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

Peter Kedron, Corresponding Author
School of Geographical Sciences and Urban Planning, Arizona State University
pkedron@asu.edu
ORCID: 0000-0002-1093-3416

Sarah Bardin
School of Geographical Sciences and Urban Planning, Arizona State University
sfbardin@asu.edu
ORCID: 0000-0001-8657-1725

Joseph Holler
Department of Geography, Middlebury College
josephh@middlebury.edu
ORCID: 0000-0002-2381-2699

Joshua Gilman
School of Life Sciences, Arizona State University
jcgilman@asu.edu
ORCID: 0000-0003-0468-2785

Bryant Grady
School of Geographical Sciences and Urban Planning, Arizona State University
bwgrady1@asu.edu

Megan Seeley
School of Geographical Sciences and Urban Planning, Arizona State University
mseeley1@asu.edu
ORCID: 0000-0003-1945-1162

Xin Wang
School of Sustainability, Arizona State University
xwang615@asu.edu

Wenxin Yang
School of Geographical Sciences and Urban Planning, Arizona State University
wyang80@asu.edu
ORCID: 0000-0002-9332-4500

Word count: 7000 words + 3 tables + 1 Figure

Submission date: April 18, 2022
Reviews Received: June, 28, 2022
Revision Submitted: September 09, 2022
Reviews Received: November 10, 2022
Revision Submitted: January, 13, 2023

Abstract

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 present three model reproductions of geographical analyses of COVID-19 that demonstrate how to achieve this goal. Each reproduction is based on a common, open access template and 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.

Keywords: Reproducibility, COVID-19, Spatial Regression, GIScience

1 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; Kedron et al., 2021c; Kedron et al., 2021b; Goodchild et al., 2021; Brunsdon and Comber, 2021). In principle, reproducible research publicly discloses the evidence base for claims from prior work not only to improve the transparency of scientific communication, but to also 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 (NASEM, 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 and have worked 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 (Kedron et al., 2021c; Tullis and Kar, 2021), offered guidelines for how to do so (Wilson et al., 2021; Hofer et al., 2019; Nüst and Pebesma, 2021), and created infrastructure to host researcher materials and recreate analyses (Wang, 2016; Yin et al., 2019; Nüst and Hinz, 2019). These activities set the stage for reproduction studies that assess the claims made in the existing geographic literature but do not themselves directly check the state of our knowledge.

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—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 the outputs of those attempts to those 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 in fact supports the claims 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 (see NASEM, 2019; Stodden, Leisch, and Peng, 2014; Stodden et al., 2016). Unfortunately, ending the evaluation of a study at an assessment of its computational reproducibility may even hinder scientific progress if others mistake the recreation of results as an affirmation of questionable decisions that led to those results.

We advocate that geographers move beyond checks of computational reproducibility and additionally begin to develop a body of reproduction studies focused on the assessment of the claims of prior work. To facilitate this transition, we make three principle contributions in this paper. 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 to this point has been under-discussed in the geographic literature.

The remainder of this paper 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 shift to 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 our three exemplar 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.

2 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; Kedron et al., 2021b; Brunsdon and Comber, 2021; Goodchild et al., 2021; Goodchild and Li, 2021; Kedron et al., 2021c). 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). During a reproduction study, the researchers conducting that study may focus on different goals. It is helpful to distinguish which of the two questions raised by the NASEM (2019) a researcher wishes to answer. First, if a reproduction study is focused on simply establishing whether the specific results of the original study can be recreated, or second, if the reproduction study is focused on whether the data, analysis, and results in fact support the conclusions and claims drawn from the original study.

When narrowly focused on 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 to recreate the results. Such reproductions 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 made.

When a researcher attempts to reproduce a study, they either have access to, or must attempt to identify, the decisions and materials 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 takes these opportunities, they may gain information about how the prior study was conceptualized, designed, and executed, which they can use to make qualified statements about whether the conclusions reached about relationships in the data 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 if errors were made during its execution (Nichols et al., 2021; Earp and Trafimow, 2015). Once identified, studies can be redesigned and errors can be corrected. In this way, reproduction studies help us progressively improve our understanding of phenomena by reducing the number of errors we make and lessening our uncertainty.

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

The reproductions that do exist in the geographic literature 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 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 in Europe have reviewed the computational reproducibility of 31 research paper results submitted to that association's annual conference for the past three years (Nüst et al., 2020; Nüst et al., 2021; Nüst et al., 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 coded analyses of submitted papers and share their findings in the form of 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 most difficult to computationally reproduce. While this research usefully summarizes technical barriers to computational reproducibility, all of these authors limit their discussion to coding errors and differences in figures and maps. Similarly, while these authors helpfully use their findings to derive guidelines for publishing computationally reproducible research, their central focus is on determining whether an independent researcher can re-execute a study's analytical code and create identical outputs.

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 stop a reproduction attempt when the code is found to fail or succeed at exactly recreating the original results and figures. In the geographical sciences, it is 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. If a study has a flawed research design, or is poorly executed, it may nonetheless be computationally reproducible. Even if 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 our 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., 2021b). 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 we think may affect the result of the original study. By introducing those changes we may 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. In a paper 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. (2021a) and Kedron et al. (2022a) 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.

3 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 own reproduction attempts. Here we introduce such an approach, and a workflow and template materials to facilitate its 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 sub-model and customized for different sub-disciplines 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.

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. Our template repository is available online as a Git repository under a BSD 3Clause License through (Kedron and Holler, Mar. 2022b). We used these materials to conduct the reproductions presented in this paper.

« Insert Figure 1 About Here »

3.1 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 that researchers clearly articulate the aspects of the prior study they intend to reproduce. For example, in the case of hypotheses-driven research the reproducing research 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 their hypotheses and provide only a partial description of their analytical plans. Researchers also often test a large number of hypotheses during the course of their study but highlight only an handful in their results. This situation leaves the reproducing researcher to choose which hypotheses to recreate and to explain why some hypotheses may have been chosen over others.

