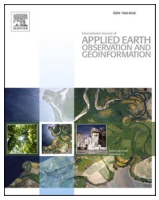
International Journal of Applied Earth Observation and Geoinformation 130 (2024) 103948

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Elevating the RRE Framework for Geospatial Analysis with Visual Programming Platforms: An Exploration with Geospatial Analytics Extension for KNIME

Lingbo Liu [a], Fahui Wang [b,*], Xiaokang Fu [a], Tobias Kötter [b,c], Kevin Sturm [b,c], Weihe Wendy Guan [a,*], Shuming Bao [d]

[a] Center for Geographic Analysis, Harvard University, MA 02138, USA
[b] Department of Geography and Anthropology, Louisiana State University, LA 70803, USA
[c] KNIME GmbH, Körtestr. 10, 10967 Berlin, Germany
[d] China Data Institute, Ann Arbor, MI 48108, USA

ARTICLE INFO

Keywords:
Geospatial Analysis
Reproducibility, replicability, and expandability (RRE)
Visual Programming
Geospatial Analytics Extension for KNIME
Geospatial Knowledge Tree
Spatial Accessibility

ABSTRACT

Reproducibility, replicability, and expandability (RRE) have emerged as fundamental concerns in the realm of scientific research and development. Wherein, devising effective solutions for RRE within geospatial analysis stands out as a particularly critical challenge that demands immediate attention. Although there has been an evolution from basic reproducibility of code and data to a more comprehensive cyberinfrastructure, this integrated solution is still grappling with issues of limited user accessibility, steep learning curves particularly in coding skills, and difficulties in achieving collaboration with other data science platforms This study proposes a framework that combines open-source GIS with visual programming platforms, grounded in principles of standardization and educationalization, to advance the RRE framework in geographic analysis. Using the Geospatial Analytics Extension for KNIME as an example, we demonstrate the platform's adaptability and utility through case studies in a recent textbook with an in-depth illustration of spatial accessibility analysis, specifically via the Generalized Two-Step Floating Catchment Area (G2SFCA) method. Our findings shed light on the transformative potential of such an integrative strategy, offer fresh perspectives for enhancing the RRE in geospatial analysis and craft a well-structured, intuitive, and extensive GIS knowledge tree.

1. Introduction

Reproducibility, replicability, and expandability (RRE) have emerged as fundamental concerns in the realm of scientific research and development (Goodchild, 2021). In the context of geographic analysis, the discourse surrounding RRE encompasses a range of topics including the precise definition of RRE concepts, the categorization of literature in terms of its replicability, the challenges posed by spatiotemporal variability, the difficulties in generalizing replication results, and potential solutions to these issues (Zaragozí et al., 2020; Nüst and Pebesma, 2021). Among these topics, devising effective solutions for RRE within geospatial analysis stands out as a particularly critical and pressing challenge that demands immediate attention (Wilson, 2020).

The development of RRE framework for geospatial analysis aligns closely with the broader evolution of RRE in data science (Sui and Kedron, 2021; Mai, 2022). This evolution is marked by a shift from basic reproducibility of code and data to a more comprehensive Cyberinfrastructure solution. This integrated solution, exemplified by platforms like CyberGIS, encapsulates data, code, metadata, and webGIS functionalities within a comprehensive open-source programming environment (Evans, 2019; Moreau et al., 2023). However, the challenges are substantial, stemming from the rapid expansion of geospatial data, the emergence of new open-source geospatial analysis packages (Shook, 2019), ongoing updates in open-source platforms (Steiniger and Bocher, 2009; Neteler and Mitasova, 2008), and the increasing demand for computational resources (Pijanowski, 2014). These challenges highlight the need for a solution that not only integrates but also simplifies and democratizes access to complex geospatial data and analysis tools.

* Corresponding authors.
E-mail addresses: lingboliu@fas.harvard.edu (L. Liu), fwang@lsu.edu (F. Wang), xiaokang_fu@fas.harvard.edu (X. Fu), tobias.koetter@knime.com (T. Kötter), kevin.sturm@knime.com (K. Sturm), wguan@cga.harvard.edu (W. Wendy Guan), sbao@umich.edu (S. Bao).

https://doi.org/10.1016/j.jag.2024.103948
Received 19 August 2023; Received in revised form 12 April 2024; Accepted 28 May 2024
Available online 31 May 2024

Integrated platforms are grappling with issues of limited user accessibility, steep learning curves particularly in coding skills, and difficulties in achieving seamless collaboration with other data science platforms (Goldberg et al., 2020). Moreover, the rapid pace of GIS knowledge expansion creates a synthesis dilemma. Despite a rich set of GIS tools, studies, and models (O'reilly, 2009), the task of effectively structuring this burgeoning body of knowledge remains formidable. The ultimate objective is not mere aggregation but a structured organization that fosters understanding and catalyzes further research.

In response to the challenges of limited user accessibility, steep learning curves associated with coding, the difficulties in seamless collaboration with other data science platforms, and the rapid expansion of GIS knowledge, this study proposes a framework that combines open-source GIS with visual programming platforms. Grounded in principles of standardization and educationalization, the framework aims to advance the RRE framework in geographic analysis. This solution, exemplified by the Geospatial Analytics Extension for KNIME, utilizes visualization to reduce the learning curve, standardizes geographic analysis modules derived from GIS knowledge tree for efficient integration with data science platforms, and promotes RRE through interdisciplinary collaboration. This approach represents a decentralized and collaborative solution, transitioning from a holistic to a more distributed methodology.

