor variables. Although the original study data was not publicly available, we were able to obtain it by request from the corresponding author. The analysis code was not made available, nor was information about the computational environment used.

Computational reproductions

We were able to partially reproduce the analyses and results of each of the three studies we investigated (Table 2). The extent to which we were able to reproduce the results of each

Table 2. Computational reproducibility of the select geographical analyses of COVID-19

	Mollalo et al. (2020)	Saffary et al. (2020)	Vijayan et al. (2020)
Descriptive statistics	Not specified	Fully	Fully
Direction of regression coefficients	Fully	Fully	Partially
Magnitude of regression coefficients	Partially	Fully	Fully
Statistical significance	Fully	Fully	Partially
Maps	Partially	Partially	Partially

study was directly related to the availability of original data and the detail of the procedural description provided. We were able to create exact reproductions of nearly all the tables and maps presented by Saffary et al. (2020) in part because these authors provided their data file. Conversely, Mollalo, Vahedi, and Rivera (2020) did not provide their data and offered limited descriptions of their data sources, which hindered our reproduction attempt and produced the least consistent results. We were similarly unable to reproduce the results of Vijayan et al. (2020) on our initial attempt, because we could not reconstruct the hexagonal tessellation, or access identical neighborhood-level COVID data. Once the authors provided these data, we were able to create an exact reproduction of their descriptive statistics and obtain consistent spatial regression estimates.

To partially reproduce the computational results each study, we had to make unplanned deviations from our initial plans. For example, while Saffary et al. (2020) published their data, that file did not include one of the key independent variables, requiring us to gather this missing variable from public sources. While we were able to collect the necessary data, some locations in the file had missing values. The authors provided no information as to how to handle those missing values, which we ultimately determined were simply omitted from analysis. While we were able to obtain consistent regression estimates when reproducing Vijayan et al. (2020), we had to adjust our original plans when reproducing the authors' LISA analyses. We found low-high and high-low clusters that were not reported by the original authors. If these clusters were purposefully omitted, this decision represents a cartographic form of observed selective inference. We similarly found inconsistencies in how Mollalo, Vahedi, and Rivera (2020) presented the variables in their paper and how they appear to have been processed in their analyses. For instance, the authors did not mention standardizing their variables prior to analysis, yet the magnitude of the reported coefficients suggest that they had been standardized. The authors also reported using the percentage of nurse practitioners as one of their independent variables, but their description of the variable calculation suggests that the count of nurse practitioners was used.

Beyond computational reproductions

Our attempts to reproduce the computational result of the three selected studies raised concerns about the internal and conclusion validity of each study. Following our model approach (see section 3), we cataloged those concerns, linked them to the stages of the research cycle (Table 3), and introduced procedural changes that allowed us to test the affect alternative decisions had on each analysis. To demonstrate the value of this form of reproduction, we discuss the issues we encountered in relation to phases of the research process.

Table 3. Points of concern identified during replication attempts

	Point of concern	Mollalo	Saffary	Vijayan
Conceptualization and design	Consideration of epistemic uncertainty	X	X	X
	Consideration of scale	X	X	X
	Justification of variable selection	X		X
Measurement and processing	Details of data processing	X	X	X
	Description of missing data procedures		X	X
Analysis and inference	Presentation of research hypotheses	X	X	X
	Atomistic fallacy and MAUP	X	X	X
	Model specification and test execution	X	X	X
	Adjustment for multiple hypothesis testing		X	
Communication	Lack of provenance information	X	X	X
	Selective inference	X	X	X

Conceptualization and design

Many of the issues we encountered when designing our reproduction attempts and interpreting their results stemmed from the conceptualization and design of the original studies. We found four overarching issues of concern. First, the authors of our target studies did not address the epistemic uncertainties potentially impacting their analyses. For example, in each study the primary response variable was the count of COVID-19 cases or deaths early in the pandemic. In principle, we could have known these counts. However in practice, testing capacity was limited and geographically variable during the study periods, and asymptomatic cases often went undetected. These factors likely contributed to a spatially varying undercount of disease prevalence. Acknowledging this systematic uncertainty in case and death counts (Halpern et al. 2021) is important because geographic variation in count reliability can impact parameter estimation. To be clear, we would not expect the authors to resolve these issues with the data available. However, understanding and explaining how uncertain critical measurements are is fundamental to placing inferences and claims in a proper context. While reproducing these studies allowed us to identify this concern, our results and inferences are similarly impacted by this issue.

