INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2020.1831514

RESEARCH ARTICLE

Understanding the multifaceted geospatial software ecosystem: a survey approach

Rebecca C. Vandewalle[a,b], William C. Barley[c], Anand Padmanabhan[a,b,d], Daniel S. Katz[d] and Shaowen Wang[a,b]

[a]Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [b]CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [c]Department of Communication, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [d]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL, USA

ABSTRACT

Understanding the characteristics of the rapidly evolving geospatial software ecosystem in the United States is critical to enable convergence research and education that are dependent on geospatial data and software. This paper describes a survey approach to better understand geospatial use cases, software and tools, and limitations encountered while using and developing geospatial software. The survey was broadcast through a variety of geospatial-related academic mailing lists and listservs. We report both quantitative responses and qualitative insights. As 42% of respondents indicated that they viewed their work as limited by inadequacies in geospatial software, ample room for improvement exists. In general, respondents expressed concerns about steep learning curves and insufficient time for mastering geospatial software, and often limited access to high-performance computing resources. If adequate efforts were taken to resolve software limitations, respondents believed they would be able to better handle big data, cover broader study areas, integrate more types of data, and pursue new research. Insights gained from this survey play an important role in supporting the conceptualization of a national geospatial software institute in the United States with the aim to drastically advance the geospatial software ecosystem to enable broad and significant research and education advances.

ARTICLE HISTORY
Received 12 December 2019
Accepted 29 September 2020

KEYWORDS
CyberGIS; cyberinfrastructure; data science; geospatial software; geospatial discovery and innovation

1. Introduction

Geospatial software and related sciences and technologies are rapidly evolving, changing the face of our human, natural, and digital worlds while shaping how we solve a variety of significant scientific and societal problems (Anselin 2012, Wang 2016, Yuan 2017, Shaw and Sui 2018). Geospatial software can be defined as digital artifacts for transforming geospatial data (data that has geo and/or spatial references) into knowledge, insights, and intelligence (Wang *et al*. 2016). As geospatial data grows at an unprecedented pace with

CONTACT Shaowen Wang shaowen@illinois.edu
The authors declare no conflicts of interests
Supplemental data for this article can be accessed here.

increasing complexity and coverage, it has become arguably one of the most challenging types of big data (Wang and Zhu 2008, Miller and Goodchild 2016, Shook *et al*. 2019). Harnessing complex geospatial data poses tremendous challenges to geospatial software, advanced cyberinfrastructure, and related research communities, due to the many formats and models available, computational considerations for various spatial characteristics, the need to often integrate capabilities across multiple tools, and the fragmentation of the geospatial software community hindering interdisciplinary and transdisciplinary research opportunities (Anselin and Rey 2012, Wang 2010, Yang *et al*. 2010, Li *et al*. 2015).

These challenges are both social and technical in nature, as the complexity of the geospatial software ecosystem causes increasing difficulties finding data and tools, building communities of scholarship interested in domain-specific applications, and making the findings of geospatial research readily accessible to and reproducible by the constituencies that would best benefit from them (Wright and Wang 2011, Wang 2013, Stodden *et al*. 2015, Kedron *et al*. 2019, Konkol *et al*. 2019). Such challenges demonstrate some of the emergent consequences associated with the rapid growth in scientific and societal value for geospatial software. As the geospatial software ecosystem continues to grow and evolve in the foreseeable future, its complexity needs to be better understood particularly in terms of how inherent difficulties and challenges may hinder efforts from researchers and the broader community to effectively solve significant scientific and societal problems (e.g., emergency management, food security, global change, and disease spread) beyond what any single domain of expertise can achieve. This paper seeks to

Table 1. Broad categories of perceived software limitations determined through thematic analysis (n = 127). Frequency indicates number of respondents mentioning that issue (codes can overlap).

Description	Frequency	Description	Frequency
Limited general functionality	49	Poor documentation, lack of customization	20
Cannot handle specific needs	44	Workflow breaking updates, poor communication between programmer and user	12
Limited functionality for specific tasks	40	Open source options are not available, limited, or buggy	10
Poor interoperability	35	Poor support for web capabilities	10
Data challenges	31	Poor support for modelling	8
Poor support for specific data types	31	Expensive	6
Poor or no big data support	29	Licensing, dependencies and, compatibility issues	6
Not user friendly, learning curve, training needed	26	Poor parallel support	6
Slow	23		

Table 2. Broad categories of limitations determined through thematic analysis (n = 146). Frequency indicates the number of respondents mentioning a specific limitation (codes can overlap).

Description	Frequency	Description	Frequency
Data availability	60	Poor workflow, compatibility, needs programming	19
Data formats, interoperability	25	Steep learning curve, lack of skills/knowledge needed, lack of specialists, hard to access support	19
Data cost	24	Too slow processing	15
Software does not handle parallel computation, big data	23	Not enough memory or storage	10
Poor computational support	22	Too big datasets	3
Software does not meet needs, limited functionality	21		

understand current limitations to inform the community of specific needs and requirements for advancing the geospatial software ecosystem in support of future research and education.

Although it is well understood that there are extensive challenges for harnessing complex geospatial data to pursue knowledge discovery (Murray 2019, Wang and Goodchild 2019), there exists limited understanding of how geospatial software developers and users perceive these challenges. Given the sheer breadth of these challenges, understanding both developer and user perceptions of the most pressing challenges can help policymakers recognize *which* challenges in the geospatial software context that if addressed hold the potential to beneficially impact that largest swaths of the community base. This report on a community survey aims to share the knowledge gained from the broad geospatial software community about their experiences and perceptions of interacting with geospatial software, and examine the research question of how to take advantage of this knowledge to guide the development of the future United States geospatial software ecosystem for enabling broad and significant research and education advances.

In the context of addressing this question, the United States National Science Foundation (NSF) has funded a project to conceptualize a national Geospatial Software Institute (GSI) that aims to create bridges across many geospatial domains by establishing a long-term cyberinfrastructure hub of excellence in geospatial software to serve diverse research and education communities. Such software institutes for multiple other science and engineering domains, including high-energy physics (https://iris-hep.org), molecular sciences (https://molssi.org), and science gateways (https://sciencegateways.org), have been established through significant NSF support ranging from 18 USD million to 25 USD million for each institute for the initial 5 years. Given the significance of the potential funding for a GSI similar to the level of these established software institutes and desirable impacts anticipated from the creation of such an institute, the GSI conceptualization project has used a survey and three cross-disciplinary workshops to gain input from broad geospatial-related communities to gain insight for the proposed GSI's mission, direction, and goals.