We suggest that researchers formally record and present their reproduction workflow as part of a pre-analysis plan that details the data collection, processing, and analysis they intend to undertake as part of their reproduction attempt. The workflow should be based on the most complete and precise understanding of procedures that is possible based upon reading the original publication and- ideally- on reading supplementary materials including data and code. In lieu of sufficient procedural detail, the plan should include the reproducing researchers' best approximation of the procedures for data processing and analysis. This pre-analysis plan should also include the criteria they intend to use to compare their results to those of the original study. Ideally, researchers will publicly register this plan before they begin their reproduction attempts. Pre-analysis plans can be shared via platforms like GitHub, or more formally registered through services like the Open Science Framework [1].

In the exemplar reproductions presented below, our student collaborators created the initial workflow models and drafted pre-analysis plans, which the lead authors reviewed and commented upon. We found iterative revisions of the plans to be vital for identifying issues of ambiguity, uncertainty, and error in the original research design. We suggest researchers catalog such issues as they arise during the reproduction attempt.

3.2 Implementation

Once pre-analysis planning is completed, most reproduction attempts will move on to data collection, preparation, and analysis. A common practical barrier is whether or not the data used in the original analysis is accessible, or whether the protocol and procedures used to gather that data is available. In some instances, if data are not available with the publication they can be accessed from an original source (e.g., US Census).

Data are almost always processed prior to analysis, and ideally the reproduction will use original code for data processing, analysis, and visualizing results. Whether the reproduction study can reuse code from the original study or needs to write new code, any changes or deviations from the original procedural plan should be documented. Whether available with the study or accessed from a common original source, data are almost always processed prior to analysis.

Pre-analysis plans are designed to be dynamic documents and communication tools for tracking unanticipated changes that occur throughout the reproduction process as researchers work to resolve the conceptual and practical challenges that arise during the reproduction process. As the pre-analysis plans are created and implemented as reproduction procedures or code, any ambiguities the reproducing researchers must resolve should be cataloged as issues that will be later categorized and evaluated following the Kedron et al. (2021b) framework. The decisions of reproducing researchers should also be included as amendments to the pre-analysis plan.

To move beyond computational reproducibility, researchers can introduce differences into their reanalyses to test the sensitivity of the original analysis to alternative conceptualizations or research designs. As these variations are introduced they should be tracked in the issues catalog along with the reasoning supporting each change. These differences can be categorized in the same way as researcher decisions, which will bring their evaluation into a common framework.

3.3 Evaluation

We suggest that researchers compare their reproduction results to those of the original study as they are created. If discrepancies arise early in the analysis (e.g., differences in descriptive statistics), we suggest revising the procedures and documenting the unplanned deviations from the original workflow in the pre-analysis plan before proceeding to subsequent analyses.

There is no universally agreed upon set of criteria to assess whether the results of an original study have been reproduced, and much of the literature related to the subject focuses on the more complicated question of replication (see Verhagen and Wagenmakers, 2014; OSC, 2015; Simonsohn, 2015; Lakens, 2017). Prior reproductions of geographic researcher have based their evaluation on either an exact match of numerical results, or the similarity of figures and maps. We suggest at minimum evaluating the direction, magnitude, and levels of uncertainty associated with both sets of results in any comparison. Differences between results

[1] See Christensen, Freese, and Miguel (2019) and Olken (2015) for a discussion of the pros and cons of pre-registration

will often motivate the introduction of further variations in research design and analysis. When those are pursued they should be cataloged and categorized like other issues and tracked as deviations from the original reproduction plan.

Critically, we view the documentation and comparison of the results of a reproduction attempt as only one part of the evaluation process. To use a reproduction to evaluate the claims made by the authors of the original study, we argue that the reproducing authors should also evaluate the complete set of decisions made during the creation of that result in light of our existing understanding of the of the phenomena under study. This task is where the cataloging and categorizing of issues and decisions that we advocate for throughout the reproduction process plays a role. That catalog can now be used to contextualize decisions made by the original and reproducing authors and the evidence of the reproduction results within the existing literature, providing a foundation for evaluating the research claims.

4 Empirical Context, and the Selection of Studies for Reproduction

4.1 Empirical Context

To demonstrate our approach and how it differs from computationally focused reproductions, we attempted to reproduce geographical analyses of COVID-19. The COVID-19 pandemic has highlighted the need to make 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; Sugg et al., 2021; Chakraborty, 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). Moreover, many of these geographical analyses have been undertaken by researchers with primary research interests and 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 are using this stream of research to revise their response to the pandemic. To ensure those groups have access to the best possible research so they can make the best possible decisions, we must know more than the results of recent geographical analyses of COVID-19. We must also know how reliable and credible those results are, because understanding the credibility of research allows us to appropriately weight findings when making decisions about pandemic response. Understanding the validity of these studies is also important because they are already becoming the foundation for future research.

Recognizing this situation, several authors (see Gustot, 2020; Sumner et al., 2020; Collins and Alexander, 2021) have emphasized how important it is that COVID-19 research be reproducible and have begun to catalog the availability of code and data within in the COVID-19 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 (see Ahasan et al., 2020; Agbehadji et al., 2020; Franch-Pardo et al., 2020; Franch-Pardo et al., 2021). To our knowledge, only one formal reproduction of geographical analyses of COVID-19 is presently available in the published literature (Kedron et al., 2021a). Conducting reproductions of COVID-19 research will allow us to assess the internal validity of selected studies and draw lessons about how we might use reproduction as a widely adopted means of geographic research assessment.

4.2 Selection of candidate studies for reproduction

4.2.1 Study selection

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 01, 2020 and March 15, 2021. The central objective of our search was to identify spatial analyses that demonstrate how reproduction can be used to critically appraise published work. 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 spatial methods frequently used in spatial epidemiology. These criteria allow our reproductions to inform our development of a practical framework and model for others to follow when reproducing empirical spatial analysis.