Second, we believe these studies would benefit from deeper consideration of how the spatial and temporal supports of their data impact analysis. Two of the studies use counties as their spatial support, while the third constructs a hexagonal grid. This selection seems to be largely a matter of data availability and convenience and is mismatched with our knowledge of the transmission dynamics of COVID-19 (Wali and Frank 2021). For example, while Vijayan et al. (2020) use a hexagonal grid as their spatial support, variable construction within that grid ignored variation in the geography of the administrative units of the original data. Moreover, the Census data used as predictors of COVID-19 incidence was collected before the pandemic raising questions about the spatial support of each study. While we expect some degree of temporal consistency in the sociodemographic profiles of these units, the pandemic also created migration patterns (Haslag

and Weagley 2021; Coven, Gupta, and Yao 2023) that may make measures from several years before the pandemic a poor match to the actual populations present in those location during the pandemic. Addressing these mismatches is difficult given data availability and the rapidity of the pandemic, but acknowledging and discussing the potential impact of measurement issues would help readers better understand the implications and reliability of each study. Moreover, recent studies in different geographic contexts (González-Leonardo et al. 2022; Rowe et al. 2022) point to alternative measures of population migration that could be now used to reassess these issues.

Third, our reproduction attempts led us to question how the original authors incorporated the current understanding of the epidemiology of COVID-19 into their operationalization of spatial relationships and their selections of the spatial scale of their analyses. Two of the three studies reproduced sought to identify ecological predictors of COVID-19, and were conducted using counties as the spatial support for all analyses. However, epidemiological research suggests that counties are not a meaningful unit of analysis for COVID-19 transmission, which happens at a much finer spatial scale (Wali and Frank 2021). Even when counties are used as proxies to measure ecological relationships, it is critical to adjust for other factors that would influence transmission within and between counties, such as population density or the presence of a large urban center. These factors were not included in the original analyses, which may have led to erroneous inferences. For example, it is not clear that these studies provide evidence of a predictive link between racial minority status and COVID-19 case counts when adjusting for urban–rural differences that were not included in the analyses. How the authors treated spatial scale also appears to have led to instances of the atomistic fallacy, which we discuss in section 6.2.3.

Fourth, when interpreting the results of our reproduction attempts, we found it difficult to identify why the authors included some ecological factors in their models but excluded others. We were unable to assess how reliable identified associations were when potentially important confounding factors were omitted from the analyses. Without understanding why the authors believed a factor would affect aggregated COVID-19 case or death counts at a particular scale, we could not assess how the patterns presented provided information about the processes that might be responsible for them.

Measurement and processing

While recreating and processing the data used in each study, we discovered concerns related to variable construction and construct validity. It was unclear how Saffary et al. (2020) handled counties missing primary care physician information. We investigated three alternative procedures when attempting to resolve this ambiguity – filtering, zero imputation, and mean imputation. Our findings suggest the authors omitted missing values. Authors were also unclear about when and how their data was standardized. Saffary et al. (2020) chose to analyze the raw count of intensive care beds in each county without adjusting for county population. The strong positive correlation between the number of such beds and county population nearly makes this unadjusted variable a measure of county population. Vijayan et al. (2020) indicate standardizing variables prior to spatial regression modeling, but are not clear about which variables were standardized and provide limited information about the specification of their spatial regressions. The authors report and discuss their coefficient estimates without referencing the fact that these results are based on standardized variables and that the model intercept was omitted from their analysis.

Our reproduction attempts also uncovered questions about how authors created the spatial support for their analyses. This concern is best illustrated by Vijayan et al. (2020) who based their statistical analyses on a 10-km hexagonal grid that they superimposed onto Los Angeles County, CA. The authors did not (1) present a clear justification as to why this grid was an appropriate unit of analysis, (2) provide the information needed to reconstruct the grid, or (3) include a discussion of how their data aggregation procedures might impact their analyses. We ultimately determined that the authors aggregated data originally linked to different areal units (e.g., Census tracts, municipalities) to their hexagonal grid based on the overlap between that grid and the centroids of the areal units of the data. This approach ignores the proportion of geographic overlap between the hexagonal grid and the source data and could lead to nonrepresentative measurements. Moreover, the age-adjusted response variables used in these analyses are problematic. Given that the age-adjustment was not based on the population within the hexagonal units but the mix of areas whose centroids fell within a given hexagon, these response variables are no longer accurately age-adjusted. Selecting a single unit of analysis and aggregating data in this way introduces unknown amounts of measurement error into any subsequent analysis and creates the possibility for inferential errors.

Analysis and inference

The reproduction attempts provide cautionary lessons about the implementation and interpretation of spatial statistical tests of COVID 19. First, as is commonly the case, the authors of all three studies did not clearly present the complete discrete set of hypotheses to be tested prior to their analyses. Authors made statements about expected associations between COVID-19 incidence and some key independent variables, but did not formalize these hypotheses. In some cases, the authors also tested other unstated hypotheses or tested the stated hypotheses multiple times. Without formal hypothesis statements, these studies are best viewed as exploratory analyses of possible spatial associations between aggregated measures.