This paper focuses on the project's survey, which was developed to gain a broad sense of the concerns and limitations related to geospatial software development and use. The survey was also intended to better understand the characteristics and diversity of the United States geospatial software community by identifying those who might have not yet been represented in the engagement and outreach efforts of the NSF GSI conceptualization project. Therefore, we designed our survey questions to identify the axes of variance for assessing where participants may have had different experiences and interactions with geospatial software. Although the survey was primarily designed to support the planned GSI, we believe knowledge gained from this exercise is valuable to the broader geospatial software community. Our central objective is to distinguish among different types and intensities of interactions with geospatial software.

The findings from the survey results indicate that 1) the community base is incredibly diverse; 2) despite prevalent discourse for developing high-performance computing capabilities at the petascale, exascale and beyond, a large percentage of users work at the desktop level; 3) despite general satisfaction with existing tools, 41% of respondents say they are limited by the capabilities of their existing tools; and 4) these limitations were patterned, where primary needs surround data access and integration, software tool interoperability, access to advanced cyberinfrastructure resources, software tool discovery and training, and solutions for buggy or fragile software.

The rest of the article is organized as follows. The background section reviews related work in literature. The method section discusses the rationale, design, and structure of the survey. The results section presents survey results and emergent patterns. Finally, the concluding summary section summarizes the findings of this research, addresses the implications of the findings, and highlights the broad relevance and importance of the research to the future of the emerging and growing geospatial software ecosystem.

2. Background

Previous efforts on developing coordinated spatial data and information infrastructures provide insight into recent emergence of geospatial software ecosystems (Rajabifard and Williamson 2001, Masser 2006, Wang 2013). In particular, spatial data infrastructure development in the United States effectively started when the National Mapping Division became aware of the imbalance between short-term data production and longer-term research needs and started to shift emphasis to emergent technologies and changing user needs (National Research Council 1991). This led to a key spatial data infrastructure proposal and the creation of the Federal Geographic Data Committee (National Research Council 1993, Office of Management and Budget 2002). Spatial data infrastructure priorities subsequently evolved from a narrower focus on efficiency and better utilization to a broader emphasis on supporting distributed data collection, storage, and management that relied on partnerships and interoperability (National Research Council 1993, Office of Management and Budget 2002).

With more focus on geospatial data interoperability at local, national, regional, and global levels, challenges in coordinating and integrating geospatial infrastructures have become apparent (Coleman and McLaughlin 1998, Rajabifard and Williamson 2001). Sustainable geospatial data infrastructures rely on cooperation and partnerships between government, industry, and academic stakeholders (National Research Council 1994, Onsrud *et al*. 2005, Masser *et al*. 2008). Consequently, there are many competing visions for what a unified geospatial data infrastructure is and should be, what it does and should do, and how it does it and how it should do it (Coleman and McLaughlin 1998, Rajabifard and Williamson 2001). These priorities may be categorized into data-driven, technology-driven, market-driven, application-driven, and institutional perspectives (Coleman and McLaughlin 1998). Although early geospatial data infrastructure development campaigns were operational and focused on geospatial data as products, later development efforts became increasingly strategic and process-based (Coleman and McLaughlin 1998, Rajabifard and Williamson 2001). Finally, recurrent discussions on data sharing standards underscored the careful consideration needed to maintain coordination, interoperability, and common structure while supporting continued diversity and innovation (National Research Council 1994, 1997, Masser 2006).

Certain lessons from geospatial data infrastructure development can be transferred to geospatial software cyberinfrastructure development. In order to create new knowledge using software, underlying datasets must be reliable, current, and interoperable (Yang *et al*. 2010, Wang *et al*. 2012). In light of this, for nurturing geospatial software ecosystems, spatial data infrastructure development histories, accomplishments, and unsolved challenges should be considered. Challenges related to duplication of software creation efforts and potential underutilization of valuable software assets may also reduce benefits from any emergent geospatial software ecosystem. Finally, software fragmentation exacerbates data incompatibilities. Using geospatial data infrastructure development as a guide, some factors to consider include approaches to integrating diverse stakeholder priorities, support for equitable software access, strategies for increasing buy-in from other stakeholders, evaluation techniques, and user priorities and needs. User perspectives are essential to the implementation and maintenance of a national geospatial software ecosystem (Schade *et al*. 2020).

As geospatial data infrastructure development efforts have matured, disconnects between researchers, managers, developers, and users have threatened ultimate project success. Profound gaps between technological and social-cultural priorities, investment and adoption rates, and success criteria have become increasingly apparent (Atkins *et al*. 2003, Bernard *et al*. 2005, Bernard and Craglia 2005, Budhathoki and Nedović-Budić 2007, Cutcher-Gershenfeld *et al*. 2016). Concerns that high-cost and intensive-effort geospatial data infrastructure programs had ultimately been underused has spurred greater attention towards user needs and barriers to their discovery and adoption of these infrastructures (Budhathoki and Nedović-Budić 2007). A particularly grave failure mode is manifest in the ‘if you build it, they will come’ mentality, which underscores a focus on technological aspects over subsequent utility (Cutcher-Gershenfeld *et al*. 2016). This mindset elucidates several tensions between domain scientist end-users and infrastructure experts, namely different capacities, communication styles, cultural foundations, priorities, funding, and incentives to participate in cyberinfrastructure efforts, as well as differences in when benefits and incentives materialized during the course of development (Finholt and Birnholtz 2006, Cutcher-Gershenfeld *et al*. 2016).

The high rate of stalled, immature, incomplete, and underutilized geospatial data infrastructure projects underscores the absolutely critical need to understand and foreground user needs at all stages of development, implementation, and assessment (Harvey and Tulloch 2006, Budhathoki *et al*. 2008, Lee *et al*. 2010, Poore 2011). This task is not easy; users are heterogeneous, can participate in multiple groups, may have multiple roles simultaneously, and may have interests beyond direct tangible benefits from particular infrastructure (Zimmerman 2007, Lee *et al*. 2010, Cutcher-Gershenfeld *et al*. 2016). Efforts to better understand user needs are further complicated due to a changing technological environment and evolving user requirements; unfortunately, current conditions cannot reliably predict future conditions and needs (Atkins *et al*. 2003, Budhathoki *et al*. 2008, Lee *et al*. 2010). Viewing users as active and integrated into the development process rather than passive and receptive contributes to bridging the technological social divide (Budhathoki *et al*. 2008, Poore 2011). There is a clear agreement that more work needs to be done across the board to gain actionable information from users (Bernard and Craglia 2005, Michener *et al*. 2012, Cutcher-Gershenfeld *et al*. 2016, Stocks *et al*. 2019). This underscores a broader need to identify user priorities and needs, understand user capabilities and perspectives, and continuously solicit and adapt to feedback.