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 that match the objective of this review. The Scopus search was run February 09, 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 then collected the abstracts and full texts for each of these articles.

Article abstracts were selected for further review if they were: 1) published in an English language journal, 2) peer-reviewed, 3) topically focused on the COVID-19 pandemic, and 4) geographically focused on the United States. These criteria narrowed the selection of articles to 60. Articles were then reviewed for a fifth criteria: application of spatial statistical methods common in spatial epidemiology and compatible with graduate student training in spatial statistics (e.g., spatial regression and pattern analysis). This review narrowed our list to 15 candidate articles that were further reviewed for their 1) complete publication details, 2) study objectives, 3) data sources, 4) data and code availability, and 5) spatial methodology.

Based on this information, we selected three articles – Mollalo, Vahedi, and Rivera (2020), Saffary et al. (2020), and Vijayan et al. (2020). These articles used spatial statistical methods common in both spatial epidemiology and the broader geographic literature and appeared feasibile to reproduce. These articles also provided a level of detail about the research process typical of the broader literature - information about data sources, hypotheses, and methodology was provided. As such these articles provided us with the opportunity to demonstrate the use of reproduction as a form of reanalysis and critique that extends beyond the matching of computational outputs.

4.2.2 Characteristics of the studies selected for reproduction

The three studies selected for reproduction use spatial regression techniques and local spatial statistics to make associational inferences about COVID-19 (Table 2). Mollalo, Vahedi, and Rivera (2020) fit 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 healthcare 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 COVID-19 crude positivity rates and a series of predictor variables. Although not publicly available, we were able to obtain the original study data after contacting the authors. The analysis code was not made available, nor was information about the computational environment used.

« Insert Table 1 About Here »

5 Implementation of the Reproduction Attempts

We followed the three-stage process of planning, implementation, and evaluation outlined by our model approach (Fig. 1) for each of the three reproductions we attempted. 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 **removed for anonymous peer-review**.

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 team member 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, as we developed a single common workflow model. Those workflows then became the foundation of our pre-analysis plans, which also identified the key hypotheses we sought to re-test and any deviations from the original analyses we anticipated due to a lack of information provided in the original study.

We used the authors' provided materials to the greatest extent possible as we implemented our reproduction attempts. When available, we used publicly available data provided by authors. When not available, we acquired the public data described in the article or contacted the corresponding author to request inaccessible data. We attempted to use the processing environment described in the original study but also translated the workflow into R code version 4.0.4 (R Core Team, 2021). Whenever possible, we followed the procedures described in the original article or provided in the form of code. When we encountered missing or uncertain information with regard to data sources or procedures, we attempted the reproduction using alternative data sources and procedural decisions.

Following our model approach, our evaluation of each reproduction attempt consisted of two assessments: (1) an assessment of result similarity and computational reproducibility, and (2) an evaluation of the execution and claims of the original study. Mirroring existing reproductions of geographic research that focus on verification and recreating results with the same data and code, we used the simple criteria of whether the results of the reproduction and original analysis were numerically or graphically identical to assess reproducibility. We applied this criteria to analyses that have analytical solutions (e.g., OLS regression or local statistics evaluated using the normality assumption). However, several of the analyses we attempted to reproduce relied on conditional permutation of the data to estimate parameters (e.g., direct/indirect effect estimation) or made statistical inferences (e.g., local statistical inferences). In these instances, we relaxed the criteria of identical reproduction and instead focused our evaluation on the comparison of parameter estimates, related uncertainty estimates, and statistical significance. (p-values)[2]. These criteria mirror those

[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.

presented by the NASEM (2019) and those used by OSC (2015). We were able to partially verify the original findings of each of the three studies by computationally reproducing a portion of the original results, albeit not without challenges (Section 6.1).

Our deep reading, attempts to recreate the workflows of the original studies, and efforts to achieve computational reproduction revealed numerous questions and concerns about the the design and implementation of each study. Those questions motivated our further investigation of the internal validity 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. (2021b) 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 plan to include a record of any unplanned deviations from the posted plan, the results of the reproduction including comparison to the original study, and a discussion of the results. All team members reviewed each report, and after revisions, final reports were posted within each reproduction compendium.

6 Lessons from the Three Reproductions

6.1 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 study was directly related to the availability of original data and the detail of the procedural description provided, as none of the authors shared their original code. 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. This gap 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 the identical neighborhood-level COVID data. Once the authors provided these data upon request, we were able to create an exact reproduction of their descriptive statistics and obtain consistent spatial regression estimates.

To achieve a partial reproduction of each study, we had to make unplanned deviations from our initial analytical plans. For example, while Saffary et al. (2020) published their analytic data file, 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. Our reproductions of the authors' LISA analysis found low-high and high-low clusters that were either not identified or 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.

« Insert Table 2 About Here »

6.2 Beyond Computational Reproductions

Our attempts to reproduce the computational result of the three selected studies uncovered procedural, statistical, and inferential questions that raised concerns about the internal validity and credibility 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.

« Insert Table 3 About Here »

6.2.1 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 there was limited and geographically variable testing capacity 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. This issue constrains the ability of our reproductions to validate these studies. While reproducing this work 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 that 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 socio-demographic profiles of these units, the pandemic also created migration patterns (Coven, Gupta, and Yao, 2020; Haslag and Weagley, 2021) 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 we 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 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.

6.2.2 Measurement and Processing

We invested substantial labor in reconstructing the original data and data processing procedures for the three reproduction studies. In the process, we discovered concerns related to construct validity and variable construction. For example, the article by Saffary et al. (2020) contains ambiguities and inconsistencies with regards to the handling of missing data and data standardization. As one example, the authors provide no information about their handling of missing county level data for their primary care physician variable. We investigated three alternative procedures to address this issue - filtering, zero imputation, and mean imputation. Our findings suggest the authors simply omitted missing values. In other instances, Saffary et al. (2020) chose not to standardize their variables by population size. For example, the authors analyzed the raw count of intensive care unit beds in each county. However, the strong positive correlation between the number of such beds and county population likely makes this essentially a measure of county population.