Second, the reproductions highlight the need to carefully consider, and explain in text, the reasoning supporting the conceptualization of scale and spatial relationships implemented during spatial statistical tests. In these three studies, the reasoning behind the implementation of the statistical tests seems to be subject to the atomistic fallacy. In each study, the authors root their variable selection and model specification decisions in knowledge and reasoning about the individual-level dynamics of COVID-19 transmission. However, a geographic area is used as the spatial support for analysis in each study and the variables used in each statistical test are aggregated to those units. These choices implicitly scale the individual-level reasoning for variable selection to the group level at these geographic scales. This scaling may be fallacious. For example, Saffary et al. (2020)'s use of the Bivariate Moran's I to measure associations across space extends assumptions about individual-level disease dynamics to the group-level and inter-county-scale. It is not clear, for example, that the evidence and reasoning supporting the belief that an individual person of color might be at a higher risk of contracting COVID would extend to all people of color in a county, or to all people of color in counties surrounding a county with COVID cases. This type of epidemiological study which is based on aggregate social data should be interpreted with caution as exploratory and should be supported by further individual-level or multiscale research.

We have no information about how sensitive each study may be to the modifiable areal unit problem because each study only reported a single spatial support and spatial extent. It may well be the case that studying these relationships at a different spatial scale would change these results.

As one example, selection of a grid size different from the 10-km hexagons adopted by Vijayan et al. (2020) could produce different association estimates. These results may be particularly sensitive to variation across spatial supports given that the centroid-based aggregation of data will produce different levels of measurement error for each hexagonal grid size. This form of uncertainty can be be better understood by testing result sensitivity to alternative spatial supports and utilizing alternative methods of spatial reaggregation based on overlap of tract areas or residential buildings.

Third, the reproductions suggest spatial statistical analyses of COVID-19 may be subject to model specification and interpretation problems. For example, Mollalo, Vahedi, and Rivera (2020) considered 34 variables for inclusion in their regression analyses, but relied on a stepwise forward selection procedure to reduce this set to a final total of four variables. This data-driven approach to variable selection positions their final model as a general, exploratory analysis. With only four variables in the final model it is likely that the model does not properly control for important confounding factors that may influence both the predictor and response variables, and thus, the model coefficients are likely to be biased. These issues are compounded by reliance on fit statistics to guide model selection and to measure explanatory power. Based solely on the higher R^2 and lower AICc of their MGWR model, the authors recommend the continued monitoring of these factors to understand spread of the disease. However, this recommendation ignores both the poor model fit of the OLS specification and the maps of the local R^2 from both the GWR and MGWR models which show large numbers of counties with negative R^2 values. Combined, these indicators suggest model underspecification while the substantial difference in the goodness of fit between the local and global models is indicative of overfitting in the local models. There is a need to balance these types of data-driven exploratory analyses with more deductive theory-based approaches to examine theorized mechanisms with inferential power.

Similarly, the reproduction of Saffary et al. (2020) revealed inconsistencies in the implementation and interpretation of the Bi-variate Local Moran's I statistic. When interpreting this statistic, the authors discuss COVID-19 rates as a measure of correlation. However, the statistic was implemented using each focal county's rate of COVID-19 incidence and the spatial lag of adjacent counties' health and demographic compositions. Contrary to the interpretations presented, this implementation suggests that the COVID-19 rates in a county are the product of variable concentrations in surrounding counties. For example, COVID-19 rates in an urban county may be influenced by the rates of minority residents in surrounding counties exclusive of the urban minority rate.

Fourth, two of the reproductions revealed that geographical analyses of COVID-19 may suffer from the problem of uncorrected multiple hypothesis testing. Saffary et al. (2020) search for spatial clustering provides the clearest example of this issue. In their study, the authors executed thousands of local univariate and bivariate tests, but included no adjustment for the number of tests in their main manuscript. As reported, the results are an example of observed selective inference, which occurs when researchers implement many statistical tests, fail to account for the effects of multiple testing, and then emphasize only a subset of their results. Making appropriate adjustments for the large amount of multiple testing done during this analysis is key to making reliable inferences. Using a $P = 0.05$ significance threshold, we would expect 156 "significant" results in a set of 3,105 tests even when no spatial pattern exists. Curiously, Saffary et al. did include Bonferroni and False Discovery Rate adjustments for multiple testing as a supplement to their analysis, but did not interpret these result in their main manuscript.

After applying these adjustments, nearly all of the spatial patterns highlighted in the manuscript disappear.

Communication

The reproduction results reinforce the importance of clearly tracking and communicating the provenance of research before, during, and after a geographical analysis. Many of the problems we encountered when reproducing these studies could have been avoided had the authors documented and shared information about the sources, quality, and uncertainty of their data and the justifications for their analytical decisions. This lack of transparency indirectly led us to more carefully deconstruct each study, which in turn led us to a deeper understanding of how the authors designed and executed their research. Indeed, many of the problems we identified were not apparent when reading the publications, and were further obscured through the lack of data, code, and sufficiently detailed procedural description.

An additional communication problem uncovered through our reproductions is the potential presence of selective inference in these geographical analyses of COVID-19. Selective inference occurs when statistical inference is focused on a finding only after observing the data (Benjamini, Heller, and Yekutieli 2009). The reproductions show the many possible avenues through which unobserved selections could occur. For example, in each study the authors selected a queens contiguity matrix at the county/hexagon scale to represent the spatial relationships underlying patterns of association with COVID-19. While a reasonable starting point, statistical results are sensitive to weights selection, and there is no reason to believe this form of contiguity was the only form tested or the form that appropriately captures the dynamics of the pandemic. Similarly, we demonstrate above how our reanalyses explored alternative missing data procedures, spatial data supports, and model specification decisions.