This paper begins with a comprehensive literature review in Section 2, delving into the concepts, current solutions, and the role of open visual programming in the RRE in geospatial analysis. Section 3 introduces the eight key features of KNIME Analytics Platform and its integration with extension development, guided by the 4E principle. We further provide an in-depth comparison between a KNIME workbook and its ArcGIS Pro-focused textbook. Delving into a detailed case study centered on spatial accessibility and leveraging the generalized two-step floating catchment method (G2SFCA), Section 5 underscores the distinct advantages of visual programming for RRE framework. Our exploration sheds light on its transformative capacity to shape a holistic knowledge tree and highlights its indispensable contribution to the evolution of the RRE framework in geospatial analysis.

2. Literature review

Fig. 1 summarizes the framework for literature analysis on the RRE framework in geospatial analysis, which comprises a comparison of concepts like reproduction, replication, reanalysis, generalization, and expansion, and an overview of the key stages in geospatial analysis and the evolution of RRE solutions.

2.1. Concepts of Reproducibility, Replicability, and expandability (RRE)

Reproducibility and replicability (R&R) are widely acknowledged as fundamental to the creation and verification of scientific theories and models (Jasny, 2011), including those in Geographic Information Sciences (GISciences) (Konkol et al., 2019). As well accepted and defined in the 2019 National Academies Report on "Reproducibility and Replicability in Science" (Kedron, 2021); reproducibility is defined as obtaining the same conclusions using the same data and analytical methods, whereas replicability involves the application of original methods to different datasets (National Academies of Sciences, E. and Medicine, Reproducibility and replicability in science., 2019). The sequential positioning of these two terms actually mirrors the scientific argumentation process, transitioning from reproducibility to replicability (Brunsdon and Comber, 2021). However, the reveres terms of replicability and reproducibility may emphasized the transition from replicability to reproducibility, particularly in studies using different datasets to achieve the same conclusions (Wilson, 2020; Stevens, 2017). Such reproduction is closely related to the concept of generalization for the (Halbert, 2022), indicating a lifecycle mantra: reproduce, replication and reproduction (Machicao, 2022).

As there exist different levels of replicability in various type of geospatial research or weak replicability due to the inherent spatiotemporal heterogeneity (Ostermann and Granell, 2016; Wainwright, 2020). The pursuit of generalizable conclusions or the laws in geographic sciences through new data utilization and model expansion is a crucial aspect of scientific research (Dangermond and Goodchild, 2019; Kedron and Holler, 2022), necessitating broader interdisciplinary collaboration and expansion. Expandability refers to the potential to extend existing research materials by reanalyzing it with new functionality (Goeva et al., 2020; Kedron and Frazier, 2022). Therefore, the progression from

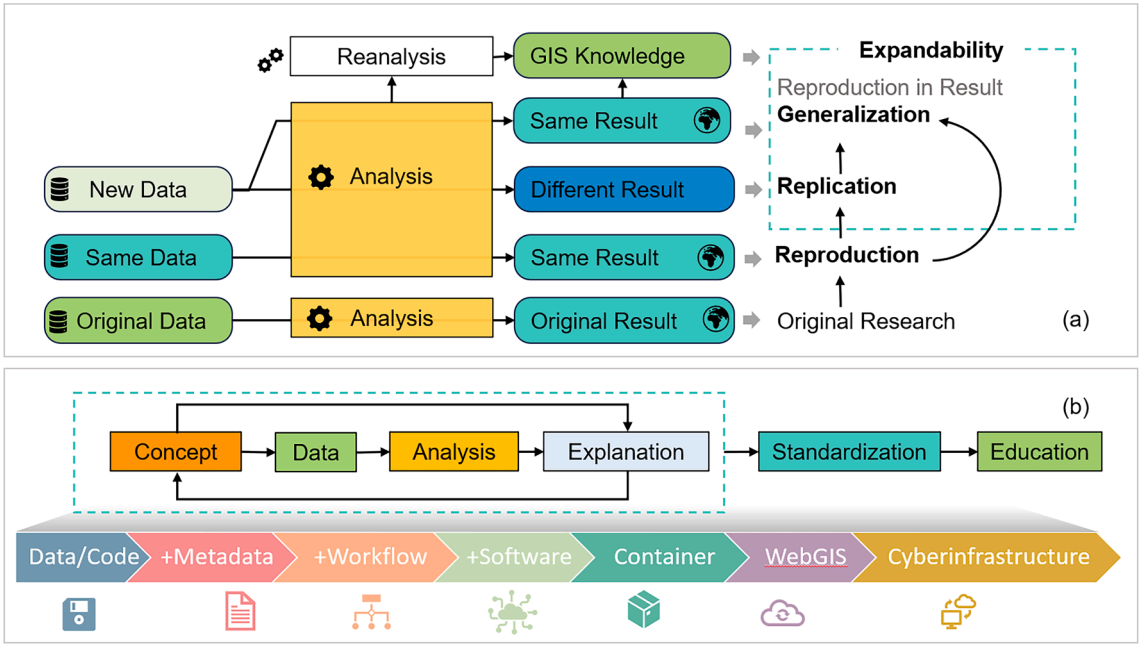


Fig. 1. Literature review summary on the Reproducible, Replicable and Extensibility (RRE) framework for geospatial analysis. (a) Concepts comparison on reproduction, replication, reanalysis, generalization, and expansion, (b) key stages for geospatial analysis and the evolution of RRE solutions.

reproducibility to replicability, to generalization, epitomizes the process of expandability in data research. The extensibility also transcends the confines of geography, extending to interdisciplinary and cross-group collaborations with the ultimate goal of enriching the spectrum of geographical knowledge (Iosifescu Enescu, 2019). Fig. 1(a) illustrates the core concepts of reproducibility, replicability, reanalysis, generalization, and expandability.

