Decentralized Infrastructure for Reproducible and Replicable Geographical Science

Joseph Holler
Department of Geography
Middlebury College
Middlebury, VT, USA
josephh@middlebury.edu

Peter Kedron
Department of Geography
University of California Santa Barbara
Santa Barbara, CA, USA
peterkedron@ucsb.edu

***Abstract*—The I-GUIDE cyberinfrastructure project for convergence science is a leading example of the possibilities the geospatial data revolution holds for scientific discovery. However, rapidly expanding access to increasingly complex data sources and methods of computational analysis also presents a challenge to the research community. With more data and more potential analyses, researchers face the possibility of jeopardizing the inferential power of convergence research with selection bias. Well-designed infrastructure that can flexibly guide researchers as they record and track decisions in their research designs opens a path to mitigating this problem, while also expanding the reproducibility and replicability of research. Much of the infrastructure needed for convergence research can be borrowed and adapted from other disciplines, but geographic convergence research confronts at least five novel challenges. These are the need for geographically-explicit project metadata, managing diverse and complex data inputs, handling restricted data, specifying and reproducing computational environments, and disclosing researcher decisions and threats to validity that are unique to geographic research. We introduce a template research compendium and analysis plan for study preregistration to address these novel challenges.**

Keywords—Reproducibility, Open Science, Geospatial, Selective Inference, Compendium, Metadata, Preregistration

I. Introduction

Over the past 30 years, we have witnessed a transformation in scientific research practices as researchers in many fields have come to rely on a wide range of inexpensive computational tools to create, structure, and analyze datasets too large to be studied using traditional techniques. Researchers use these tools as they work in teams and search for solutions to pressing societal problems that are too complex to address using the resources and frameworks available in a single discipline. This combination of interdisciplinary integration driven to address a specific and compelling societal problem is the hallmark of the convergence research approach at the center of the Integrated Discovery Environment (I-GUIDE).

Built on the backbone of JupyterHub, CyberGIS-Compute, and GeoEDF, the I-GUIDE platform provides the flexible infrastructure needed to pursue convergence research projects. However, the computational power and flexibility of I-GUIDE can also be a double-edged sword. On one edge, advances in computational power and statistical science have made it possible for researchers to select among numerous permutations of a rich variety of models. On the other, the I-GUIDE platform may reduce the barriers to interdisciplinary exploration to such an extent that researchers may stumble into unstructured exploratory analysis and the inadvertent selection of outcomes that fail to meet the scientific standards required to support societal needs. Therefore, a central challenge of developing I-GUIDE—and delivering on the promises of the convergence research it supports—is guiding researchers toward practices that minimize the potential for unobserved selection while maximizing the examination of justifiable and meaningful analytical pathways and solutions. While other disciplines have encountered and addressed many facets of this challenge, the geographic research community has yet to address facets of the problem that are uniquely pronounced in geographic and convergent research.

In the remainder of this paper, we present a research template designed to help users of the I-GUIDE platform avoid selective inference and disclose the solution space examined during any particular analysis. Our work is organized into three sections. First, we briefly present the issue of selective inference and discuss how and why it might manifest in an I-GUIDE convergence research project. Second, we differentiate between centralized and decentralized solutions to selective inference and irreproducibility to frame our own resource development and approach. Finally, we catalog a series of uniquely geographic challenges we encountered while interacting with the I-GUIDE platform and attempting reproduction and replication studies. We present the decentralized solutions we developed to overcome these challenges and discuss how others might adapt our solutions in their own future work using the platform.

II. Selective Inference and Convergence Research

Selective inference refers to the practice of focusing on a subset of findings that were identified as interesting only after viewing the available data and alternative analytical paths. Benjamini [1] has linked selective inference to the increasingly computational nature of scientific research and called the practice the silent killer of replicability. The argument, which

has been mirrored by many others [2-4], proceeds as follows. Pursuing complex questions, researchers have automated selected portions of the research process and created a situation in which it is common to have extensive flexibility when designing and executing a project. These 'researcher degrees of freedom' [5] can be used in a number of ways. In the best cases, researchers use this freedom to consider alternative research designs and analytical pathways, communicate the thinking and motivations behind those alternatives, and even combine the plausible alternative forms of evidence generated throughout a project. In less ideal cases, researchers might use this freedom to select and report the results of a single 'best' analysis even when many alternatives were tested through practices known as p-hacking, HARKing, data dredging, etc. This selective reporting is problematic because it removes potentially informative results from view and masks the uncertainty surrounding the results independent research can see. This omission creates a systematic distortion in the available evidence base that makes it difficult to assess the credibility of individual claims and the existence of corroborating evidence.