As published, we cannot know whether selective inferences occurred during these studies, and have no evidence to suggest the authors intentionally or unintentionally made any selective inferences. The important point is that in many instances we do not know what the authors did, but we do have clear evidence that it is very easy to make unintentional selective inferences in any geographical analysis. To provide evidence that selection did not occur, the complete provenance of the research needs to be recorded and shared. This sharing should include any sensitivity analyses or specification check the authors preformed as they focus inference on some models rather than others. Ideally, authors would also preregister or share their hypotheses and analytical plans before they observe their data, thus creating a need to justify any deviations from those plans. Conducting these type of sensitivity analyses and communicating their outcomes frames research decisions and lends credibility to claims.

Conclusion

In this article, we present a model workflow and corresponding materials to help geographic researchers move beyond using reproduction to simply answer whether the computational results of a study can be recreated to assessing whether the data, analysis, and results presented in a study in fact support the claim(s) made by the study authors. We demonstrate how reproduction studies can act as the foundation for testing alternative research designs, problem conceptualizations, and analytical pathways, which can lead to improvements in the quality of geographic research and knowledge production in the discipline. Over the course of this article, we make three principle contributions.

First, we introduce a model workflow for conducting reproduction studies aimed at assessing the claims of published research. The conceptual foundation of our approach is Kedron et al. (2021a)'s representation of the research process as a series of decisions researchers make in the face of uncertainty. We adopt the authors' four part segmentation of the research process, and their discussion of some of the challenges particular to reproducing geographical analyses, as a means of tracking and categorizing decisions made by both the original authors and the researchers attempting to reproduce their work. This approach provides a means of linking the existing literature on challenges and uncertainties in geographical analyses to aspects of the reproduction process. This approach also matches an understanding of research as a continuous process aimed at refining degrees of confidence in our understanding of phenomena, rather than establishing complete certainty.

Second, to demonstrate the use of our approach and materials, we report the findings of our attempts to reproduce and assess the claims of three published geographical analyses of COVID-19 in the United States. We were able to partially reproduce each study, and the reproduction process led us to identify a number of conceptual and methodological concerns that raise questions about the predictive value and the magnitude of the associations presented in each study. Overall, while already highly cited, we believe the studies we reproduced and reanalyzed are best viewed as exploratory analyses of spatial patterns of COVID-19 early in the pandemic. In our view, they provide limited reliable evidence of meaningful associations of substantial magnitude.

In each reproduction study, we go beyond reviewing the availability of data and methods and executing code. Rather, we attempt to recreate all aspects of the procedures of each study regardless of an absence of, or errors in, data and code. By retracing each study's procedure, we scrutinize the work, including details and decisions not communicated in the published manuscript. We highlight questions about the spatial reasoning used when designing these studies and problems in the application of spatial statistical techniques used regularly in the geographic literature. As we encounter shortcomings in the research design and discrepancies between the manuscript, the procedures, and the reported results, we reanalyze the study and correct errors. Identified errors and uncertainties are presented and discussed in reports. Each of the three reproduction studies is published with open-source licensing as a reproducible research compendium composed of data, code, pre-analysis plans and detailed reports of our results (Kedron et al. 2022c, d, e). We thereby improve the computational reproducibility of these published studies, provide an enriched assessment of their claims, and facilitate any future research attempting to replicate or extend these studies.

Third, we review the reproduction process and use the information gathered during our attempts to identify how we might systematically use reproduction studies to assess and enhance future geographical research. We identify threats to conclusion and internal validity involving geographic space and connect those threats to de cision points in the research process. The concerns highlighted in this article can serve as a guide for others seeking to implement original research with these techniques in a principled manner. We similarly believe our work can be incorporated into coursework when training future geographic analysts, as these analyses were conducted under the supervision of the lead authors in collaboration with graduate students early in their respective programs. To our knowledge, this article is one of the first attempts to push reproduction attempts beyond computation in the geographical sciences.

Despite the questions revealed by our reproductions, these papers all passed through peer-review and, in some cases, are garnering significant positions in the literature. As of

September 7, 2022, Mollalo, Vahedi, and Rivera (2020) has received 300 citations on Scopus and 472 citations on Google Scholar. Our work therefore raises questions about the peer-review process, while demonstrating the potential value of incorporating reproductions into that process. We believe that had reviewers reproduced these studies or had access to fully reproducible research compendia complete with data and code, they would have found at least some of the issues we raise. We hope that further revisions would have addressed some of our identified concerns. However, simply re-executing the code and data used in these studies would not have identified many of the issues raised in this article.