2.2. RRE framework and solutions for geospatial analysis

The RRE framework is intrinsically aligned with the entire scientific research process, encompassing concept development, data collection, analysis, and interpretation (Kedron, 2021). This framework is continuously evolving in response to the increasing complexity of data, computational codes, platforms, and research content. Central to the RRE framework are data and analysis, with a growing trend in the geographic RRE field towards utilizing open-source, programming-based code. This shift is largely attributed to its potential to minimize manual operations (Grieve, 2020).

Emphasizing minimum reproducible standards for data and code (Peng, 2011), the RRE framework has expanded its scope to include comprehensive workflows, software versioning, random parameter settings, and adherence to the "10 simple rules" for effective text processing (Sandve, 2013). In this evolving landscape, CyberGIS has emerged as a pivotal development, integrating the strengths of WebGIS and container technology to become a significant trend in RRE platform solutions (Gahegan, 2019; Yin, 2018). CyberGIS is particularly effective in facilitating basic reproduction and replication functions (Wang, 2019).

However, the challenges for CyberGIS as an integrated solution are substantial, which include limited user accessibility, steep learning curves particularly in coding skills, and difficulties in achieving seamless collaboration with other data science platforms (Goldberg et al., 2020). Consequently, it becomes imperative to standardize data and functional modules, along with reducing the learning barriers associated with geospatial analysis modules (Bush, 2020; Lin, 2020). Such measures are crucial not only for promoting interdisciplinary collaboration but also for lowering the learning barriers associated with the platform. Ultimately, these efforts contribute to the advancement of GIS education and the further development of the RRE framework (Kedron, 2021; Leek and Peng, 2015).

2.3. Analysis standardization and reducing learning barriers through visual programming

Visual programming is increasingly acknowledged as an effective method for standardizing tools and reducing the learning barriers in programming education (Eronen, 2002; Dillon et al., 2012; Olsson, 2015). Popular platforms like Scratch (Xinogalos, 2015)and Minecraft (Saito, 2016)have demonstrated their effectiveness in facilitating collaborative and constructivist learning through mind mapping techniques. ESRI ArcGIS Model Builder also exemplifies visual programming for GIS. Facilitating GIS through visual programming could be a crucial pathway for the 'Next Generation of GIS' (Zhu, 2020).

In the realm of data science, open visual programming software like KNIME Analytics Platform, Orange, and RapidMiner emerges as a promising solution (Hirudkar and Sherekar, 2013). These platforms offer scalable modules and visual programming, and reduce the technical barriers for data processing (Egger, 2022). They allow seamless integration of Python or R scripting, facilitate exploration of advanced geographic analysis techniques, such as AI or ML methodologies, thus cater to a broad spectrum of user expertise (Dietz, 2020). Their visual interfaces present a structured, graphical blueprint for project organization, and champion a modular approach ideal for crafting knowledge trees—a marked departure from the linear content in Jupyter notebooks (Chauhan and Sehgal, 2018).

However, most of these open tools are not fundamentally designed for GIS. The pressing question is: can these tools be adapted and expanded to satisfy the distinct requirements of GIS applications and, consequently, bolster the RRE framework?

3. KNIME Analytics platform and geospatial Analytics Extension for RRE framework

3.1. Harnessing KNIME Analytics platform's capabilities for RRE framework in geospatial analysis

The KNIME Analytics Platform presents a distinct advantage in implementing the RRE framework in geospatial analysis (Di Martino, 2024). This is exemplified by the comprehensive workflow depicted in Fig. 2, which illustrates a typical case from reading geographic data to data manipulation, visualization, geographic modeling, machine learning modeling, and the use of R and Python extensions, along with components composed of multiple nodes. The workflow demonstrates KNIME's modular architecture, enabling a complete end-to-end visual programming process. It showcases data flow represented by arrows, the capability for parallel computation of multiple data streams, programming extensibility, and the scalability of combining multiple components. Furthermore, both the workflow and these customized components can be effectively shared on the open KNIME Hub. With the support of KNIME Webportal, workflows can also be transformed into WebGIS applications, enhancing their applicability and reach.

The platform's feature set, dubbed MVP-S5, is concluded as Modular Architecture, Visual Programming, Parallel Computation, Streaming Data, Scripting Extensibility, Scalable Nodes, Server Synchronization, and Seamless Sharing (Fig. 2).

Modular Architecture: Within KNIME, each node signifies a distinct operation, contributing to a high degree of modularity that facilitates the analysis phase of the RRE framework (Jagla et al., 2011). Nodes can be manipulated without disrupting the overall structure, promoting reproducibility by enabling scientists to reuse and rearrange pre-built modules and compare them with modules of other geospatial tools or models.