Previous research on learning user needs and capabilities within the geospatial software landscape has been undertaken within geospatial data infrastructure and cyberinfrastructure development initiatives (Atkins *et al*. 2003, Zimmerman and Finholt 2008, Stewart *et al*. 2011, Michener *et al*. 2012, Cutcher-Gershenfeld *et al*. 2016, Stocks *et al*. 2019). Although a broad and diverse set of software tools are desired and used within diverse geospatial domains, existence, accessibility, cost, performance, and interoperability remain pressing concerns (Stewart *et al*. 2011, Stocks *et al*. 2019). Software development may disproportionately support certain portions of data lifecycles more than others (Madhavji *et al*. 2015, Kumar and Alencar 2016). Needed tools are often under-developed, under-discovered or simply do not exist (Stocks *et al*. 2019). User capabilities vary in terms of accessibility to computing resources, technical knowledge necessary to adopt geospatial software, and access to institutionally gathered computational experience that could make up for technical knowledge gaps (Zimmerman 2007). A tendency for siloed sharing and a lack of knowledge of other software projects leads to duplication of software development efforts (Stewart *et al*. 2013). Finally, incentives to create and sustain geospatial software ecosystems are hindered by lack of institutionalized rewarding mechanisms, challenges of collaborative software development, publicizing new tools that are available, and flexible access to advanced cyberGIS and cyberinfrastructure (Budhathoki and Nedović-Budić 2007, Michener *et al*. 2012, Kim and Stanton 2013, Wang 2013).

3. Method

The geospatial software community is diverse and rapidly evolving. Therefore, it is important for the survey to understand the characteristics and needs of the community (Kemp and Frank 1996), by looking into both respondents' perceptions of the status and future of geospatial software, and how they approach developing and using geospatial software for different purposes. We intended to reach a wide audience and purposefully used a broad definition for geospatial software (GSS) as software that interacts with geospatial data.

The survey text was drafted during a series of meetings and piloted with a wide range of individuals. We also compiled a set of geospatial community mailing lists that were used to seek survey participants and completed the IRB approval process. See the supplementary materials for the survey text and number of respondents per question. Survey data were analyzed through descriptive charts and statistics available from Qualtrics-based reports and also through quantitative analysis. Categories for free response questions were determined by 'thematic analysis' produced via an iterative coding process, whereby the first author coded the entire dataset separately from the second author (Corbin and Strauss 2015, Tracy 2020). Then, the entire research team explored and discussed themes to ensure they had internal consistency. Some responses match multiple codes. As our research questions were exploratory in nature, analyses in this paper are largely descriptive and inductive.

We approached sampling with as broad a representation of geospatial software experts and users within the United States as possible and also welcomed insight from international participants. Survey participation invitations were distributed through pertinent community mailing lists, including lists for the American Association of Geographers, the Association of Geographic Information Laboratories for Europe, CyberGIS, Geo4all, ACM SIGSPATIAL, the National Socio-Environmental Synthesis Center, the University Consortium for Geographic Information Science, NSF Extreme Science and Engineering Discovery Environment (XSEDE) and CRYOLIST (an international email list

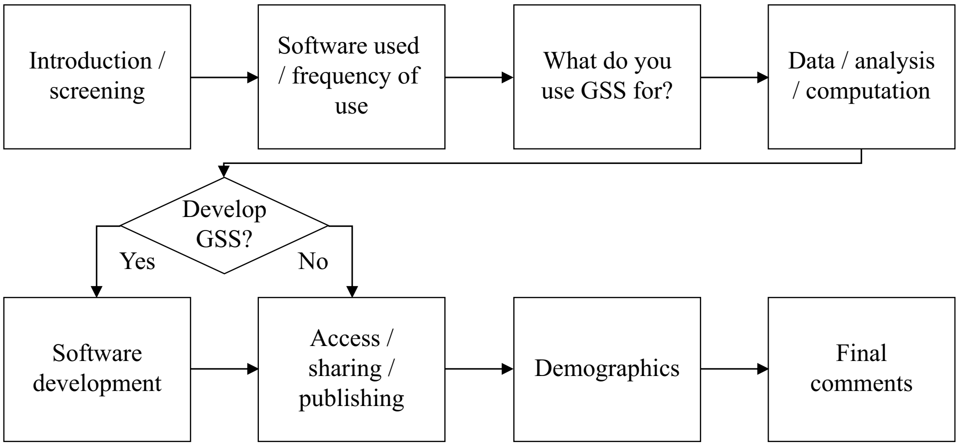

Figure 1. Survey structure and flow.

highlighting snow and ice). Respondents accessed this survey using the Qualtrics web-based platform between January and December 2018.

This survey aimed to gather foundational information on respondents' experience with geospatial software. The survey had five primary sections, designed to explore the respondents' current geospatial software use patterns and use cases, data and computational characteristics of a typical project, software development approaches, and experiences with publishing geospatial software. These divisions were informed by survey team members' prior understandings of the geospatial software landscape. The survey contained a total of 56 questions. Some questions were displayed based on prior responses, and the software development section was only shown to respondents who indicated they had developed geospatial software. Figure 1 demonstrates the overall survey structure.

Survey participation was voluntary and respondents could exit at any time. As a screening question, we asked whether users had either used or developed software tools that engage geospatial data within the past 3 years. 454 respondents passed the screening question, and 270 respondents answered the last question. Uncomplete survey data were recorded after two inactive weeks. On average it took approximately 1 h for those who finished to reach the end of the survey (this metric includes time between active sessions).

For gaining a better sense of our respondents' backgrounds we categorized open-ended responses for occupational fields and organizational affiliations. 57% of respondents reported a broadly geospatial field. Other common categories included computer science, environmental sciences, informatics, and social sciences. For organizational affiliation, roughly two-thirds of respondents were affiliated with universities, 20% with government organizations, and most of the rest with industry or non-profits.

4. Results

4.1 Overview

Most respondents perceived geospatial software as extremely important to their work (Figure 2), were reasonably satisfied with the geospatial software they used, and used geospatial software daily or weekly. Although users were seemed generally content with available tools, 42% of respondents indicated that they felt that their work was limited by inadequacies in geospatial software (Figure 3). As a whole, survey results show critical limitations for geospatial software access, interoperability, functionality, and training. In the following portions of this section, we discuss survey findings related to software, data, computation, geospatial community engagement, general challenges, and envisioned research gains without limits.

4.2 Software

4.2.1 Software used and frequency of usage

Our first aim was to learn how often our respondents used and accessed geospatial software so we could later investigate the relationships between how they used software and limitations they encountered while using it. Respondents reported using a wide range and diverse combinations of software tools. Most commonly listed were Esri products (66%), followed by open-source options such as QGIS and GRASS GIS (45%), and Python and/or R (45%). Additionally, 15% of respondents specifically reported using Google products (such as Maps, Earth, and Earth Engine). Approximately 20% of respondents relied on programing and custom tools. There was a general agreement that necessary geospatial software was mostly accessible; however, the main hindrances were cost, steep learning curves, and insufficient documentation. Most respondents used geospatial software frequently while fewer regularly developed it.