The discussion and practice of reproducibility in geography should not be limited to matters of sharing research artifacts and recomputing results. This insight has implications that extend beyond the reproduction of a single study to the institutional changes we might pursue to improve the creation and accumulation of geographic knowledge. For one, our findings support a case for geographic journals considering not just requiring the submission of research materials but also incentivizing comprehensive reproduction studies. For example, editors could commission reproduction studies of selected articles, pair publications of reproductions and original author response, or create recurrent special issues of reproductions or replications in their given field. These institutional changes can help to identify, communicate, and improve recurrent issues with geographic analyses in geography and adjacent disciplines.

We might similarly incorporate comprehensive reproduction studies into our graduate coursework. Conducting rigorous reproduction is a time-consuming endeavor that is currently not incentivized by academic review process. As such it seems likely that many academics do not conduct formal reproductions, or if they do conduct them do not pursue the publication of those results, creating the present shortage. We have demonstrated that graduate students can conduct high-quality reproductions using our practical framework to structure their approach. Although we did not formally document their experiences, we found graduate students interested to engage in the reproduction studies as they provided an opportunity to both learn techniques and contribute formally to the geographic literature. To this end, folding reproductions into coursework may produce the dual benefit of introducing more reproductions into the literature while preparing the next generation of geographic researchers to work in a reproducible manner. Our hope is that this work will start a culture or reproduction and replication in geography, and the open sharing of any such efforts.

Author Contribution

Peter Kedron led study conceptualization, methodology, writing; and supervision and administration of reproductions. **Sarah Bardin** performed reproduction attempts and contributed substantially to reproduction review and all writing tasks. She also led data curation and software development. **Joseph Holler** contributed to methodology development, writing - review and editing. **Joshua Gilman**, **Bryant Grady**, **Megan Seeley**, **Xin Wang**, **Wenxin Yang** undertook the initial reproductions.

Acknowledgements

We would like to recognize and thank Middlebury graduates Derrick Burt and Drew An-Pham who undertook independent reviews of our three reproduction attempts and corresponding repositories as part of their undergraduate research experiences. We similarly acknowledge ASU

graduate students Summer Cliff, Kim Fuller, and Lev VanZanderberg who participated in our initial reproduction attempts. This article is based on work supported by the National Science Foundation under Grant No. (BCS-2049837), and a fellowship awarded to Kedron and Holler under Grant No. (OAC-1743184).

Note

1 See Christensen, Freese, and Miguel (2019) and Olken (2015) for a discussion of the pros and cons of preregistration.

2 Gelman and Stern (2006) present the challenge of comparing statistical significance across studies and caution against basing conclusions on changes in significance alone. We incorporated this thinking into our analyses but retained comparisons of significance levels, because as reproductions our work uses the same data and methods, which should lead to the same or very similar significance levels.

References

Agbehadji, I. E., B. O. Awuzie, A. B. Ngowi, and R. C. Millham. (2020). "Review of Big Data Analytics, Artificial Intelligence and Nature-Inspired Computing Models towards Accurate Detection of COVID-19 Pandemic Cases and Contact Tracing." *International Journal of Environmental Research and Public Health* 17(15), 5330.

Ahasan, R., M. S. Alam, T. Chakraborty, and M. M. Hossain. (2020). "Applications of GIS and Geospatial Analyses in COVID-19 Research: A Systematic Review." *F1000 Research* 9, 1379.

Andersen, L. M., S. R. Harden, M. M. Sugg, Ph.D., J. D. Runkle, Ph.D., and T. E. Lundquist. (2021). "Analyzing the Spatial Determinants of Local COVID-19 Transmission in the United States." *Science of the Total Environment* 754, 142396.

Benjamini, Y., R. Heller, and D. Yekutieli. (2009). "Selective Inference in Complex Research." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367(1906), 4255–71.

Brunsdon, C. (2016). "Quantitative Methods I: Reproducible Research and Quantitative Geography." *Progress in Human Geography* 40(5), 687–96.

Brunsdon, C., and A. Comber. (2021). "Opening Practice: Supporting Reproducibility and Critical Spatial Data Science." *Journal of Geographical Systems* 23(4), 477–96.

Brunsdon, C., and A. Singleton. (2015). *Geocomputation: A Practical Primer*, Vol 2015. London, UK: Sage.

Chakraborty, J. (2021). "Social Inequities in the Distribution of COVID-19: An Intra-Categorical Analysis of People with Disabilities in the US." *Disability and Health Journal* 14(1), 101007.

Christensen, G., J. Freese, and E. Miguel. (2019). *Transparent and Reproducible Social Science Research*. Oakland, CA: University of California Press.

Clemens, M. A. (2017). "The Meaning of Failed Replications: A Review and Proposal." *Journal of Economic Surveys* 31(1), 326–42.

Collins, A., and R. Alexander. (2022). "Reproducibility of COVID-19 Pre-Prints." *Scientometrics* 127.8, 4655–73. https://doi.org/10.1007/s11192-022-04418-2

Coven, J., A. Gupta, and I. Yao. (2023). "JUE Insight: Urban Flight Seeded the COVID-19 Pandemic across the United States." *Journal of Urban Economics* 133, 103489. https://doi.org/10.1016/j.jue.2022.103489

Davies, W. K. D. (1968). "The Need for Replication in Human Geography: Some Central Place Examples." *Tijdschrift Voor Economische en Sociale Geografie* 59(3), 145–55.