Visual Programming: KNIME employs visual interfaces to represent programming constructs and enables users to craft programs by manipulating elements graphically rather than textually (Berthold, 2007). This visual approach fosters reproducibility by making workflows intuitive and accessible to both programmers and non-programmers, simplifying the understanding and reproduction of complex geospatial concepts.

Parallel Computation: KNIME's capability to perform multiple computations simultaneously not only enhances computational efficiency but also supports the reproducibility of complex or large-scale tasks. It allows the user to inspect the results of different models simultaneously, and thus fosters a more thorough and efficient analysis process.

Streaming Data: In KNIME, each connection between nodes in the data flow represents the data stream. This transparency in data flow aids reproducibility by providing a clear and explicit map of the data's journey through the algorithm. As geospatial data often involves various scales and formats, data flow visualization helps users understand how data is processed, transformed, and measured.

Scripting Extensibility: This flexibility allows researchers to create reproducible workflows that can be extended to suit unique geospatial analysis needs using custom scripts with Python, R or Java, thereby increases the compatibility of KNIME with new models. It also allow users to build a new standardized extension based on Python packages and KNIME Python API (KNIME Python API, 2023).

Scalable Nodes: Nodes in KNIME can be wrapped and configured as components that work as nodes, and thus enable high scalability for specific functions (Knime, 2023). This scalability supports reproducibility for complex geospatial analyses that involve multiple steps or

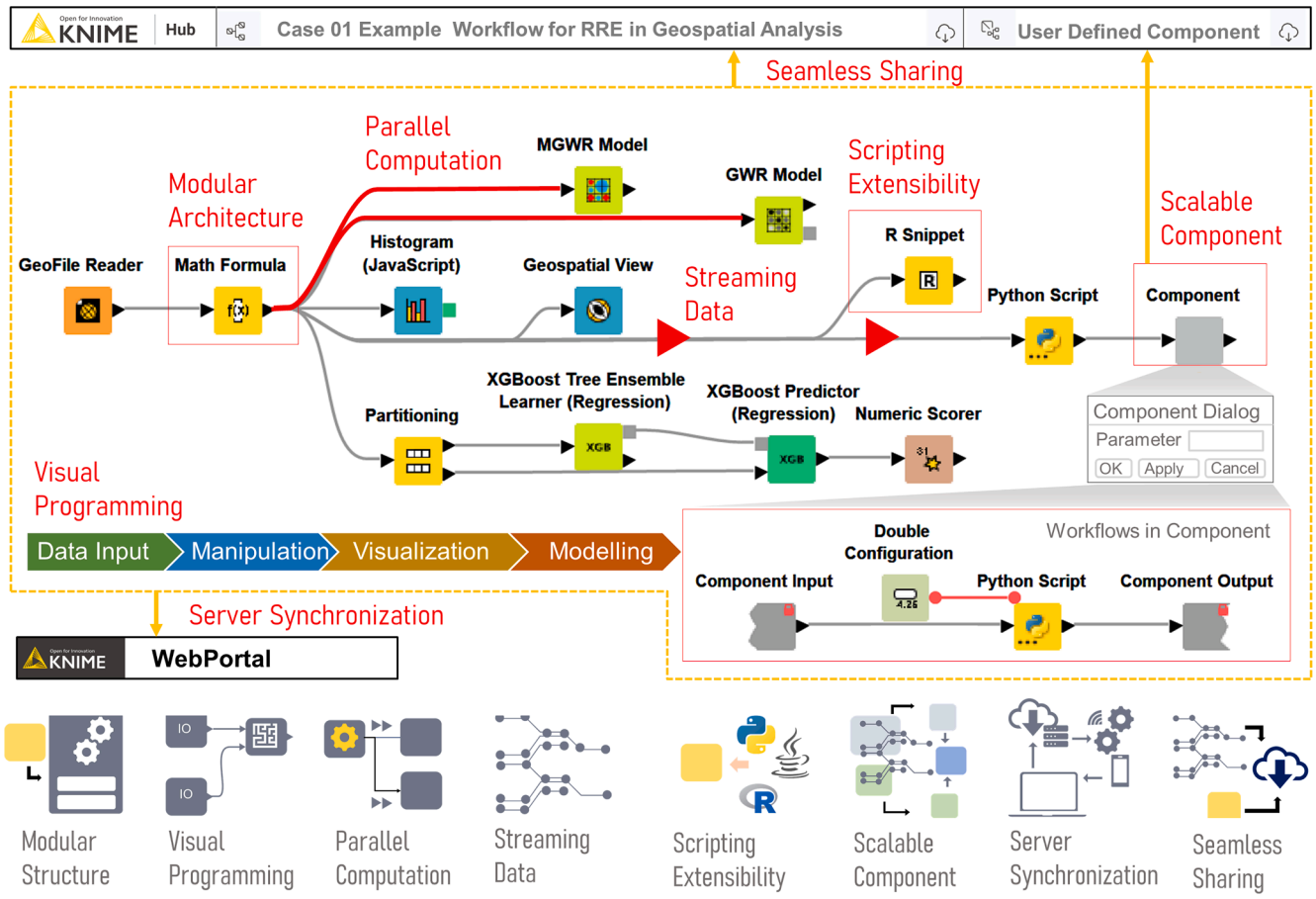


Fig. 2. The MVP-S5 Feature Set of the KNIME Analytics Platform for the RRE Framework of Geospatial Analysis.

large datasets by allowing complex analyses to be encapsulated into reusable components.