These results underscore software's fundamental role in geospatial research. Findings in this study complement prior user reports, which show high diversity in tools used,

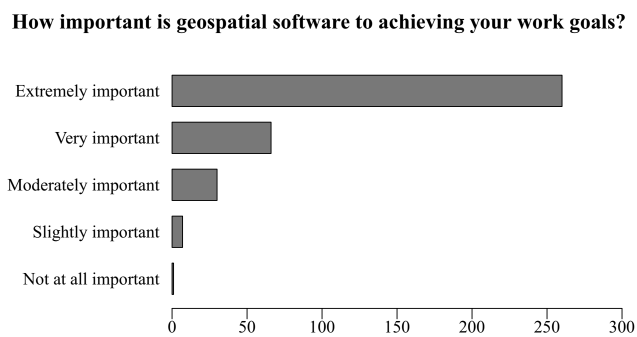

Figure 2. Geospatial software importance for achieving work goals.

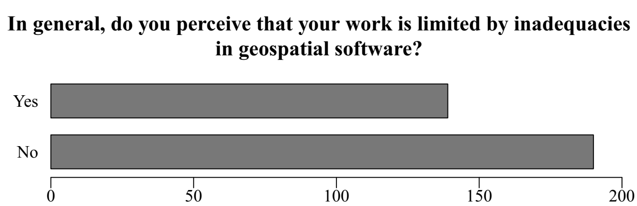

Figure 3. Perceived limitations of geospatial software.

regular reliance on custom in-house tools, and also that often needed software does not exist (Stewart *et al*. 2011, Stocks *et al*. 2019). Software tool diversity may indicate a set of interrelated gaps: 1) mainstream tools provide insufficient support for specific research needs, 2) researchers face difficulties discovering, accessing, and integrating needed tools, or 3) researchers lack adequate time and access to software experts who would enable them to streamline geospatial workflows. These results suggest an urgent need for improved software education, training, and consulting.

4.2.2 Software use cases

After obtaining a general picture of geospatial software use, we investigated differences in use and satisfaction relative to geospatial data lifecycle phases. We used five overarching phases: pre-processing, running analyses, visualization, integration, and conducting modeling and simulation. Although respondents used geospatial software throughout the entire data lifecycle and were fairly satisfied overall with software capabilities, the modeling and simulation phase was used much less frequently and viewed the least satisfactorily (see Figure 4). Two scenarios might explain data for modeling and simulation: 1) either there is insufficient demand or 2) demand exists but there is insufficient support. Drawing from our open-ended responses, we expect that the second scenario is likely as respondents indicated they wanted more modeling and simulation capabilities but, due to a lack of support, instead often shifted usage patterns.

4.2.3 Limitations of geospatial software

A major aim of this survey was to determine whether geospatial software users thought that geospatial software limited their work capability and, if so, what knowledge gains they could make without these limitations. Broadly speaking, users perceived patterned limitations in geospatial software tools that are seen as unfriendly and difficult to discover, and that inadequate documentation and training are available for using available tools efficiently (Table 1). At the same time, they had concrete understandings of how efforts to ameliorate these challenges would impact their work and support their research and education.

Challenges reported ranged from accessing and acquiring software to using software and exporting results. Respondents brought up interoperability issues, such as that many different tools and programs were needed for their workflow and that individual tools were too specialized and needed to be painstakingly stitched together. Different platforms often required different software solutions, and the lack of software and tool standardization hindered interoperability. The vast number of available software tools made it hard to stay current; simultaneously, not enough targeted software existed to fulfill certain needs. Taken together, these results indicate our respondents perceived the current geospatial software landscape as characterized by a fragmented constellation of software tools with an uneven distribution of capabilities. This highlights a real opportunity for support through enabling greater interoperability and between tool workflow.

A noticeable subset of responses specifically highlighted software deficiencies for handling geospatial data. Respondents noted software challenges in dealing specifically with data interoperability, integration, aggregation, and transfer. A common view was that software tools did not handle specific types of data well, such as big and multidimensional data, data from atypical geospatial sources, and qualitative data. Some

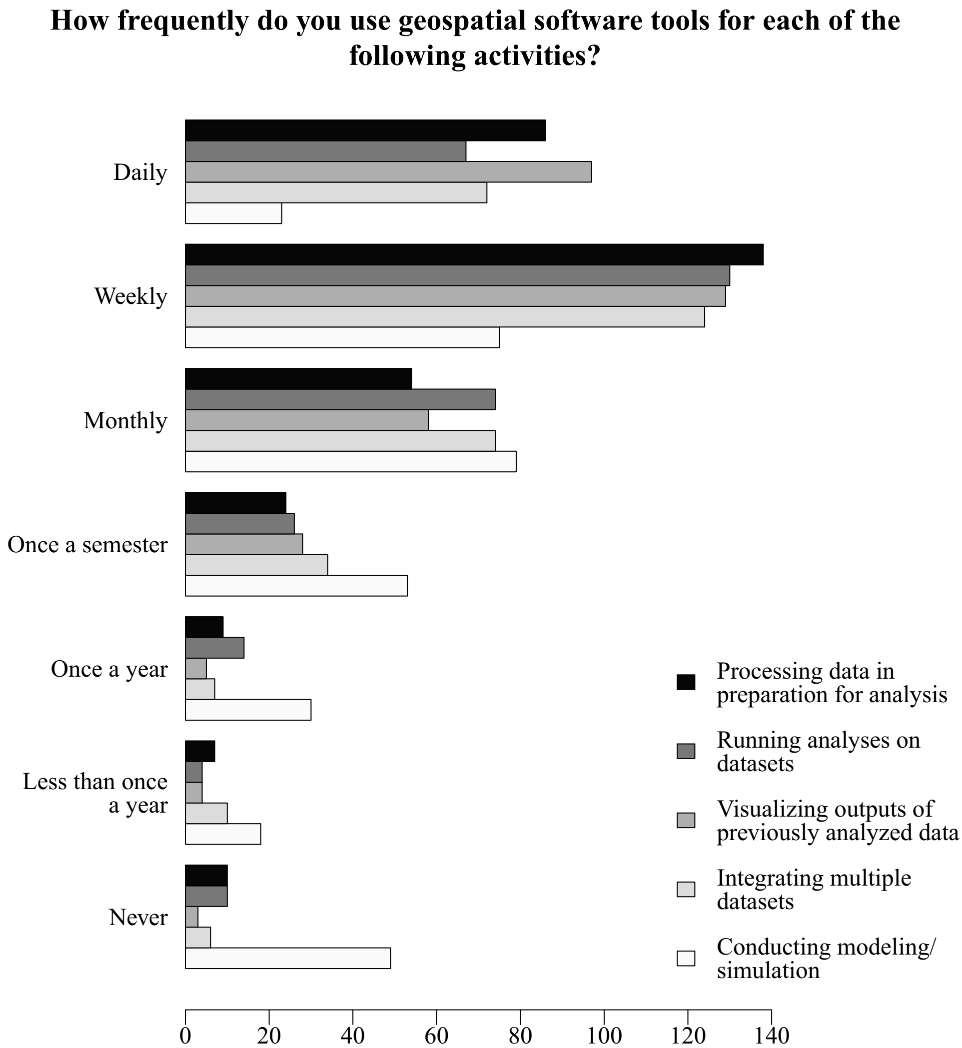

Figure 4. Frequency of geospatial software use.