The prevalence of selective inference in convergence research is unknown, but there is no reason to believe this type of research is immune to this issue. The explicit focus of convergence research on intentionally integrating diverse intellectual traditions is likely to expand the decision space considered and the number of analytical pathways explored within any particular project. As experts from different disciplines pursue a common challenge, they ask questions in new ways and make novel connections that take the research process in directions it would not otherwise follow. Indeed, the National Science Foundation characterizes successful convergence projects as those that demonstrate a co-development of integrated tools, techniques, and solutions that are often new to science. To deliver on their promise of actionable solutions, convergence researchers must recognize and communicate how those solutions were developed, which analytical pathways they explored when creating them, and why those paths were selected over others. Absent the rigorous tracking and reporting of this information, convergence research will likely be subject to the same pitfalls that have troubled scientific research–-irreproducibility, non-extensibility, and unclear rates of false positive findings.

One promising pathway to this type of replicable and extensible form of inquiry is the further development and integration of veridical data science practices into convergence projects. Initially proposed by Yu and Kumbier [6] and elaborated spatially by Kedron and Bardin [7], veridical data science seeks to extract reproducible information from data using an enriched technical language to communicate and evaluate empirical evidence in the context of human decisions, further supported by the purposeful replication of findings across locations and times. This approach not only rigorously records research decisions, but also simultaneously evaluates and tracks alternative solution sets by following alternative analytical paths to their endpoints and communicating the distribution of alternative outcomes. This approach not only makes apparent the range of potential research outcomes, but also facilitates the evaluation of those outcomes in light of their supporting decisions. A convergence research team may then further establish the applicability and scalability of their solutions by replicating across space and time.

III. Centralized and Decentralized Infrastructure

Identifying selective inference as a key contributor to the replication crisis and a threat to convergence research may be surprising, as efforts across the sciences to standardize project reporting, increase research transparency, and facilitate the sharing of research components are ultimately designed to address this problem [1]. However, centralized responses that involve funding agencies, publishers, and cross-cutting institutions provide only generalized resources or incentives to address selection. These resources must be supported by decentralized solutions that help researchers adapt these advances to the needs individual convergence research projects.

In our interpretation, centralized responses involve funding agencies, publishers, or organizations creating new forms of research services available to researchers from diverse institutional and disciplinary affiliations. For example, funding agencies are modifying grant proposal and award administration guidelines to promote the FAIR open science data sharing principles [8] and are increasingly requiring open access publication. Academic publishers are increasingly expecting reproducible supplementary materials or even preregistered analysis plans associated with research publications for enhanced replicability. Funding agencies, academic publishers are expanding their infrastructure for archiving and disseminating open access publications and supplementary materials. New centralized infrastructure has also emerged in the form of digital archives for research data and code, preprints, preregistrations, and registered reports; and these include the Open Science Foundation, Figshare, AsPredicted, arXiv, and more.

In the geographic sciences, centralized responses also include shared cyber-infrastructure resources to provide common computational environments (e.g., CyberGISX, I-GUIDE, and KNIME Geospatial), open data resources (e.g. Copernicus) and data enclaves for managing credentialed access to sensitive research data [see 9]. For example, the Hydroshare project bundles the data, computational environment, code, and metadata for a project together with a server that can not only store, but also execute, computational hydrological research. The o2r project developed a reproducibility service to load, inspect and execute research projects in Docker containers. These centralized responses are facilitated by a robust open source geospatial community creating and maintaining standards and software for storing, analyzing and visualizing geospatial data, including packages for three popular data science languages: R, Python, and SQL.

As researchers confront the challenge of selective inference and irreproducibility, they have also developed decentralized infrastructure. Decentralized infrastructure includes resources for creating and managing research compendia, which consist of the data and code required to execute the computational analysis. Decentralized infrastructure may also support tracking and controlling project versions, managing a computational software environment, citing internal and external research products, and writing analysis plans for preregistration or reports for post-analysis registration. An executable research

compendium is capable of executing all computations required to transform raw data into final results and of interweaving results with narrative in a computational notebook to render the research manuscript [10]. This decentralized infrastructure is most commonly deployed in the form of templates, research guides, and/or software packages in R or Python, and is often designed to interface with centralized infrastructure introduced above.