Server Synchronization: The cloud service can be directly accessed by the desktop version of KNIME Analytics Platform, and workflows can be run in the Webportal across all platforms with adjustable parameters (Knime, 2023). This capability promotes reproducibility by ensuring that everyone can replicate and explore the research, regardless of their location or device.

Seamless Sharing: The workflows, components and nodes can be easily shared and installed via KNIME Hub, and thus facilitate the effortless and efficient sharing of data, tools, and results among users (University, 2023).

In essence, the MVP-S5 features of the KNIME Analytics Platform intricately align with the tenets of the RRE framework in geospatial analysis. The Modular Architecture and Scalable Nodes enhance the precision in the conceptualization of geospatial tasks, and allow more flexible and customizable structures. Streaming Data ensures clarity in measurement by making data flow transparent and traceable. Meanwhile, the combined capabilities of Visual Programming, Parallel Computation, and Scripting Extensibility support robust analytical processes and simplify complex analyses. Finally, Server Synchronization and Seamless Sharing promote effective communication by ensuring that findings and methodologies are easily accessible and shared across platforms and users.

3.2. Applying the 4E approach in the geospatial Analytics Extension for KNIME: Merging RRE strategy with GIS knowledge tree

In the rapidly evolving field of geospatial analysis, effectively integrating the innovation of data and models from newly published research is key to the successful application of the RRE framework. A systematic and organized approach is essential to ensure efficiency, consistency, and scalability. In this context, we propose the 4E approach – Examine Innovation, Engineer Workflow, Establish Nodes, and Embed Structure (Fig. 3). This strategy, designed intertwine the principles of the RRE framework with the development of the Geospatial Analytics extension for KNIME, not only emphasizes technical innovation in geospatial analysis but also highlights how these innovations are effectively organized, refined, standardized and supplemented within an existing GIS knowledge tree.

Examine Innovation: This initial step involves a comprehensive evaluation of either data innovation or model innovation, framed within the context of an existing geospatial knowledge tree in KNIME. This process is essential for understanding the unique aspects and requirements of the new data or model, which aids users in comprehending the methods for replicability and expandability.

Engineer Workflow: Depending on the nature of the innovation, the workflow engineering process is determined. For model innovations, an initial assessment evaluates whether the model can be constructed using existing KNIME nodes. If feasible, a component encapsulating these nodes is created. If not, Python or R script nodes are embedded within the component to act as the key function extracted from the academic paper or package. For data innovations, it is vital to ascertain whether the data can be retrieved via an API. If so, the data can be directly converted into a node. If API access isn't available, the data will be pushed to Dataverse for subsequent use.

Establish Nodes: After engineering the data or model using the KNIME workflow, these nodes are then established as standard KNIME components or new nodes in the Geospatial Analytics Extension for KNIME. This process involves packaging the nodes into reusable

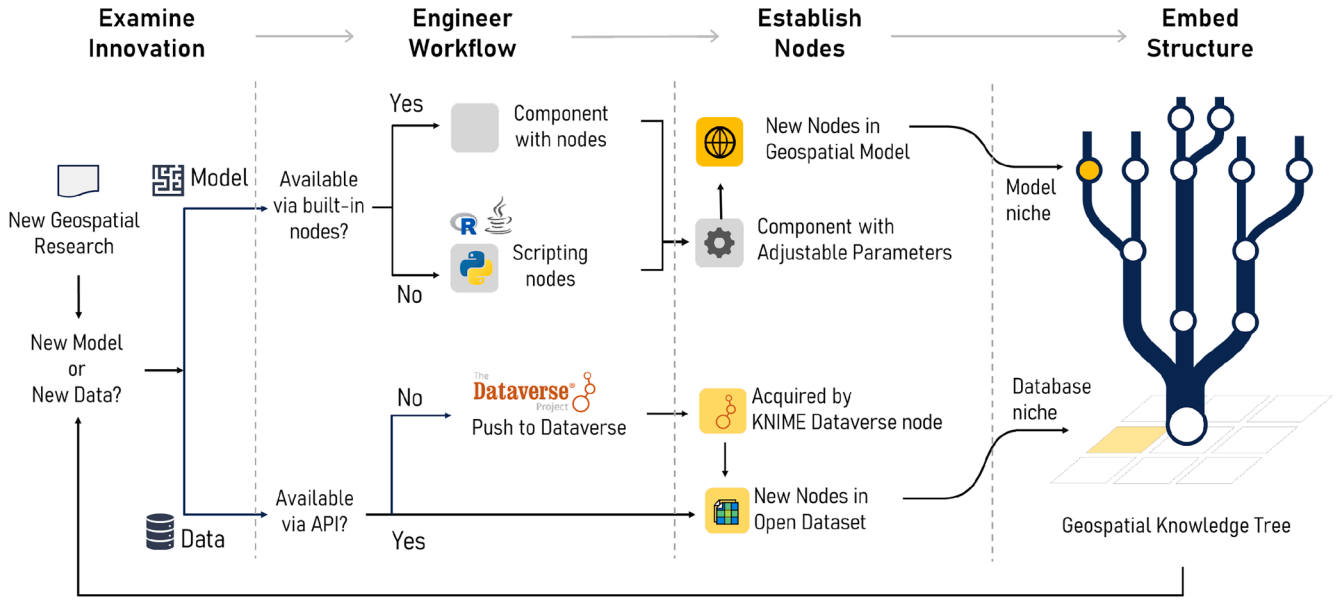


Fig. 3. The 4E Approach (Examine Innovation, Engineer Workflow, Establish Nodes, and Embed Structure) for the Geospatial Analytics Extension for KNIME, based on the RRE Framework with Geospatial Knowledge Tree.

components, which can be shared and reused across different projects or by different teams. By doing so, we promote best practices, ensure consistency, and improve efficiency in geospatial analysis, thereby further strengthening the replicable and reproducible nature of the process. This method not only fosters open-source collaboration but also enhances the reproducibility and expandability of research by sharing nodes across different disciplines.