In other instances, although Vijayan et al. (2020) indicate the standardization of variables prior to spatial regression modeling, they are not clear as to whether this standardization was applied to both the response and predictor variables. Moreover, the authors provide limited information about the specification of their spatial regressions, which makes it difficult to interpret their results. Nonetheless, 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 the original 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 10km 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 non-representative measurements. For example, 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.

6.2.3 Analysis and Inference

Our reproduction attempts allowed us to draw several cautionary lessons about the implementation and interpretation of spatial statistical tests of COVID 19. First, the authors of all three studies did not clearly present the complete discrete set of hypotheses they tested prior to their analyses. Each study 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, our 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 their assumptions about individual-level associations and 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. We suggest that this type of epidemiological study that is based on aggregate social data should be interpreted with caution as exploratory and should be supported by further individual-level or multi-scale research.

Additionally, 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, Vijayan et al. (2020) use of a single 10km hexagonal grid as the spatial support for their analysis provided. Selection of another grid size may produce different estimated associations between predictor variables and COVID-19 rates. The results of this analysis may be particularly sensitive to variation across spatial supports given that the centroid overlap-based aggregation of data will produce different levels of measurement error for each hexagonal grid size. We suggest that 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, our 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 4 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 the authors reliance on fit statistics to guide model selection and to measure explanatory power. Based solely on the higher R-squared 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-squared from both the GWR and MGWR models which show large numbers of counties with negative R-squared 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. We propose a need to balance data-driven exploratory analyses with more deductive theory-based approaches to spatial epidemiological modelling with registered pre-analysis plans in order to develop theorized mechanisms with inferential power.

Similarly, our 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 we attempted 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. After applying these adjustments nearly all of the spatial patterns highlighted in the manuscript disappear. While we were able to reproduce the authors' FDR-adjusted results, our attempt to reproduce the Bonferroni-adjusted results failed.

6.2.4 Communication

Our 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; the decisions and justifications for their analytical decisions; and the foundations of their conclusions. However, 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. That process of reanalysis led us to the results presented in the prior sections and reinforces the value of reproduction as a tool to check the internal validity and credibility of 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 we 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). While we could not directly observe selective inference in these studies, our 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. The potential for selective inference is also exacerbated by the absence of a clear and complete set of research hypotheses in each study.

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. However, the important point is that while we do not know what the authors did, 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 pre-register 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.

7 Conclusion

In this paper, we present a model workflow and corresponding materials to help geographic researchers move beyond using reproduction to simply answer whether the 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 paper, 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. (2021b)'s representation of the research process as a series of decisions researchers make in the face of uncertainty about the phenomenon under study. 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. In doing so, our approach provides a means of linking the existing literature on challenges and uncertainties in geographical analyses to aspects of the reproduction process. This approach 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 an code. By retracing each study's procedure, we scrutinize every detail of 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. All identified errors and uncertainties in each study 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., Apr. 2022c, Apr. 2022d, Apr. 2022b). 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 a series of such threats to conclusion and internal validity involving geographic space present in the three studies we attempt to reproduce and connect those threats to decision points in the research process. The concerns highlighted in this paper 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 paper is one of the first attempts to push reproduction attempts beyond computation in the geographical sciences.

Despite the concerns 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 value incorporating reproductions into that process might bring. 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 paper.

The discussion and practice of reproducibility in geography should not be limited to matters of sharing research artifacts and re-computing 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 are necessary 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 endeavour that is currently not incentivized by academic review process. As such it seems likely that many academics do not conduct formal reproduction, 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.

CRediT Author Statement

Kedron led study conceptualization, methodology, writing; and supervision and administration of reproductions. **Bardin** performed reproduction attempts and contributed substantially to reproduction review and all writing tasks. She also led data curation and software development. **Holler** contributed to methodology development, writing - review and editing. **Gilman**, **Grady**, **Seeley**, **X. Wang**, **W. 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.

Funding

This material 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**).

References

Agbehadji, Israel Edem et al. (2020). "Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing". In: *International journal of environmental research and public health* 17.15, 5330.

Ahasan, Rakibul et al. (2020). "Applications of GIS and geospatial analyses in COVID-19 research: a systematic review". In: *F1000Research* 9.1379, 1379.

Andersen, Lauren M et al. (2021). "Analyzing the spatial determinants of local Covid-19 transmission in the United States". In: *Science of the Total Environment* 754, 142396.

Benjamini, Yoav, Ruth Heller, and Daniel Yekutieli (2009). "Selective inference in complex research". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1906, 4255–4271.

Brunsdon, Chris (2016). "Quantitative methods I: Reproducible research and quantitative geography". In: *Progress in Human Geography* 40.5, 687–696.

Brunsdon, Chris and Alexis Comber (2021). "Opening practice: supporting reproducibility and critical spatial data science". In: *Journal of Geographical Systems* 23.4, 477–496.

Brunsdon, Chris and Alex Singleton (2015). *Geocomputation: a practical primer*. Sage, 2015.

Chakraborty, Jayajit (2021). "Social inequities in the distribution of COVID-19: An intra-categorical analysis of people with disabilities in the US". In: *Disability and health journal* 14.1, 101007.

Christensen, Garret, Jeremy Freese, and Edward Miguel (2019). *Transparent and reproducible social science research*. University of California Press, 2019.

Clemens, Michael A (2017). "The meaning of failed replications: A review and proposal". In: *Journal of Economic Surveys* 31.1, 326–342.

Collins, Annie and Rohan Alexander (2021). "Reproducibility of COVID-19 pre-prints". In: arXiv: `2107.10724 [stat.AP]`.

Coven, Joshua, Arpit Gupta, and Iris Yao (2020). "Urban flight seeded the covid-19 pandemic across the united states". In: *Available at SSRN 3711737*.

Davies, Wayne KD (1968). "The need for replication in human geography: some central place examples". In: *Tijdschrift voor economische en sociale geografie* 59.3, 145–155.