IV. The Unique Challenges of Geographic and Convergence Research

At the outset of our work with the I-GUIDE platform, we sought to use and adapt the decentralized infrastructure already developed by researchers to the challenges of geospatial and convergence research. However, early in our project we encountered limitations with each resource when we attempted to use them for geographic and convergent research. These challenges included the 1) additional requirements of geographic project metadata, 2) diversity and complexity of geographic data inputs, 3) restrictions on selected data, 4) complexity of computational environments and software dependencies, and 5) uncertainty and additional researcher degrees of freedom associated with geographic research. We present each of these issues in turn and briefly discuss the solutions we developed. These solutions are integrated into a template research compendium for human-environment and geographical sciences [11], designed as a GitHub template inclusive of a readme file, metadata, and analysis plan.

A. Geographic Project Metadata

It is well documented that spatiotemporal dependency, heterogeneity, and nonstationarity are important confounding factors for meta-analysis, and that discovering and confirming theorized mechanisms will likely require systematic protocols for replicating studies across different spatial and temporal contexts [12]. It is therefore essential to be able to determine the geographic extent, geographic resolution, temporal extent, and temporal resolution of a study to be included in meta-analysis or to be replicated. Unfortunately, this essential metadata about empirical geographic research remains ambiguous in the published literature [13], limiting our ability to search for research with geographic criteria or to design geographically explicit meta-analyses or systematic reviews.

In response, we adopt the Dublin Core metadata elements [14] for describing research project metadata. We expand the coverage Dublin Core element to explicitly describe the spatial reference system, spatial extent, spatial scale, temporal extent, and temporal scale of the research project. This metadata not only prepares researchers with the metadata they will need to register their project and its components in open science archives, but also enhances the findability of the project and ability to determine project suitability for replication and meta-analysis. Furthermore, the metadata will help independent researchers review the validity of researcher decisions in choosing data and transforming data into the desired spatial-temporal extent and resolution for analysis.

B. Diverse and Complex Data Inputs

One approach to reproducibility in science is to separately register or archive each research component with open access and a DOI, and then to cite each of these in the final manuscript. In this approach, the research components are all open and accessible, but distributed across many different digital resources. This strategy may suffice for a research design with just one data source created by the author, e.g. the results from an online survey. However, convergence research and geographic studies frequently use multiple and complex data inputs, implying a need to bundle the data inputs together in a research compendium. When gathered within a research compendium, data inputs also need to be organized and indexed.

The FAIR principles are one set of guidelines that outline how to index and preserve research inputs. The FAIR principles for open science rely heavily on metadata to support discovery and appropriate use of open data. The complexity of primary and secondary data sources in geographic research requires detailed attention to metadata about geographic and temporal extents and resolution, lineage, access and distribution, use constraints, quality, and variables and their accuracy. According to Wilson et al [15], the key to achieving five-star reproducibility is geospatial metadata documented according to international standards. The collection of primary data sources must be planned to avoid geographic bias and maximize potential for redistribution, and secondary data sources must be scrutinized for bias and uncertainty prior to use. Many disciplines have developed their own metadata standards for disciplinary-specific data (e.g. the Data Documentation Initiative for surveys in the social behavioral economic and health sciences or the Ecological Metadata Language for the ecological sciences). The authoritative standard for geographic metadata has been developed for spatial data infrastructures and encoded in the ISO 191** suite of geospatial metadata standards [16], with which both the European INSPIRE standards and United States FGDC standards comply. The standards are supported in commercial GIS software and in open source geospatial content management servers, but they are not well supported in open source research applications [11] and are scarcely applied by geographic researchers [17].

We provide directory structures specifically for raw input data and for metadata. An index table in comma-separated values (CSV) format links all data files in the compendium with short descriptions and detailed metadata files. As a first step to support decentralized geographic metadata documentation, we provide a template metadata document compatible with international standards. This human-readable documentation should allow researchers with different domain expertise to understand the nature and limitations of research data inputs, and allow version-tracking software to visualize changes in research data characteristics as projects evolve. While our solution still places the onus to generate metadata on the researcher, a next step in open science infrastructure development could be to create semi-automated open source software for users to create, update, and validate geospatial metadata.

C. Managing Restricted Data

Geographic research projects often use data with one or more of three restrictions: proprietary licenses, confidentiality, or very large file size. Researchers customarily deal with these challenges by storing data on password-restricted and encrypted servers and accessing data from repositories with authenticated logins. Some data providers prohibit redistribution of raw data and require use agreements and authentication to access restricted data repositories. Examples include the United States Agency for International Development's Demographic and Health Surveys and the Minnesota Population Center's Integrated Public Use Microdata Series. These data agreements limit how researchers can handle, archive, and redistribute data, but we have only observed careful attention to restricted data in one other decentralized infrastructure—the WORCS package for R [18].