Earp, B. D., and D. Trafimow. (2015). "Replication, Falsification, and the Crisis of Confidence in Social Psychology." *Frontiers in Psychology* 6, 621.

Franch-Pardo, I., B. M. Napoletano, F. Rosete-Verges, and L. Billa. (2020). "Spatial Analysis and GIS in the Study of COVID-19. A Review." *Science of the Total Environment* 739, 140033.

Franch-Pardo, I., M. R. Desjardins, I. Barea-Navarro, and A. Cerdà. (2021). "A Review of GIS Methodologies to Analyze the Dynamics of COVID-19 in the Second Half of 2020." *Transactions in GIS* 25(5), 2191–239.

Gelman, A., and H. Stern. (2006). "The Difference between "Significant" and "Not Significant" Is Not itself Statistically Significant." *The American Statistician* 60(4), 328–31.

González-Leonardo, M., A. López-Gay, N. Newsham, J. Recaño, and F. Rowe. (2022). "Understanding Patterns of Internal Migration during the COVID-19 Pandemic in Spain." *Population, Space and Place* 28(6), e2578. https://doi.org/10.1002/psp.2578

Goodchild, M. F., and W. Li. (2021). "Replication across Space and Time Must be Weak in the Social and Environmental Sciences." *Proceedings of the National Academy of Sciences* 118, 35.

Goodchild, M. F., A. S. Fotheringham, P. Kedron, and W. Li. (2021). "Introduction: Forum on Reproducibility and Replicability in Geography." *Annals of the American Association of Geographers* 111(5), 1271–4.

Gustot, T. (2020). "Quality and Reproducibility during the COVID-19 Pandemic." *JHEP Reports* 2, 4.

Halpern, D., Q. Lin, R. Wang, S. Yang, S. Goldstein, and M. Kolak. (2021). "Dimensions of Uncertainty: A Spatiotemporal Review of Five COVID-19 Datasets." *Cartography and Geographic Information Science*, 1–22. https://doi.org/10.1080/15230406.2021.1975311

Haslag, P. H., and D. Weagley. (2021). "From LA to Boise: How Migration Has Changed during the COVID-19 Pandemic." SSRN 3808326.

Hofer, B., K. W. Broman, C. Granell, A. Graser, K. Hettne, D. Nust, and M. Teperek. (2019). "Reproducible Publications at AGILE Conferences–Proposed Guidelines for Authors and Reviewers." In *Accepted Short Papers and Posters from the 22nd AGILE Conference on Geo-Information Science, Limassol, Chipre, Editorial, Stichting AGILE*. 2019.

Kedron, P., and J. Holler. (2022a). "Replication and the Search for the Laws in the Geographic Sciences." *Annals of GIS* 28(1), 45–56.

Kedron, P. and J. Holler (2022b). "Template for Reproducible and Replicable Research in Human-Environment and Geographical Sciences." 2022. https://doi.org/10.17605/OSF.IO/W29MQ.

Kedron, P., A. E. Frazier, A. B. Trgovac, T. Nelson, and A. S. Fotheringham. (2021a). "Reproducibility and Replicability in Geographical Analysis." *Geographical Analysis* 53(1), 135–47.

Kedron, P., W. Li, S. Fotheringham, and M. Goodchild. (2021b). "Reproducibility and Replicability: Opportunities and Challenges for Geospatial Research." *International Journal of Geographical Information Science* 35(3), 427–45.

Kedron, P., Z. Hilgendorf, M. Sachdeva, and M. Sachdeva. (2022). "Using the Specification Curve to Teach Spatial Data Analysis and Explore Geographic Uncertainties." *Journal of Geography in Higher Education* 46, 304–14.

Kedron, P., S. Bardin, T. D. Hoffman, M. Sachdeva, M. Quick, and J. Holler. (2022a). "A Replication of DiMaggio et al. (2020) in Phoenix, AZ." *Annals of Epidemiology* 74, 8–14. https://doi.org/10.1016/j.annepidem.2022.05.005

Kedron, P., J. Holler, S. Bardin, and Z. Hilgendorf. (2022b). "Reproducibility, Replicability, and Open Science Practices in the Geographical Sciences." 2022. https://doi.org/10.17605/OSF.IO/C5A2R.

Kedron, P., S. Bardin, J. Holler, J. Gilman, B. Grady, M. Seeley, X. Wang, and W. Yang. (2022c). "Reproduction of Mollalo et al. 2020." 2022. https://doi.org/10.17605/OSF.IO/e43kq.

Kedron, P., S. Bardin, J. Holler, J. Gilman, B. Grady, M. Seeley, X. Wang, and W. Yang. (2022d). "Reproduction of Saffary et al 2020." 2022. https://doi.org/10.17605/OSF.IO/qfkg4.