Earp, Brian D and David Trafimow (2015). "Replication, falsification, and the crisis of confidence in social psychology". In: *Frontiers in psychology* 6, 621.

Franch-Pardo, Ivan et al. (2020). "Spatial analysis and GIS in the study of COVID-19. A review". In: *Science of The Total Environment* 739, 140033.

Franch-Pardo, Ivan et al. (2021). "A review of GIS methodologies to analyze the dynamics of COVID-19 in the second half of 2020". In: *Transactions in GIS* 25.5, 2191–2239.

Gelman, Andrew and Hal Stern (2006). "The difference between "significant" and "not significant" is not itself statistically significant". In: *The American Statistician* 60.4, 328–331.

González-Leonardo, Miguel et al. (2022). *Understanding Patterns of Internal Migration During the COVID-19 Pandemic in Spain*. 2022. `https://doi.org/10.31235/osf.io/svtw7`.

Goodchild, Michael F and Wenwen Li (2021). "Replication across space and time must be weak in the social and environmental sciences". In: *Proceedings of the National Academy of Sciences* 118.35.

Goodchild, Michael F et al. (2021). "Introduction: Forum on reproducibility and replicability in geography". In: *Annals of the American Association of Geographers* 111.5, 1271–1274.

Gustot, Thierry (2020). "Quality and reproducibility during the COVID-19 pandemic". In: *JHEP Reports* 2.4.

Halpern, Dylan et al. (2021). "Dimensions of uncertainty: a spatiotemporal review of five COVID-19 datasets". In: *Cartography and Geographic Information Science*, 1–22.

Haslag, Peter H and Daniel Weagley (2021). "From LA to Boise: How migration has changed during the COVID-19 pandemic". In: *Available at SSRN 3808326*.

Hofer, Barbara et al. (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, Peter and Joseph Holler (2022a). "Replication and the search for the laws in the geographic sciences". In: *Annals of GIS* 28.1, 45–56.

— (Mar. 2022b). *Template for Reproducible and Replicable Research in Human-Environment and Geographical Sciences*. Mar. 2022. https://doi.org/10.17605/OSF.IO/W29MQ.

Kedron, Peter et al. (2021a). "A Replication of Dimaggio Et Al. (2020) in Phoenix, AZ". In: *SSRN Electronic Journal*, 0–2. ISSN: 1556-5068. https://doi.org/10.2139/ssrn.3995296.

Kedron, Peter et al. (2021b). "Reproducibility and replicability in geographical analysis". In: *Geographical Analysis* 53.1, 135–147.

Kedron, Peter et al. (2021c). "Reproducibility and replicability: opportunities and challenges for geospatial research". In: *International Journal of Geographical Information Science* 35.3, 427–445.

Kedron, Peter et al. (2021d). "Using the specification curve to teach spatial data analysis and explore geographic uncertainties". In: *Journal of Geography in Higher Education*, 1–11.

Kedron, Peter et al. (2022a). *Reproducibility, Replicability, and Open Science Practices in the Geographical Sciences*. 2022. https://doi.org/10.17605/OSF.IO/C5A2R.

Kedron, Peter et al. (Apr. 2022b). *Reproduction of Mollalo et al 2020*. Apr. 2022. https://doi.org/10.17605/OSF.IO/e43kq.

— (Apr. 2022c). *Reproduction of Saffary et al 2020*. Apr. 2022. https://doi.org/10.17605/OSF.IO/qfkg4.

— (Apr. 2022d). *Reproduction of Vijayan et al 2020*. Apr. 2022. https://doi.org/10.17605/OSF.IO/my5dz.

Kmoch, Alexander, Daniel Nust, and Evalyn Uuemma (2020). "Reproducible submission workflow". In: *Proceedings of the 5th AGILE PhD School 2019*. 2020, 1–10.

Konkol, Markus, Christian Kray, and Max Pfeiffer (2019). "Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study". In: *International Journal of Geographical Information Science* 33.2, 408–429.

Konkol, Markus, Daniel Nüst, and Laura Goulier (2020). "Publishing computational research-a review of infrastructures for reproducible and transparent scholarly communication". In: *Research integrity and peer review* 5.1, 1–8.

Lakens, Daniël (2017). "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses". In: *Social psychological and personality science* 8.4, 355–362.

Lee, Jieun and Ivan J. Ramírez (2022). "Geography of Disparity: Connecting COVID-19 Vulnerability and Social Determinants of Health in Colorado". In: *Behavioral Medicine* 0.0, 1–13. ISSN: 0896-4289. https://doi.org/10.1080/08964289.2021.2021382.

Mollalo, Abolfazl, Behzad Vahedi, and Kiara M Rivera (2020). "GIS-based spatial modeling of COVID-19 incidence rate in the continental United States". In: *Science of the total environment* 728, 138884.

Muenchow, Jannes, Susann Schäfer, and Eric Krüger (2019). "Reviewing qualitative GIS research—Toward a wider usage of open-source GIS and reproducible research practices". In: *Geography Compass* 13.6, e12441.

NASEM (2019). *Reproducibility and Replicability in Science*. National Academies Press, 2019.

Nichols, James D et al. (2021). "Opinion: A better approach for dealing with reproducibility and replicability in science". In: *Proceedings of the National Academy of Sciences* 118.7.

Nosek, Brian A, Jeffrey R Spies, and Matt Motyl (2012). "Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability". In: *Perspectives on Psychological Science* 7.6, 615–631.

Nüst, Daniel and Matthias Hinz (2019). “containerit: Generating Dockerfiles for reproducible research with R”. In: *Journal of Open Source Software* 4.40, 1603.

Nüst, Daniel and Edzer Pebesma (2021). “Practical reproducibility in geography and geosciences”. In: *Annals of the American Association of Geographers* 111.5, 1300–1310.

Nüst, Daniel et al. (2018). “Reproducible research and GIScience: an evaluation using AGILE conference papers”. In: *PeerJ* 6, e5072.