Embed Structure: The final stage involves embedding the established plugin nodes into the geographical knowledge tree. The organization and management of the geospatial knowledge tree is achieved by building a GIS Nodes repository in GitHub. Similar to other software development in GitHub involving collaborative efforts through crowdsourcing, the core team is responsible for the main development, while user teams can optimize nodes by submitting issues and pull requests. By embedding the plugin nodes into this knowledge tree, they become part of this organized structure, and thus enhance accessibility and usability for users in their geospatial analysis tasks. This structured approach ensures the process's adaptability and readiness to incorporate new nodes as they emerge.

Collectively, these 4E steps delineate the entire journey from case replication (Examine Innovation) to knowledge discovery and organization. This comprehensive process, encapsulated within the expansion of the knowledge tree (Embed GIS Knowledge Tree) and leveraging visual programming workflows (Engineer Workflow) and node standardization (Establish Nodes), successfully extends from specific case replication to broader knowledge generalization. Importantly, it underscores the significance of reproducibility and expandability throughout the entire process.

3.3. Geospatial Analytics Extension for KNIME as an RRE framework for replicating an ArcGIS-Pro centric textbook

Building upon the geospatial knowledge tree outlined in the textbook, *Computational Methods and GIS Application in Social Science* (3rd Edition) (Wang and Liu, 2023); we leveraged the KNIME Analytics platform to recreate case studies and developed geospatial analysis nodes, forming the Geospatial Analytics Extension for KNIME. Currently in its 1.2 version, the extension is composed of 12 categories and 86 nodes. This diverse set of nodes accommodates a wide range of tasks, including data import, cleaning, transformation, analysis, and visualization. The extension is continually developed and updated, with the latest nodes and source code available on GitHub and the KNIME Hub (Liu, 2024).

From a comprehensive review of the nodes incorporated into KNIME workflows, 775 nodes were deployed, stemming from 146 distinct nodes (Liu and Wang, 2023). The top 10 nodes frequently leveraged include Math Formula, Joiner, GroupBy, GeoFile Reader, Linear Regression Learner, Geospatial View, Column Filter, CSV Reader, and Row Filter. Beyond basic functionalities, Spatial Join, Euclidean Distance, Projection, and Dissolve emerge as the primary nodes in the Geospatial Analytics Extension for KNIME (Refer to Fig. 4).

A noteworthy observation is the overlap between geospatial analysis data operations and conventional data science tasks, evident in nodes like Math Formula, Joiner, and GroupBy. This overlap underscores the dual advantages of the KNIME Analytics platform in geospatial contexts. On one hand, it allows seamless integration with KNIME's overarching data science capabilities and amplifies the extensibility of geospatial analyses. For instance, nodes related to deep learning or Explainable AI can be easily integrated and enrich the reproduction and expansion potential of intricate geospatial analyses. On the other hand, a nuanced analysis of the generic data science nodes paves the way for a clearer delineation of the unique attributes of geographic operations or models when replicating a case study for novel research.

3.4. Case study on measuring spatial accessibility within the RRE framework context

In this section, a specific case study is introduced to demonstrate the practical application of the Geospatial Analytics Extension for KNIME. It aims to validate its efficacy through textbook replication and a focused study on spatial accessibility.

Anchored in the RRE framework, we present a case study on measuring spatial accessibility to hospitals in East Baton Rouge Parish (EBRP). The study evaluates access for residents across census block groups to five acute hospitals (Fig. 5). Data of locations and bed sizes of hospitals is from the Louisiana Hospital Association's 2021 directory. Supplementary data concerning census block groups, including population demographics and road networks in EBRP, is informed by the