Kedron, P., S. Bardin, J. Holler, J. Gilman, B. Grady, M. Seeley, X. Wang, and W. Yang. (2022e). "Reproduction of Vijayan et al 2020." 2022. https://doi.org/10.17605/OSF.IO/my5dz.

Kedron, Peter, S. Bardin, and J. Holler (2023). "A Framework for Moving Beyond Computational Reproducibility: Lessons from Three Reproductions of Geographical Analyses of COVID19." MetaArXiv Preprints 2023. https://doi.org/10.17605/OSF.IO/WXYF5.

Kmoch, Alexander, D. Nust, and E. Uuemma (2020). "Reproducible Submission Workflow". In: *Proceedings of the 5th AGILE PhD School 2019*. 1–10.

Konkol, M., C. Kray, and M. Pfeiffer. (2019). "Computational Reproducibility in Geoscientific Papers: Insights from a Series of Studies with Geoscientists and a Reproduction Study." *International Journal of Geographical Information Science* 33(2), 408–29.

Konkol, M., D. Nüst, and L. Goulier. (2020). "Publishing Computational Research-a Review of Infrastructures for Reproducible and Transparent Scholarly Communication." *Research Integrity and Peer Review* 5(1), 1–8.

Lakens, D. (2017). "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses." *Social Psychological and Personality Science* 8(4), 355–62.

Lee, J., and I. J. Ramírez. (2022). "Geography of Disparity: Connecting COVID-19 Vulnerability and Social Determinants of Health in Colorado." *Behavioral Medicine*, 48(2), 72–84. https://doi.org/10.1080/08964289.2021.2021382

Mollalo, A., B. Vahedi, and K. M. Rivera. (2020). "GIS-Based Spatial Modeling of COVID-19 Incidence Rate in the Continental United States." *Science of the Total Environment* 728, 138884.

Muenchow, J., S. Schäfer, and E. Krüger. (2019). "Reviewing Qualitative GIS Research – Toward a Wider Usage of Open-Source GIS and Reproducible Research Practices." *Geography Compass* 13(6), e12441.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

Nichols, J. D., M. K. Oli, W. L. Kendall, and G. S. Boomer. (2021). "Opinion: A Better Approach for Dealing with Reproducibility and Replicability in Science." *Proceedings of the National Academy of Sciences* 118, 7.

Nosek, B. A., J. R. Spies, and M. Motyl. (2012). "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability." *Perspectives on Psychological Science* 7(6), 615–31.

Nüst, D., and M. Hinz. (2019). "Containerit: Generating Dockerfiles for Reproducible Research with R." *Journal of Open Source Software* 4(40), 1603.

Nüst, D., and E. Pebesma. (2021). "Practical Reproducibility in Geography and Geosciences." *Annals of the American Association of Geographers* 111(5), 1300–10.

Nüst, D., C. Granell, B. Hofer, M. Konkol, F. O. Ostermann, R. Sileryte, and V. Cerutti. (2018). "Reproducible Research and GIScience: An Evaluation Using AGILE Conference Papers." *PeerJ* 6, e5072.

Nüst, D., F. Ostermann, C. Granell, and A. Kmoch. (2020). "Reproducibility Reviews AGILE 2020." https://doi.org/10.17605/OSF.IO/6K5FH.

Nüst, D., F. Ostermann, B. Hofer, C. Granell, and R. Sileryte. (2021). "Reproducible Research at AGILE".

Nüst, D., F. Ostermann, E. Koukouraki, P. Friese, J. Krukar, R. Decoupes, C. Granell, and E. Tomai. (2022). "Reproducibility Reviews AGILE 2022."

Olken, B. A. (2015). "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29(3), 61–80.

Open Science Collaboration. (2015). "Estimating the Reproducibility of Psychological Science." *Science* 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Ostermann, F. O., and C. Granell. (2017). "Advancing Science with VGI: Reproducibility and Replicability of Recent Studies Using VGI." *Transactions in GIS* 21(2), 224–37.

Ostermann, F. O., D. Nüst, C. Granell, B. Hofer, and M. Konkol. (2021). "Reproducible Research and GIScience: An Evaluation Using GIScience Conference Papers." In *11th International Conference on Geographic Information Science (GIScience 2021) - Part II*. Leibniz International Proceedings in Informatics (LIPIcs) Vol 208, 2, 1–2:16, edited by K. Janowicz and J. A. Verstegen. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Paez, A. (2022). "Reproducibility of Research during COVID-19: Examining the Case of Population Density and the Basic Reproductive Rate from the Perspective of Spatial Analysis." *Geographical Analysis* 54(4), 860–80. https://doi.org/10.1111/gean.12307

Rowe, F., A. Calafiore, D. Arribas-Bel, K. Samardziev, and M. Fleischmann. (2022). "Urban Exodus? Understanding Human Mobility in Britain during the COVID-19 Pandemic Using Facebook Data." https://doi.org/10.48550/ARXIV.2206.03272.