Nüst, Daniel et al. (2020). *Reproducibility Reviews AGILE 2020*. 2020. https://doi.org/10.17605/OSF.IO/6K5FH.

Nüst, Daniel et al. (2021). *Reproducible Research at AGILE*. 2021.

Nüst, Daniel et al. (2022). *Reproducibility Reviews AGILE 2022*. 2022.

Olken, Benjamin A (2015). “Promises and perils of pre-analysis plans”. In: *Journal of Economic Perspectives* 29.3, 61–80.

OSC (2015). “Estimating the reproducibility of psychological science”. In: *Science* 349.6251.

Ostermann, Frank O and Carlos Granell (2017). “Advancing science with VGI: Reproducibility and replicability of recent studies using VGI”. In: *Transactions in GIS* 21.2, 224–237.

Ostermann, Frank O. et al. (2021). “Reproducible Research and GIScience: An Evaluation Using GIScience Conference Papers”. In: *11th International Conference on Geographic Information Science (GIScience 2021) - Part II*. Ed. by Krzysztof Janowicz and Judith A. Verstegen. Vol. 208. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, 2:1–2:16. ISBN: 978-3-95977-208-2. https://doi.org/10.4230/LIPIcs.GIScience.2021.II.2.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021.

Rowe, Francisco et al. (2022). *Urban Exodus? Understanding Human Mobility in Britain During the COVID-19 Pandemic Using Facebook Data*. 2022. https://doi.org/10.48550/ARXIV.2206.03272.

Saffary, T et al. (2020). “Analysis of COVID-19 cases’ spatial dependence in US Counties reveals health inequalities”. In: *Frontiers in public health* 8.

Schmidt, S (2009). “Shall we really do it again? The powerful concept of replication is neglected in the social sciences”. In: *Review of General Psychology* 813 (2), 90–100.

Schmitt, Rolf R (1978). “Threats to validity involving geographic space”. In: *Socio-Economic Planning Sciences* 12.4, 191–195.

Simonsohn, Uri (2015). “Small telescopes: Detectability and the evaluation of replication results”. In: *Psychological science* 26.5, 559–569.

Singleton, Alex David, Seth Spielman, and Chris Brunsdon (2016). “Establishing a framework for Open Geographic Information science”. In: *International Journal of Geographical Information Science* 30.8, 1507–1521.

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

Stodden, Victoria, Friedrich Leisch, and Roger D Peng (2014). *Implementing reproducible research*. Vol. 546. CRC Press Boca Raton, FL, 2014.

Stodden, Victoria et al. (2016). “Enhancing reproducibility for computational methods”. In: *Science* 354.6317, 1240–1241.

Sugg, Margaret M et al. (2021). “Mapping community-level determinants of COVID-19 transmission in nursing homes: A multi-scale approach”. In: *Science of The Total Environment* 752, 141946.

Sui, Daniel and Peter Kedron (2021). “Reproducibility and Replicability in the Context of the Contested Identities of Geography”. In: *Annals of the American Association of Geographers* 111.5, 1275–1283.

Sumner, Josh et al. (2020). "Reproducibility and reporting practices in COVID-19 preprint manuscripts". In: *medRxiv*. https://doi.org/10.1101/2020.03.24.20042796. eprint: https://www.medrxiv.org/content/early/2020/03/27/2020.03.24.20042796.full.pdf.

Sun, Feinuo et al. (2020). "A spatial analysis of the COVID-19 period prevalence in US counties through June 28, 2020: where geography matters?" In: *Annals of Epidemiology* 52, 54–59.

Tullis, Jason A and Bandana Kar (2021). "Where is the provenance? Ethical replicability and reproducibility in GIScience and its critical applications". In: *Annals of the American Association of Geographers* 111.5, 1318–1328.

Verhagen, Josine and Eric-Jan Wagenmakers (2014). "Bayesian tests to quantify the result of a replication attempt." In: *Journal of Experimental Psychology: General* 143.4, 1457.

Vijayan, Tara et al. (2020). "Beyond the 405 and the 5: Geographic variations and factors associated with SARS-CoV-2 positivity rates in Los Angeles County." In: *Clinical Infectious Diseases: an Official Publication of the Infectious Diseases Society of America*.

Wainwright, Joel (2020). "Is critical human geography research replicable?" In: *Annals of the American Association of Geographers*, 1–7.

Wali, Behram and Lawrence D Frank (2021). "Neighborhood-level COVID-19 hospitalizations and mortality relationships with built environment, active and sedentary travel". In: *Health & Place* 71, 102659.

Wang, Shaowen (2016). "CyberGIS and spatial data science". In: *GeoJournal* 81.6, 965–968.

Wilson, John P et al. (2021). "A five-star guide for achieving replicability and reproducibility when working with GIS software and algorithms". In: *Annals of the American Association of Geographers* 111.5, 1311–1317.

Wolf, Levi John et al. (2021). "Quantitative geography III: Future challenges and challenging futures". In: *Progress in Human Geography* 45.3, 596–608.

Yeo-Teh, Nicole Shu Ling and Bor Luen Tang (2020). "An alarming retraction rate for scientific publications on Coronavirus Disease 2019 (COVID-19)". In: *Accountability in Research* 28.1, 47–53. ISSN: 15455815. https://doi.org/10.1080/08989621.2020.1782203.

Yin, Dandong et al. (2019). "CyberGIS-Jupyter for reproducible and scalable geospatial analytics". In: *Concurrency and Computation: Practice and Experience* 31.11, e5040.