specific areas that could benefit improvement included dynamic visualization, scripting and batch processing, and handling time, coordinate systems, and map projections. Importantly, scientist-user needs appear to be overlooked and unfulfilled by geospatial software developers. This disconnect mirrors tensions uncovered in infrastructure building efforts, where failing to account for different incentives and measures of success between these two stakeholder groups had the potential to collapse the initiative (Cutcher-Gershenfeld *et al.* 2016).

4.2.4 Software development

This survey section was only presented to respondents who had indicated they developed or customized geospatial software. Software development across data lifecycle phases often occurred weekly; however, similar to trends in software use, developers contributed to the modeling and simulation phase much less often than other phases. Although big data are frequently considered something that should be harnessed by geospatial software, in practice tools for handling big data appear insufficient even to software developers. When asked about which scaling limitations that could affect software performance, 41% of respondents indicated limitations in software design that caps maximum data and problem size, as well as a lack of parallel computing capability. Other scaling problems include institutional limitations, cost, difficulty of use or access to scalable resources, poor software design and inefficient base algorithms, and model complexity. The most common scalable computing model reported was batch processing, at 32%, followed by MPI (11%), OpenMP (9%), CUDA (8%), and MapReduce (8%). However, 25% of respondents did not use any scalable computing models. Altogether, results from this section builds on findings from other survey sections which call for better integration of high-performance infrastructure into geospatial software workflows. This is strengthened by our findings here that even those who have experience customizing geospatial tools find current high-performance capabilities limiting and lacking.

4.3 Data

From questions about past data use, we aimed to gain a patterned understanding of the scale and scope of geospatial software work that users take on in their current projects. We asked respondents to consider a recent geospatial project they had worked on while answering the subsequent questions for more concrete and specific responses. Most respondents indicated that this was a typical project. When we asked respondents whether their example project had been limited by certain factors such as data availability, geospatial software, computational support, and 'other factors' were chosen as responsible for the limitations encountered to a similar degree. More detailed categorizations of limitations are listed in Table 2. These answers suggest that even if software limitations were addressed, progress towards supporting interoperability and complex analyses needs equal attention paid to data handling and other challenges.

Handling spatial data was recognized as a challenge before explicit geospatial software functionally existed (Dangermond and Goodchild 2020). Accordingly, early calls for geospatial infrastructure focused primarily on issues surrounding data production, collection, storage, standardization, interoperability, sharing, and discovery (National Research Council 1993). Although this survey is primarily in support of a software ecosystem, responses indicate that fundamental challenges surrounding data still remain unsolved and still must be grappled within the context of software development and use. Even a well-constructed software infrastructure will likely stay underused if critical data availability and interoperability issues are not also acceptably addressed. Our survey responses reaffirm the critical gap between the importance of finding, accessing, and integrating data, models, and tools and the ease of actually doing so (Cutcher-Gershenfeld *et al.* 2016).

4.4 Computational limitations and resources

Effective software is highly dependent on computational underpinnings, so we asked respondents about what computational resources they needed and limitations they encountered. As 43% of respondents indicated that computational capabilities limit what they can achieve in their work, this area demands improvement (Figure 5). Concerning systems used, a slight majority of respondents (52%) indicated that they used primarily their personal desktop or laptop to run geospatial software, followed by departmental servers (24%), and campus-wide computing resources (11%). Relatively fewer respondents reported primarily using cloud computing resources (7%) or national computing resources (5%). However, several clarified that they did not have a single preferred system and often utilized different systems concurrently.

The top two contenders for high-performance, big data, or cloud computing environments used by respondents were MPI and OpenMP, which were often listed together. Amazon was also often mentioned, as well as GPUs, private cloud resources, Spark, and Hadoop. There was some reliance on university, department, and lab-specific clusters, cloud resources, and servers, as well as national computing resources and dedicated private servers. Approximately 30% of respondents said they did not use any of these environments or were not aware of which environments they used.

A variety of computational limitations were described. Respondents noted issues with computational components and resources, such as limited memory, bandwidth, storage space, not enough GPUs, and too few CPU cycles. Additionally, respondents mentioned issues with high-performance interfaces, noting difficulties with transferring and visualizing data. Respondents brought up processing issues, such as difficulties setting up processing routines, ability to deal with data that does not fit into memory, inefficiency, and specific requirements such as processing a very high volume of data for a short time. Respondents reported issues trying to utilize high-performance computing, such as needing a different skillset to run analyses on computing clusters, realizing some tools do not work on clusters, not having the correct type of computing cores. Several respondents noted that they primarily use desktop computers, which fail to scale well. Additionally, respondents brought up issues with cloud computing, noting the cost, and time it takes to set up cloud analyses, lack of reliability, and issues with cloud I/O. Respondents also pointed out broader structural issues, such as lack of access to high-performance resources granted by an institution or organization, lack of support, and lack of knowledge to be able to take advantage of these resources, and budgetary limitations.

As desktop computers have gained memory and power, many routine geospatial tasks are less reliant on specialized and expensive computing resources. High-performance geospatial computing relies on a large body of foundational studies; therefore, support for geospatial researchers working at smaller scales and complexity levels must not be discounted. However, some believe that too much focus is placed on raw computational

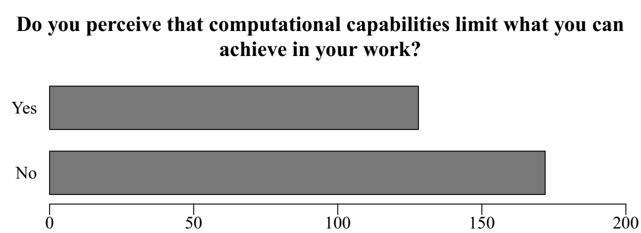

Figure 5. Perceived limitations due to computational capabilities.

performance in cyberinfrastructure efforts (Atkins *et al*. 2003). Results from this survey show that although effort should be made to support visionary computationally intensive work, effort is also needed to support capabilities of current researchers working at desktop or cluster levels in addition to bridging gaps that hinder adoption of high-performance workflows. Attention should be paid to market forces, funding, education, and culture, all of which hinder geospatial researchers from effectively expanding beyond the desktop (Gahegan 2018). A digital divide in access to computational resources still exists and must be taken into consideration (Stewart *et al*. 2011). For example, high-performance computing may be more widely used if it was facilitated through a client-focused approach, rather than a local computational environment.