We have implemented a decentralized solution to work on local research computers and bridge to restricted data providers. This solution is threefold. First, researchers should document thorough metadata for the raw data source, as detailed in the previous section. Second, the restricted data can be stored in our template private folders, which are not version-tracked and therefore will not be included when pushing data to remote servers with Git. This implies that restricted data can be downloaded and analyzed locally without risk of transferring the data with the compendium, or that the entire compendium could be transferred to a remote server with Git, executed, and transferred back with results and any unrestricted outputs. Third, once data is sufficiently deidentified, derived data can be stored in public folders.

D. Complex Computational Environments

Geographic research conducted with open science software uses scripting languages like R and Python with a multitude of different software package dependencies, and these packages are updated frequently. Moreover, there are multiple algorithmic approaches to implementing some geographic analyses, implying that the often overlooked selection of the computational environment embodies a significant researcher decision and possible source of uncertainty. In order to reproduce the research in the future, it will be necessary to reconstruct a computational environment similar to the original, or to invest substantial effort into modernizing the code to work with contemporary software packages. Unfortunately, most decentralized infrastructure for R&R are lacking resources for documenting the computational environment, much less reconstructing or containerizing it. Meanwhile, centralized infrastructure may provide server-based computational environments, but does not necessarily document or package the environment for reproduction elsewhere, implying dependence on maintenance of and access to the centralized server.

To begin solving this problem, we create a dedicated folder of the compendium and section of the analysis plan for documenting the computational environment. We also add code to template computational notebooks in R and Python to save metadata about software package versions and dependencies and to install and load an identical set of packages. Collectively, when researchers use these resources in conjunction with the version control system Git, a provenance record of the research process is created. The completeness of this record will depend on whether researchers work exclusively in this environment, and whether they commit to creating metadata files and project narratives. However, the resource should lower the cost of these tasks, hopefully increasing the odds of compliance.

E. Uncertainty and Researcher Degrees of Freedom

The goal of pre-analysis registration is to require researchers to register a research plan prior to observing data, thus removing the bias of researcher degrees of freedom from the research process. We know that the integration of multiple data sources and research disciplines for convergence research adds exponentially to the range of possible researcher decisions and forking paths for the research design, especially with regards to techniques for wrangling data sources into a common spatial-temporal analytical frame for the analysis. Unfortunately, extant templates for writing an analysis plan for pre-registration tend to emphasize sources of bias in observational and experimental designs, e.g. plans for participant recruitment, outlier treatment, and grouping criteria. Additionally, extant templates do not consider major threats to validity in geographic research, including the modifiable areal unit problem, spatial autocorrelation, spatial heterogeneity, boundary effects, and more.

Therefore, we have developed a template analysis plan for the preregistration of geographic research designed to prompt researchers to specify decisions related to major sources of error and threats to validity in our discipline. The template includes Dublin Core metadata for the project and ISO 191** series metadata for the input data sources so that both the spatio-temporal support of the analytical models and the input data sources is specified. The data transformations section specifies the geographic and attribute transformations required to wrangle input data into the analysis frame. A section devoted to bias and threats to validity invites researchers to specify sources of uncertainty and specify checks and steps to mitigate it. Although research designs may change as unanticipated challenges with data and methods are discovered, our approach to saving metadata and analysis plans in a version-tracked repository enable transparent provenance records of these changes and their possible influence on research findings.

V. Conclusions

We have outlined selected advances in centralized infrastructure for reproducible and replicable research and a need for parallel advances in decentralized infrastructure which individual researchers and research groups can use to structure their research practices and fulfill the goals of open science. While we have learned much from other scientific disciplines that have taken the lead in addressing the replication crisis, we have also identified several unique challenges in the nature of geographic and convergence science. These challenges are rooted in the complexity of geographic phenomena, the diversity and complexity of data sources in geographic research, and the practical challenges of working with restricted data in rapidly developing computational environments.

Fortunately, the challenges are not so dire as the wicked problems that convergence research aims to solve. Geographers already have a long history of addressing these challenges through development of open source GIS and metadata standards for spatial data infrastructures, and are capable of

applying those lessons to their own research projects. As of yet, the majority of geographic research is not readily reproducible, nor are the findings easily interpretable in the context of replication or meta-analysis. However, if geographic researchers can adopt a flexible yet structured approach to research compendium design, metadata documentation, and analysis plan registration, then they will drastically expand the possibilities for metascience and of geography's central role in convergence science moving forward.