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	USA	USA	LA County
Spatial Support	County	County	10km Hexagons
Temporal Extent	Jan-Apr 2020	Feb-May 2020	Feb-June 2020
Hypothesis Tests	1000s	1000s	1000s
Methods	SEM, SLM, GWR, MGWR	Moran's I, Bivariate Moran's I	Moran's I, SLM

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

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

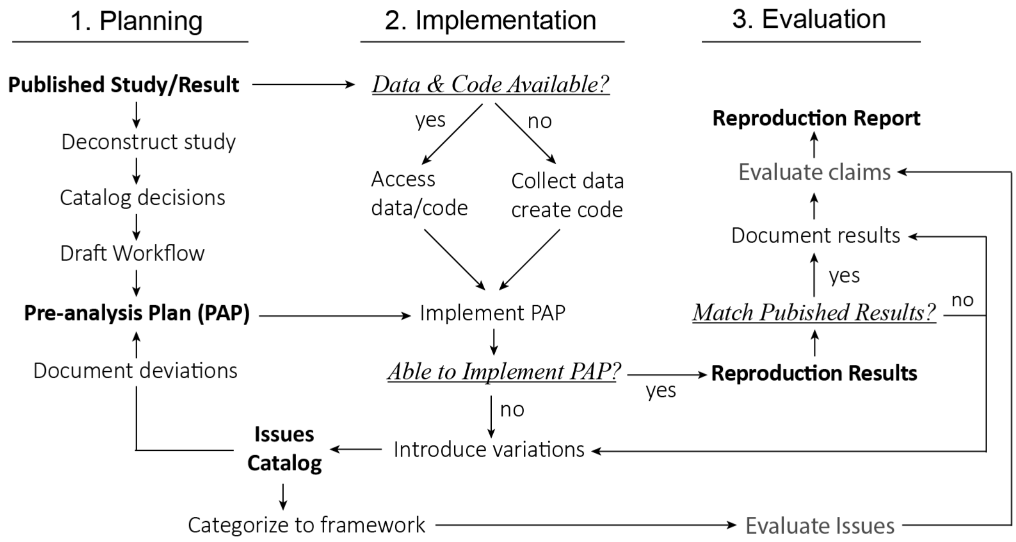


Figure 1: Three stage approach to reproduction

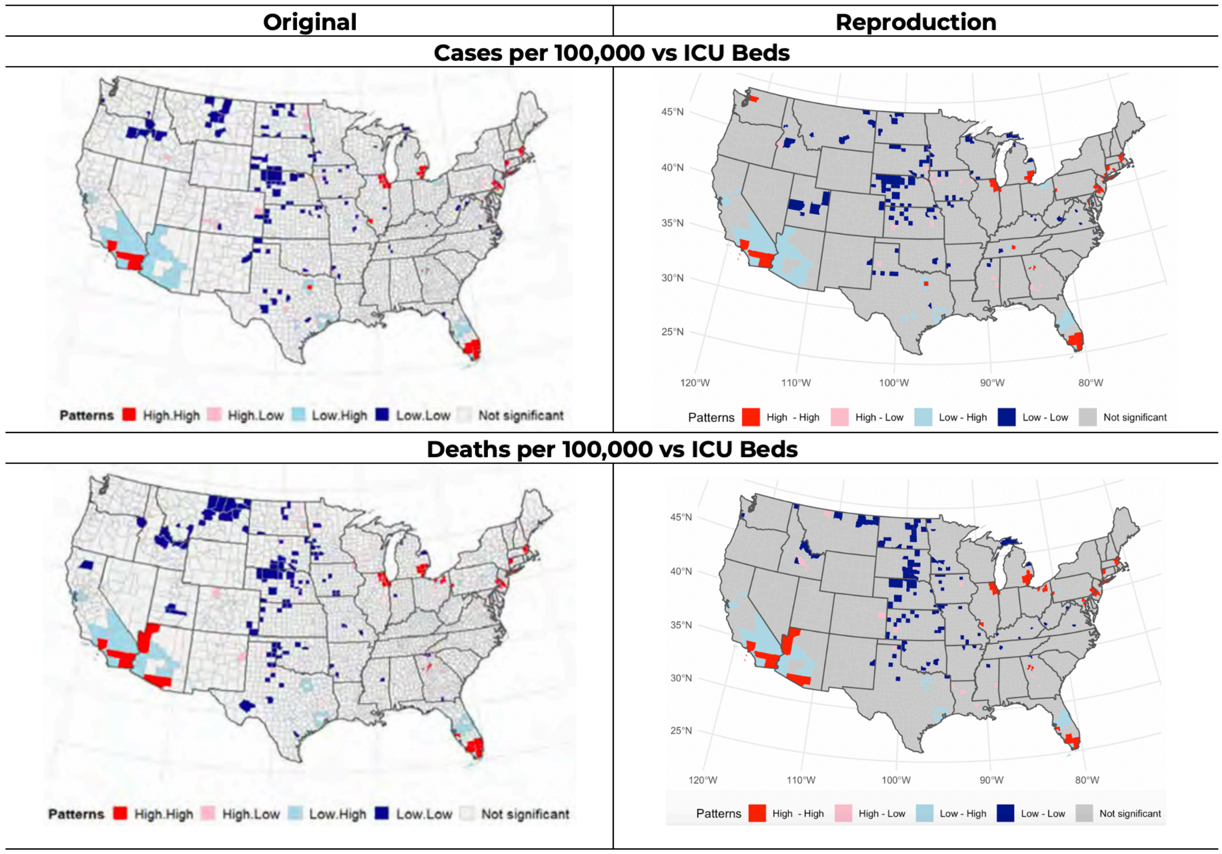


Figure 2: Results from bivariate local Moran's I analysis of number of intensive care unit beds and rate of COVID-19 cases (top) and rate of COVID-19 deaths (bottom) from Saffary et al. (left) and reproduction analysis (right). High-high clusters are denoted in red, high-low clusters are denoted in pink, low-high clusters are denoted in light blue, low-low clusters are denoted in dark blue, and non-significant clusters are denoted in grey.