When faced with computational challenges, a plurality of respondents (40%) reported developing code to fix the issues. Respondents (23%) also indicated that they sought out a group with computational expertise to collaborate. Respondents also sought support in-house or from a specific support team, contacted software companies directly, interrogated personal network connections or mailing lists, or worked with specific collaborators. Broadly, this supports a view of active and engaged users who can take steps to handle challenges, however suggests support may be lacking for those without the skillset or time to take care of the issues.

4.5 Geospatial community engagement

Aspects of geospatial community engagement we covered in this survey relate to software and data sharing practices and collaboration. Data and software sharing efforts face both cultural and practical challenges. Similar publishing levels were reported for data and software; in the past three years, 44% of respondents published data and 40% of respondents who had developed geospatial software published or shared software. Although many respondents were motivated to publish or share software, 34% found the process difficult. Software and data repositories were most often used for publication. Software repositories were preferred compared to other software publishing outlets with 41% of respondents who published software using software repositories. In comparison, data publishing outlets were more evenly utilized. When we asked respondents to list what they considered to be the top five outlets for publishing software and data, the resulting list was very diverse. These results suggest that while software repositories are a frequent choice for geospatial software sharing, publishing data are more closely tied to publishing research articles, which leads to a sprawling set of potential publication outlets. This huge data outlet diversity can frustrate researchers looking to keep abreast of the latest developments.

We also found robust patterns of collaboration among researchers who reported developing or customizing geospatial software. 59% of respondents who had developed geospatial software reported that they had computational science expertise in their development team and 55% collaborated with other research groups to develop geospatial software.

4.6 Challenges

To get a general assessment from respondents and to cover any unaddressed aspects, we asked them to describe the most pressing challenges they had encountered. Respondents reported challenges related to learning and interacting with software. Also mentioned was the difficulty in finding and hiring experienced professionals for research and support. Additionally, the often ad-hoc short-term solutions found and lack of unified approaches to development often leads to fragile software solutions and unnecessary duplication of work. Respondents also noted that the lack of good documentation hindered sharing and reuse.

Data related challenges were also problematic, with common issues being in discovery, availability, poor metadata, challenges surrounding data publishing, interoperability, lack of standards, and poor support. High-performance computation also was challenging, as respondents mentioned issues including scaling and handling big data, lack of scalable computing support, not enough storage space, memory, processing power, software incompatibility, and poor I/O support.

More general challenges were also listed. Examples include poor standards, poor interoperability, and the need for too many separate tools, slow development, not enough open source and too many black-box tools, poor support for high-resolution data, few affordable solutions, poor visualization capabilities, difficulty with reuse, non-user-friendly interfaces and poor GUI design, and poor support for specific types of data or analyses. Other concerns brought up by respondents included issues with operating system updates and software dependencies, security concerns, difficulty choosing the right tool for tasks, and computational reproducibility.

4.7 Envisioning unhindered potential research advances

In order to gain a vision of the benefits of an improved geospatial software ecosystem, we asked respondents what scientific questions they would choose to pursue if they were not hampered by geospatial software limitations. Example questions that could be addressed spanned natural sciences, social sciences, and geospatial research areas, covering topics such as movement and trajectories, urban sprawl, flood forecasting, species distribution, healthcare access and health outcomes, and climate change. 32 respondents mentioned broader types of data they would ideally have the improved capability to harness, such as high-resolution images, larger datasets, temporal data, and data from multiple data sources. Finally, respondents envisioned making better use of virtual computing environments and finding it easier to share analyses with a broad audience.

However, for 27 respondents there was not a specific question they would newly be able to tackle with better software, rather they could continue to focus on questions they are currently investigating but do this better, faster, more efficiently, and with a higher volume of data, more data sources, and larger study areas. They anticipate utilizing smoother workflows, faster response from software tools, easier modeling capabilities, more capacity to use exploratory data analysis, and the ability to provide higher quality output. These gains would apply to practically any question, even those that have not yet been asked. Geospatial researchers have the imagination needed to support cutting-edge work; the issue is more likely the lack of functional tools, guidance, funding, and support.

5. Concluding discussion

The primary goal of this survey is to gain an in-depth understanding of the geospatial software community in support of the GSI conceptualization process, as the success of the

GSI will be dependent on both engaging and supporting a variety of geospatial software experts and users with diverse and evolving needs. By analyzing the results of this survey, we have mapped community concerns and limitations of geospatial software to key areas of need for the GSI. Some general issues raised by respondents include concerns about data privacy and security, steep learning curves and insufficient time available for mastering geospatial software, difficulty finding technical support, and often restricted or limited access to high-performance computing resources. When asked to contemplate future directions for geospatial software, key areas envisioned are software capability to handle big data and cover broader study areas, and integrate diverse data. More broadly, with a quicker and easier response time, respondents thought they would be better able to get necessary answers in a reasonable time scale and could perform more replications to solidify conclusions and findings.

Specific findings worth highlighting follow. Availability and access to needed software tools, data, and computational resources are major limiting factors for geospatial-related research and education. Respondents were hindered by overly expensive software, the need to develop custom tools, and reliance on prohibitively expensive datasets. Some, by necessity, depend on free or relatively inexpensive software solutions. Large spatial and temporal coverage gaps frustrate those doing regional or site-specific work, as did outdated and otherwise unsuitable data. Furthermore, suitable data are often stored in poorly compatible formats. Many respondents primarily used desktop computers and noted difficulties gaining access to high-performance computing resources.

The geospatial software learning curve and development rate also presented a set of challenges. Many reported simply not having enough time to learn to use updated tools for their research or work. It was also considered hard to find support or outside expertise to aid in gaining competency. For some respondents, geospatial software developed too fast making it hard to stay current. For others, the rate of development was not fast enough, where recent technical or algorithmic developments were slow to be included in major software applications.

Finally, geospatial software scope and ease of use were reported issues. Many respondents preferred commercial solutions as they were familiar, while others used only open-source tools, believing black-box procedures in commercial software can hinder computational reproducibility. Some preferred more one-size-fits-all geospatial software applications to simplify analysis workflows. Respondents viewed software tools with too broad of coverage as lacking depth with poor fit for specific use cases. Similarly, greater general standardization would help to support greater software interoperability. Also, the current geospatial software ecosystem is perceived as buggy and fragile. Tenuous connections are easily broken by operating system, programming language, and other software updates, often leading to cascading failures.