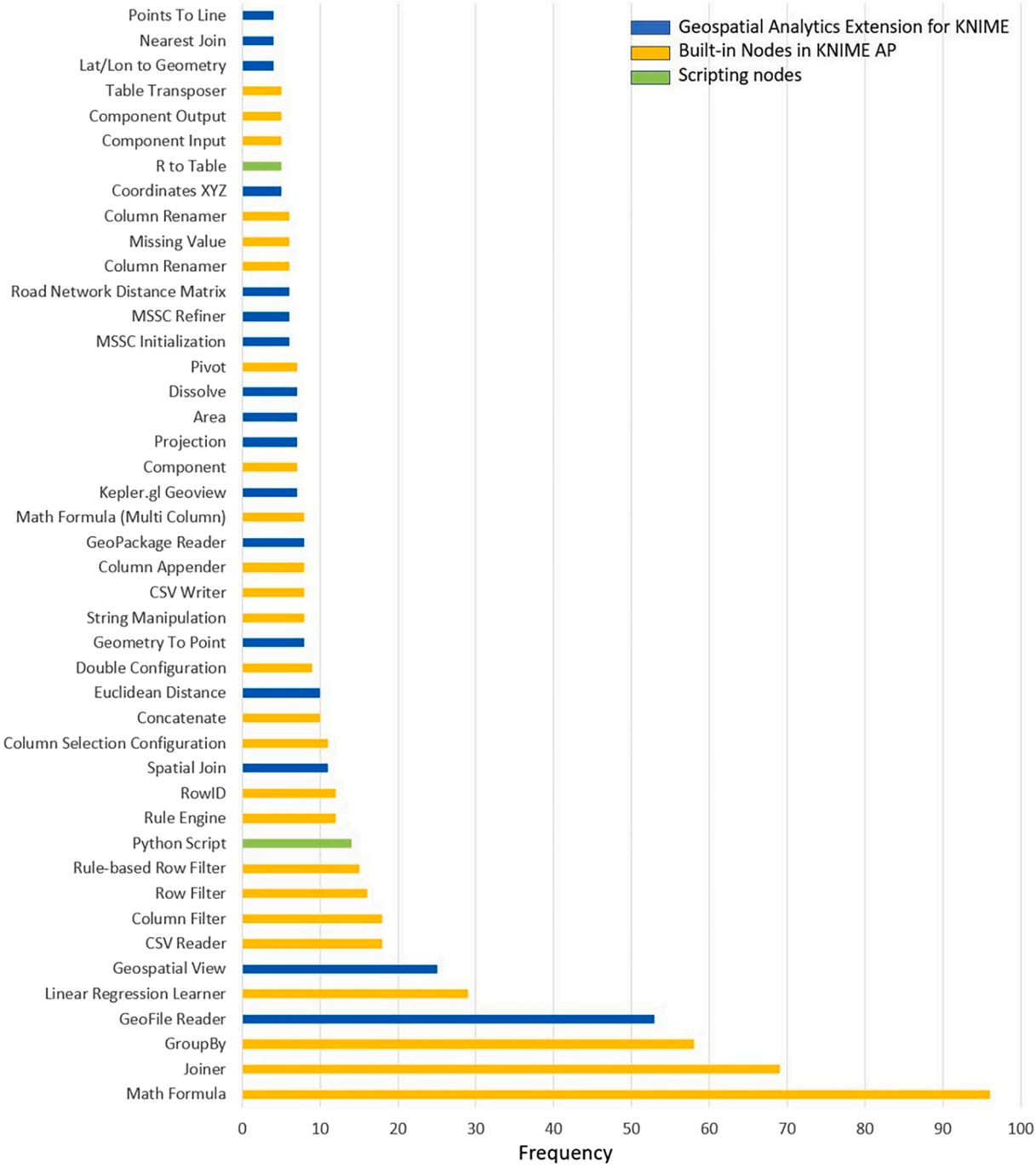


Fig. 4. The frequency of the top 45 nodes in the KNIME lab manual for *Computational Methods and GIS Application in Social Science*.

2020 Census.

3.5. KNIME workflow for G2SFCA based on an RRE framework

Spatial accessibility to healthcare services, measured by the generalized two-step floating catchment method (G2SFCA), requires three elements: demand (D), supply (S), and the spatial impedance or travel cost (d) between them, such as travel time or distance (Wang, 2012; Wang, 2014). The spatial accessibility at demand location i, SA_i, is written as

$$SA_i = \sum_{j=1}^{n} \left[S_j f(d_{ij}) / \sum_{k=1}^{m} (D_k f(d_{kj})) \right] \quad (1)$$

where hospital capacity at supply location j is denoted by S_j, population at demand location k (or i) is denoted by D_k (or D_i), and the interactions between them is a declining function of their physical distance or travel cost d_{kj} (or d_{ij}).

Emergent uncertainties within the G2SFCA model, when contextualized within the RRE framework, are depicted in Fig. 6. They encompass conceptual, measurement, modeling, and communication dimensions. Supplementary Table S1 encapsulates these uncertainties alongside the pertinent data or methodologies applied in the study.

In the G2SFCA implementation within the KNIME platform, the workflow is segmented into six methodical phases, as depicted in Fig. 7. These phases encompass the entirety of the process, from the initial data acquisition and distance computations, through to the visualization of