Saffary, T., O. A. Adegboye, E. Gayawan, F. Elfaki, M. A. Kuddus, and R. Saffary. (2020). "Analysis of COVID-19 Cases' Spatial Dependence in US Counties Reveals Health Inequalities." *Frontiers in public health* 8, 579190.

Schmidt, S. (2009). "Shall we Really Do it Again? The Powerful Concept of Replication Is Neglected in the Social Sciences." *Review of General Psychology* 813(2), 90–100.

Schmitt, R. R. (1978). "Threats to Validity Involving Geographic Space." *Socio-Economic Planning Sciences* 12(4), 191–5.

Simonsohn, U. (2015). "Small Telescopes: Detectability and the Evaluation of Replication Results." *Psychological science* 26(5), 559–69.

Singleton, A. D., S. Spielman, and C. Brunsdon. (2016). "Establishing a Framework for Open Geographic Information Science." *International Journal of Geographical Information Science* 30(8), 1507–21.

SPARC. (2019). "2019 Workshop on Replicability and Reproducibility in Geospatial Research at SPARC."

Stodden, V., F. Leisch, and R. D. Peng. (2014). *Implementing Reproducible Research*, Vol 546 2014. Boca Raton, FL: CRC Press.

Stodden, V., M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M. A. Heroux, J. P. A. Ioannidis, and M. Taufer. (2016). "Enhancing Reproducibility for Computational Methods." *Science* 354(6317), 1240–1.

Sugg, M. M., T. J. Spaulding, S. J. Lane, J. D. Runkle, S. R. Harden, A. Hege, and L. S. Iyer. (2021). "Mapping Community-Level Determinants of COVID-19 Transmission in Nursing Homes: A Multi-Scale Approach." *Science of the Total Environment* 752, 141946.

Sui, D., and P. Kedron. (2021). "Reproducibility and Replicability in the Context of the Contested Identities of Geography." *Annals of the American Association of Geographers* 111(5), 1275–83.

Sumner,, J., L. Haynes, S. Nathan, C. Hudson-Vitale, and L. McIntosh, (2020). "Reproducibility and Reporting Practices in COVID-19 Preprint Manuscripts." *medRxiv*. https://doi.org/10.1101/2020.03.24.20042796

Sun, F., S. A. Matthews, T. C. Yang, and M. H. Hu. (2020). "A Spatial Analysis of the COVID-19 Period Prevalence in US Counties through June 28, 2020: Where Geography Matters?" *Annals of Epidemiology* 52, 54–9.

Tullis, J. A., and B. Kar. (2021). "Where Is the Provenance? Ethical Replicability and Reproducibility in GIScience and its Critical Applications." *Annals of the American Association of Geographers* 111(5), 1318–28.

Verhagen, J., and E.-J. Wagenmakers. (2014). "Bayesian Tests to Quantify the Result of a Replication Attempt." *Journal of Experimental Psychology: General* 143(4), 1457.

Vijayan, T., M. Shin, P. C. Adamson, C. Harris, T. Seeman, K. C. Norris, and D. Goodman-Meza. (2020). "Beyond the 405 and the 5: Geographic Variations and Factors Associated with SARS-CoV-2 Positivity Rates in Los Angeles County." *Clinical Infectious Diseases* ciaa1692, e2970–75. https://doi.org/10.1093/cid/ciaa1692

Wainwright, J. (2021). "Is Critical Human Geography Research Replicable?" *Annals of the American Association of Geographers* 111(5), 1284–90. https://doi.org/10.1080/24694452.2020.1806025

Wali, B., and L. D. Frank. (2021). "Neighborhood-Level COVID-19 Hospitalizations and Mortality Relationships with Built Environment, Active and Sedentary Travel." *Health & Place* 71, 102659.

Wang, S. (2016). "CyberGIS and Spatial Data Science." *GeoJournal* 81(6), 965–8.

Wilson, J. P., K. Butler, S. Gao, Y. Hu, W. Li, and D. J. Wright. (2021). "A Five-Star Guide for Achieving Replicability and Reproducibility when Working with GIS Software and Algorithms." *Annals of the American Association of Geographers* 111(5), 1311–7.

Wolf, L. J., S. Fox, R. Harris, R. Johnston, K. Jones, D. Manley, E. Tranos, and W. W. Wang. (2021). "Quantitative Geography III: Future Challenges and Challenging Futures." *Progress in Human Geography* 45(3), 596–608.

Yeo-Teh, N. S. L., and B. L. Tang. (2020). "An Alarming Retraction Rate for Scientific Publications on Coronavirus Disease 2019 (COVID-19)." *Accountability in Research* 28(1), 47–53. https://doi.org/10.1080/08989621.2020.1782203

Yin, D., Y. Liu, H. Hu, J. Terstriep, X. Hong, A. Padmanabhan, and S. Wang. (2019). "CyberGIS-Jupyter for Reproducible and Scalable Geospatial Analytics." *Concurrency and Computation: Practice and Experience* 31(11), e5040.