Overall, this survey was a first step for taking the temperature of the geospatial software community. Knowledge of goals and frustrations in practice can be used to plan a more supportive and better-integrated geospatial software infrastructure. Although the primary goal of this survey was to support the conceptualization of a national GSI, insights from this survey can be of importance to a variety of geospatial constituents.

Acknowledgments

This material is supported in part by the National Science Foundation under the grant number: 1743184. Any opinions, findings, and conclusions or recommendations expressed in the material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors are grateful for insightful comments on the earlier versions of the manuscript received from Editor Stephen Hirtle and multiple anonymous reviewers.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Science Foundation [1743184].

Notes on contributors

Rebecca Vandewalle is a PhD student at the University of Illinois at Urbana-Champaign. Her research interests include spatially-explicit agent-based modeling, spatial network analysis, coupled human and natural systems in emergency contexts, and cyberGIS.

William C. Barley is an Assistant Professor in the Department of Communication at the University of Illinois Urbana-Champaign. His research interests include organizational communication, collaboration and coordination, data representation, and field studies of technology design, adoption, and use.

Anand Padmanabhan is a Research Associate Professor in the Department of Geography and Geographic Information Science at the University of Illinois Urbana-Champaign. His research interests include distributed systems, cyberinfrastructure, and cyberGIS.

Daniel S. Katz is Assistant Director for Scientific Software and Applications at the National Center for Supercomputing Applications and Research Associate Professor in Computer Science, Electrical and Computer Engineering, and the School of Information Sciences at the University of Illinois Urbana-Champaign. His research interests include the interaction of people and software.

Shaowen Wang is a Professor and Head of the Department of Geography and Geographic Information Science; and an Affiliate Professor of the Department of Computer Science, Department of Urban and Regional Planning, and School of Information Sciences at the University of Illinois at Urbana-Champaign. His research interests include geographic information science and systems (GIS), advanced cyberinfrastructure and cyberGIS, complex environmental and geospatial problems, computational and data sciences, high-performance and distributed computing, and spatial analysis and modeling.

ORCID

William C. Barley http://orcid.org/0000-0001-5438-9961
Daniel S. Katz http://orcid.org/0000-0001-5934-7525

Data and codes availability statement

The research uses summary statistics and response counts generated from the Qualtrics survey platform. R scripts developed to create graphs are available via the following link [https://doi.org/10.13012/B2IDB-6834324_V1]. The raw survey data cannot be shared due to the survey consent agreement. The survey protocol is available as a supplementary file.

References

Anselin, L., 2012. From SpaceStat to CyberGIS: twenty years of spatial data analysis software. *International Regional Science Review*, 35 (2), 131–157. doi:10.1177/0160017612438615

Anselin, L. and Rey, S.J., 2012. Spatial econometrics in an age of CyberGIScience. *International Journal of Geographical Information Science*, 26 (12), 2211–2226. doi:10.1080/13658816.2012.664276

Atkins, D.E., *et al*., 2003. *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Washington, D.C.: National Science Foundation.

Bernard, L., *et al*., 2005. Towards an SDI research agenda. In: K. Fullerton, ed. *11th EC GI & GIS Workshop ESDI: Setting the Framework*, Sardinia, June 2005: *Abstracts Handbook*. 147–151.

Bernard, L. and Craglia, M., 2005. SDI - From Spatial Data Infrastructure to Service Driven Infrastructure. *In*: *Research Workshop on Cross-Learning Between Spatial Data Infrastructures and Information Infrastructures*. Enschede, The Netherlands.

Budhathoki, N.R. and Nedović-Budić, Z., 2007. Expanding the Spatial Data Infrastructure Knowledge Base. *In*: H. Onsrud, ed.. *Research and Theory in Advancing Spatial Data Infrastructure Concepts*. Redlands, CA: ESRI Press, 7–31.

Budhathoki, N.R., Bruce, B.C., and Nedović-Budić, Z., 2008. Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal*, 72, 148–160. doi:10.1007/s10708-008-9189-x

Coleman, D.J. and McLaughlin, J.D., 1998. Defining global geospatial data infrastructure (GGDI): components, stakeholders and interfaces. *Geomatica*, 51 (2), 129–143.

Corbin, J. and Strauss, A., 2015. *Basics of qualitative research: techniques and procedures for developing grounded theory. 4th ed*. Los Angeles: SAGE.

Cutcher-Gershenfeld, J., *et al*. 2016. Build it, but will they come? A geoscience cyberinfrastructure baseline analysis. *Data Science Journal*, 15 (8), 1–14. doi:10.5334/dsj-2016-008

Dangermond, J. and Goodchild, M.F., 2020. Building geospatial infrastructure. *Geo-spatial Information Science*, 23 (1), 1–9. doi:10.1080/10095020.2019.1698274

Finholt, T.A. and Birnholtz, J.P., 2006. If We Build It, Will They Come? The Cultural Challenges of Cyberinfrastructure Development. *In*: W.S. Bainbridge and M.C. Roco, eds.. *Managing Nano-Bio-Info-Cogno Innovations: converging Technologies in Society*. The Netherlands: Springer, 89–101.

Gahegan, M., 2018. Our GIS is too small. *The Canadian Geographer/Le Géographe Canadien*, 62 (1), 15–26. doi:10.1111/cag.12434

Harvey, F. and Tulloch, D., 2006. Local-government data sharing: evaluating the foundations of spatial data infrastructures. *International Journal of Geographical Information Science*, 20 (7), 743–768.

Kedron, P., *et al*., 2019. Reproducibility and replicability in geographical analysis. *Geographical Analysis*. doi:10.1111/gean.12221

Kemp, K.K. and Frank, A.U., 1996. Toward consensus on a European GIS curriculum: the international post-graduate course on GIS. *International Journal of Geographical Information Systems*, 10 (4), 477–497. doi:10.1080/02693799608902091

Kim, Y. and Stanton, J.M., 2013. Institutional and individual influences on Scientists' data sharing behaviors: a multilevel analysis. In: *ASIST* 1- 6 November 2013, 2013. Montreal, Quebec, Canada.

Konkol, M., Kray, C., and Pfeiffer, M., 2019. Computational reproducibility in geoscientific papers: insights from a series of studies with geoscientists and a reproduction study. *International Journal of Geographical Information Science*, 33 (2), 408–429.