We have already developed prototypes of decentralized R&R infrastructure, deployed it in our research and our methods curriculum, and revised it based on experience. We have already observed how it can improve the transparency, legibility, reproducibility, and replicability of our own research projects and those of our students. The next challenge is to cultivate an expanding research community to use and refine this infrastructure to enhance reproducibility and reduce the risk of selective inference in geographic and convergent research.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. (BCS-2049837). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

.

REFERENCES

[1] Y. Benjamini, "Selective inference: The silent killer of replicability. Harvard Data Science Review, vol. 2, iss. 4, 2020, https://doi.org/10.1162/99608f92.fc62b261.

[2] A. Gelman, and E. Loken, "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time." Department of Statistics, Columbia University, vol. 348, pp. 1-17, 2013.

[3] V. Stodden, F. Leisch, and R.D. Peng (Eds.), "Implementing reproducible research" CRC Press, 2014.

[4] G. Christensen, and M. Edward, "Transparency, reproducibility, and the credibility of economics research." Journal of Economic Literature, vol. 56, iss. 3, pp. 920-80., 2018, http://doi.org/10.1257/jel.20171350

[5] J.P. Simmons, L.D. Nelson, and U. Simonsohn, "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science, vol. 22, iss. 11, pp. 1359-1366, 2011, https://doi.org/10.1177/0956797611417632

[6] B. Yu., and K. Kumbier, "Veridical data science." Proceedings of the National Academy of Sciences, vol. 117, iss. 8, pp. 3920–3929, February 2020, https://doi.org/10.1073/pnas.1901326117

[7] P. Kedron, and S. Bardin, "A Vision for Veridical Spatial Data Science". UC Santa Barbara, 2021, Spatial Data Science Symposium 2021 Short Paper Proceedings, 2021 http://doi.orf/10.25436/E20016

[8] M.D. Wilkinson et al. "Comment: The FAIR Guiding Principles for scientific data management and stewardship." Scientific Data, vol. 3, pp. 1–9.

[9] D.B. Richardson, M.P. Kwan, G. Alter, and J.E. McKendry. "Replication of scientific research: addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research." Annals of GIS, vol. 21, iss. 2, pp. 101–110, 2015.

[10] D. Nüst, D., and E. Pebesma, "Practical Reproducibility in Geography and Geosciences." Annals of the American Association of Geographers, vol. 111, iss. 5, pp. 1300–1310, 2021, https://doi.org/10.1080/24694452.2020.1806028

[11] P. Kedron, and J. Holler, "Template for Reproducible and Replicable Research in Human-Environment and Geographical Sciences." Open Science Framework, 2022, https://doi.org/10.17605/OSF.IO/W29MQ

[12] P. Kedron, and J. Holler, "Replication and the search for the laws in the geographic sciences." Annals of GIS, vol. 28, iss. 1, pp. 45–56, 2022, https://doi.org/10.1080/19475683.2022.2027011

[13] J.D. Margulies, N.R. Magliocca, M.D. Schmill, and E.C. Ellis, "Ambiguous geographies: Connecting case study knowledge with global change science." Annals of the American Association of Geographers, vol. 106, iss. 3, pp. 572–596, 2016, https://doi.org/10.1080/24694452.2016.1142857

[14] Dublin Core Metadata Innovation, "Using Dublin Core - The Elements." 2005. https://www.dublincore.org/specifications/dublin-core/usageguide/elements/

[15] J.P. Wilson, K. Butler, S. Gao, Y. Hu, W. Li, and D.J. Wright, "A five-star guide for achieving replicability and reproducibility when working with GIS software and algorithms." Annals of the American Association of Geographers. Vol. 111, iss. 5, pp. 1311-1317, 2021

[16] International Standards Organization. "ISO 19115-1 Geographic information – Metadata" 2014, https://www.iso.org/standard/53798.html

[17] P. Kedron, J. Holler, and S. Bardin, "Reproducible Research Practices and Barriers to Reproducible Research in Geography: Insights from a Survey." 2023, https://doi.org/10.31219/osf.io/nyrq9

[18] C.J. Van Lissa, A.M. Brandmaier, L. Brinkman, A.-L. Lamprecht, A. Peikert, M.E. Struiksma, and B.M.I. Vreede. "WORCS: A workflow for open reproducible code in science." Data Science, vol. 4, iss. 1, pp. 29–49, 2021.