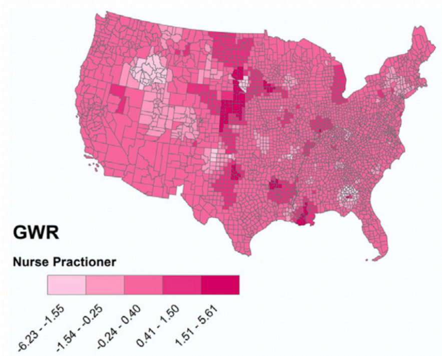
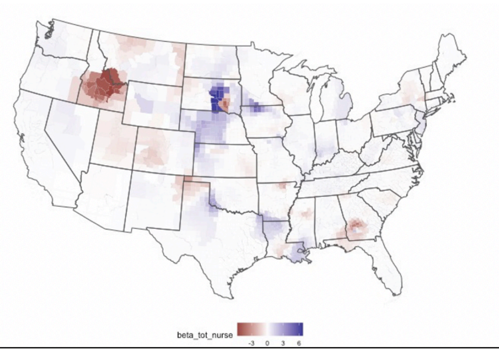
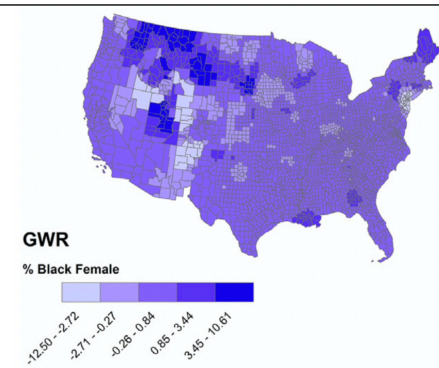
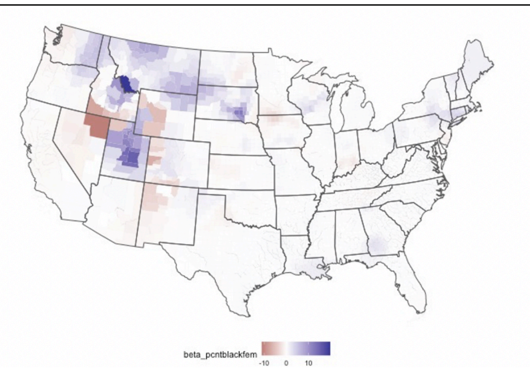
Original	Reproduction
Nurse Practitioner	
	
Percent Black Female	
	

Figure 3: Parameter estimates from Geographically Weighted Regression analysis for the number of nurse practitioners (top) and percent Black females (bottom) from Mollalo et al. (left) and reproduction analysis (right).

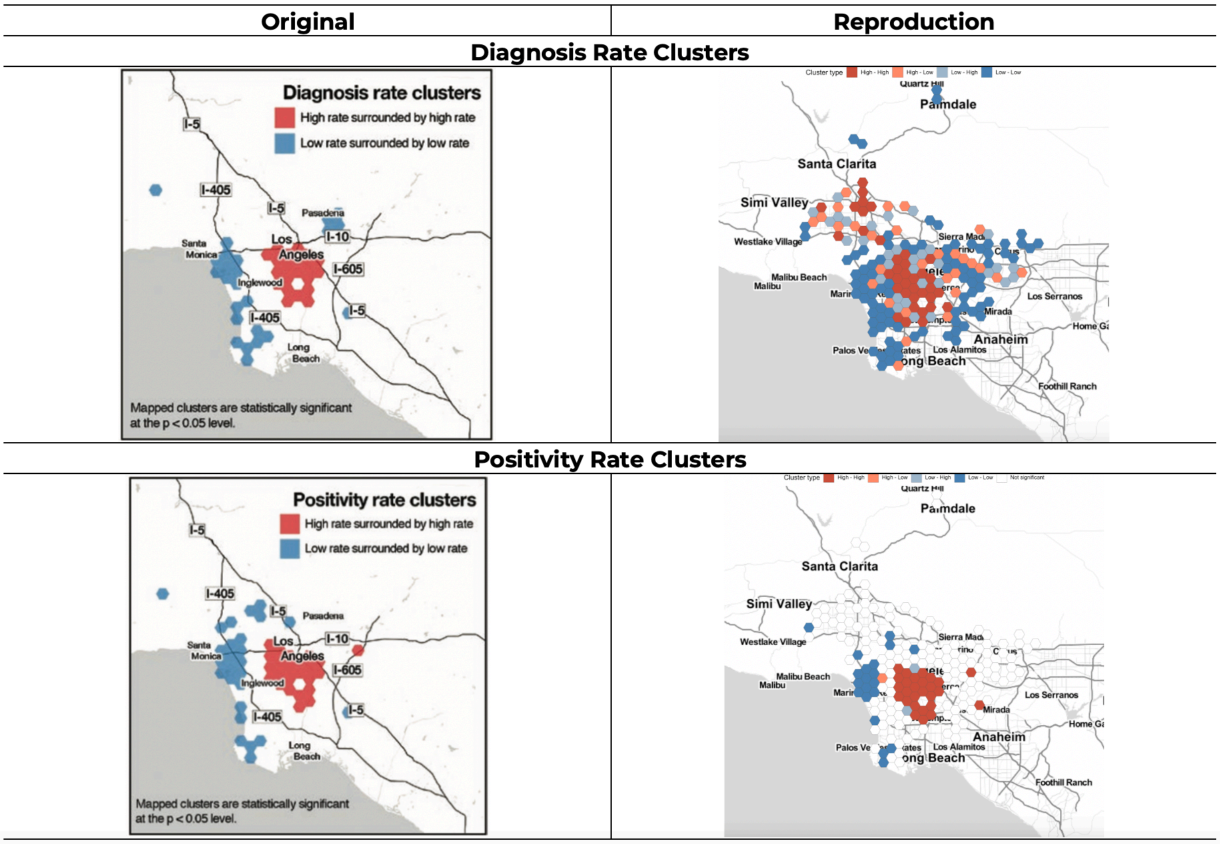


Figure 4: Results from LISA analysis of COVID-19 diagnosis rates (top) and positivity rates (bottom) from Vijayan et al. (left) and reproduction analysis (right). High-high clusters are denoted in red, low-low clusters are denoted in blue, non-significant clusters denoted in white.