Kumar, V.D. and Alencar, P., 2016. Software engineering for big data projects: domains, methodologies and gaps. In: *2016 IEEE International Conference on Big Data (Big Data)*. Washington D.C., 2886–2895.
Lee, C.P., Bietz, M.J., and Thayer, A., 2010. Research-driven stakeholders in cyberinfrastructure use and development. In: *2010 International Symposium on Collaborative Technologies and Systems*. Chicago, 163–172.
Li, A., *et al*. 2015. Space-time analysis: concepts, quantitative methods, and future directions. *Annals of the Association of American Geographers*, 105 (5), 891–914. doi:10.1080/00045608.2015.1064510
Madhavji, N.H., Miranskyy, A., and Kontogiannis, K., 2015. Big picture of big data software engineering: with example research challenges. In: *2015 IEEE/ACM 1st International Workshop on Big Data Software Engineering*. Florence, Italy, 11–14.
Masser, I., 2006. What's Special about SDI Related Research? *International Journal of Spatial Data Infrastructures Research*, 1, 14–23.
Masser, I., Rajabifard, A., and Williamson, I., 2008. Spatially enabling governments through SDI implementation. *International Journal of Geographical Information Science*, 22 (1), 5–20. doi:10.1080/13658810601177751
Michener, W.K., *et al*., 2012. Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, 11, 5–15. doi:10.1016/j.ecoinf.2011.08.007
Miller, H.J. and Goodchild, M.F., 2016. Data-driven geography. *GeoJournal*, 80, 449–461. doi:10.1007/s10708-014-9602-6
Murray, A., 2019. Anticipating the next half-century of geographical analysis, *Geographical Analysis*, https://doi.10.1111/gean.12200.
National Research Council, 1991. *Research and Development in the National Mapping Division, USGS: trends and Prospects*. Washington, DC: The National Academies Press. doi:10.17226/10986
National Research Council, 1993. *Toward a Coordinated Spatial Data Infrastructure for the Nation*. Washington, DC: The National Academies Press. doi:10.17226/2105
National Research Council, 1994. *Promoting the National Spatial Data Infrastructure Through Partnerships*. Washington, DC: The National Academies Press. doi:10.17226/4895
National Research Council, 1997. *The Future of Spatial Data and Society: summary of a Workshop*. Washington, DC: The National Academies Press. doi:10.17226/5581
Office of Management and Budget, 2002. *Coordination of Geographic Information, and Related Spatial Data Activities*. OMB Circular A-16. Washington, DC.
Onsrud, H., *et al*. 2005. The Future of the Spatial Information Infrastructure. *In*: R. McMaster and E. L. Usery, eds.. *A Research Agenda for Geographic Information Science*. Boca Raton, Florida: CRC Press, 225–255.
Poore, B.S., 2011. Users as essential contributors to spatial cyberinfrastructures. *Proceedings of the National Academy of Sciences*, 108 (14), 5510–5515. doi:10.1073/pnas.0907677108
Rajabifard, A. and Williamson, I., 2001. Spatial data infrastructures: concept, SDI hierarchy and future directions. In *Proceedings of GEOMATICS'80 Conference*, Tehran, Iran.
Schade, S., *et al*. 2020. Geospatial Information Infrastructures. *In*: H. Guo, M. Goodchild, and A. Annoni, eds.. *Manual of Digital Earth*. Singapore: Springer, 161–190.
Shaw, S.L. and Sui, D., 2018. GIScience for human dynamics research in a changing world. *Transactions in GIS*, 22 (4), 891–899.
Shook, E., *et al*. 2019. Cyber literacy for GIScience: toward formalizing geospatial computing education. *The Professional Geographer*, 71 (2), 221–238. doi:10.1080/00330124.2018.1518720
Stewart, C.A., *et al*., 2011. Technical report: *survey of cyberinfrastructure needs and interests of NSF-funded principal investigators*.
Stewart, C.A., *et al*., 2013. Initial findings from a study of best practices and models for cyberinfrastructure software sustainability. *arXiv preprint arXiv:1309.1817*.
Stocks, K.I., *et al*. 2019. Geoscientists' perspectives on cyberinfrastructure needs: a collection of user scenarios. *Data Science Journal*, 18 (21), 1–15. doi:10.5334/dsj-2019-021

Stodden, V., *et al*. 2015. Enhancing reproducibility for computational methods. *Science*, 354 (6317), 1240–1241. doi:10.1126/science.aah6168

Tracy, S.J., 2020. *Qualitative research methods: collecting evidence, crafting analysis, communicating impact. 2nd ed*. Hoboken, NJ: John Wiley & Sons.

Wang, S., 2010. A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers*, 100 (3), 535–557. doi:10.1080/00045601003791243

Wang, S., 2013. CyberGIS: blueprint for integrated and scalable geospatial software ecosystems. *International Journal of Geographical Information Science*, 27 (11), 2119–2121. doi:10.1080/13658816.2013.841318

Wang, S., 2016. CyberGIS and spatial data science. *GeoJournal*, 81 (6), 965–968.

Wang, S. and Goodchild, M.F., 2019. *CyberGIS for Geospatial Innovation and Discovery*. Dordrecht, Netherlands: Springer. doi:10.1007/978-94-024-1531-5

Wang, S., Liu, Y., and Padmanabhan, A., 2016. Open cyberGIS software for geospatial research and education in the big data era. *SoftwareX*, 5, 1–5. doi:10.1016/j.softx.2015.10.003

Wang, S., Wilkins-Diehr, N.R., and Nyerges, T.L., 2012. CyberGIS—Toward synergistic advancement of cyberinfrastructure and GIScience: a workshop summary. *Journal of Spatial Information Science*, 4, 125–148.

Wang, S. and Zhu, X.-G., 2008. Coupling cyberinfrastructure and geographic information systems to empower ecological and environmental research. *BioScience*, 58 (2), 94–95.

Wright, D.J. and Wang, S., 2011. The emergence of spatial cyberinfrastructure. *Proceedings of the National Academy of Sciences*, 108 (14), 5488–5491. doi:10.1073/pnas.1103051108

Yang, C., *et al*. 2010. Geospatial cyberinfrastructure: past, present and future. *Computers, Environment and Urban Systems*, 34 (4), 264–277. doi:10.1016/j.compenvurbsys.2010.04.001

Yuan, M., 2017. 30 years of *IJGIS*: the changing landscape of geographical information science and the road ahead. *International Journal of Geographical Information Science*, 31 (3), 425–434.

Zimmerman, A., 2007. A socio-technical framework for cyberinfrastructure design. In: e-Social Science Conference, 7-9 October 2007, Ann Arbor, MI.

Zimmerman, A. and Finholt, T.A., 2008. *Report from the TeraGrid Evaluation Study, Part 1: project Findings*. Ann Arbor, MI: Collaboratory for Research on Electronic Work.