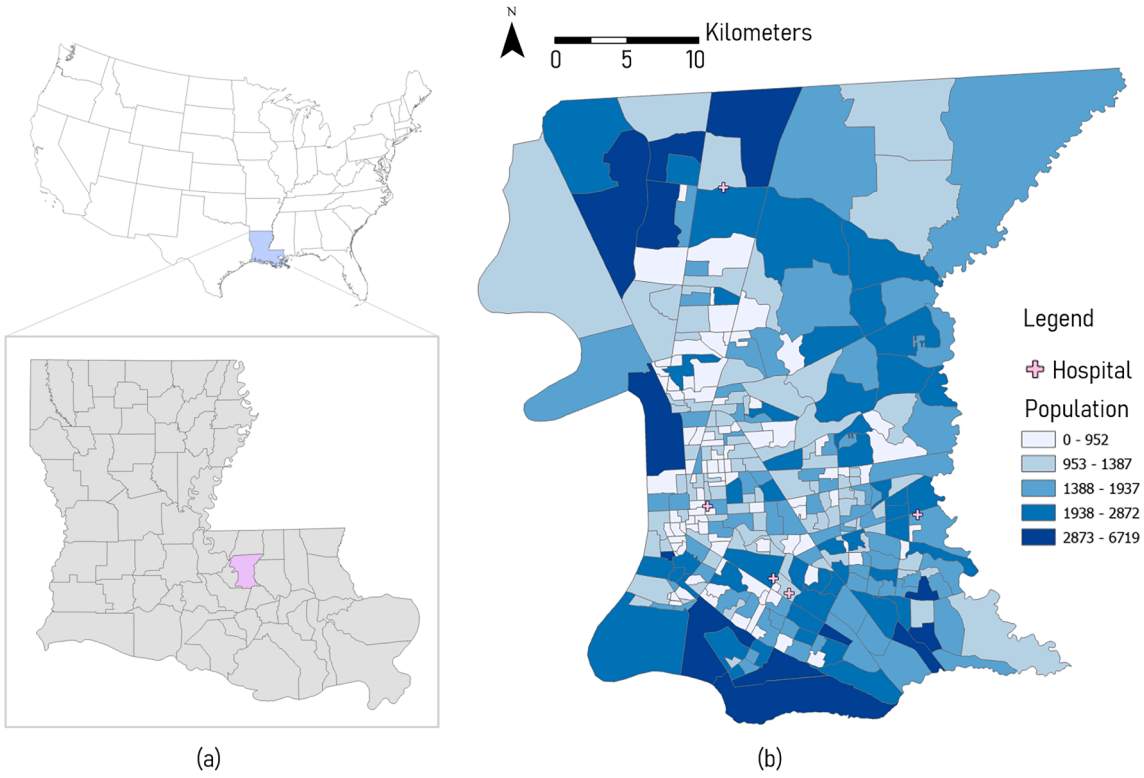


Fig. 5. (a) Location of East Baton Rouge Parish, and (b) Population and hospitals in EBRP.

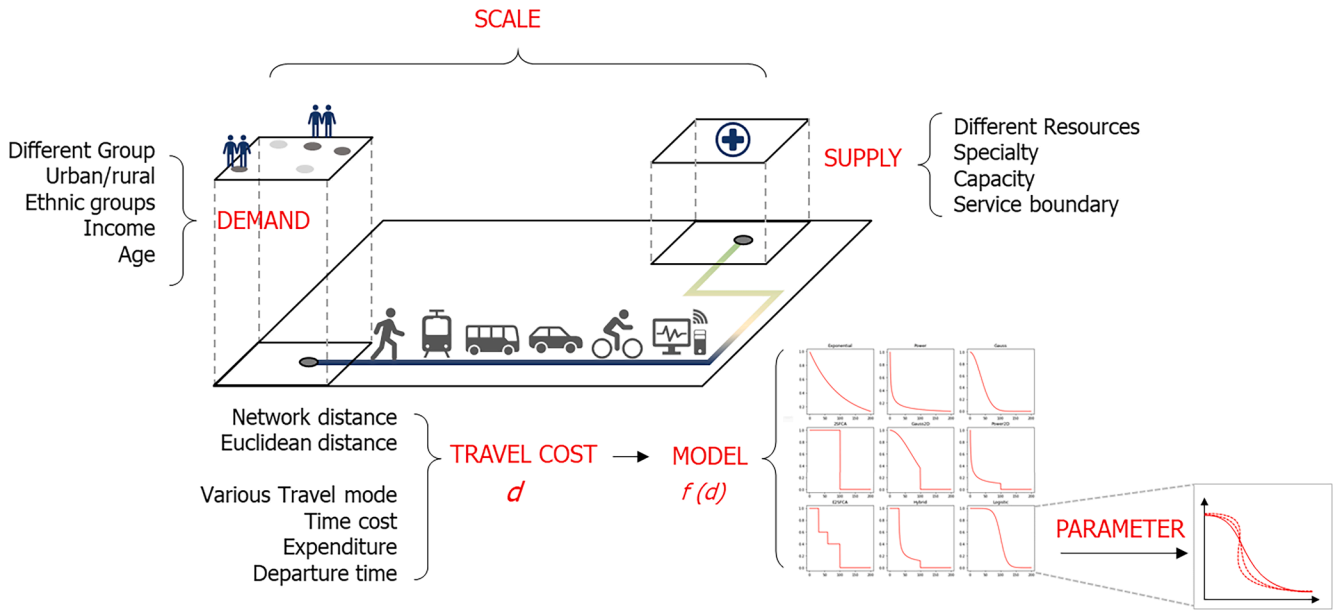


Fig. 6. Illustration of the uncertainty in G2SFCA Model.

accessibility scores (Table 1).

The workflow's design is inherently adaptable and aligns with the principles of the RRE framework and the knowledge tree approach to healthcare accessibility. Such adaptability is evident in the choices offered for distance computations and decay functions (as seen in Nodes 5–9, and Nodes 12–16), and facilitates an easy integration of alternative methodologies and metrics. Supplementary Figures S1 and S2 show the interfaces of the key nodes along with their parameters.

3.6. Component Development, sensitivity analysis and WebPortal interface

With a keen focus on replicability and adaptability, specific nodes tailored for geospatial analytical tasks are condensed into a compact component within the RRE framework. A refined workflow was built to illustrate the process of node standardization for 2SFCA model, G2SFCA model that integrates multiple models and parameters, and a model sensitivity analysis based on looping parameters. The standardization process is primarily executed through Components, which utilize nodes

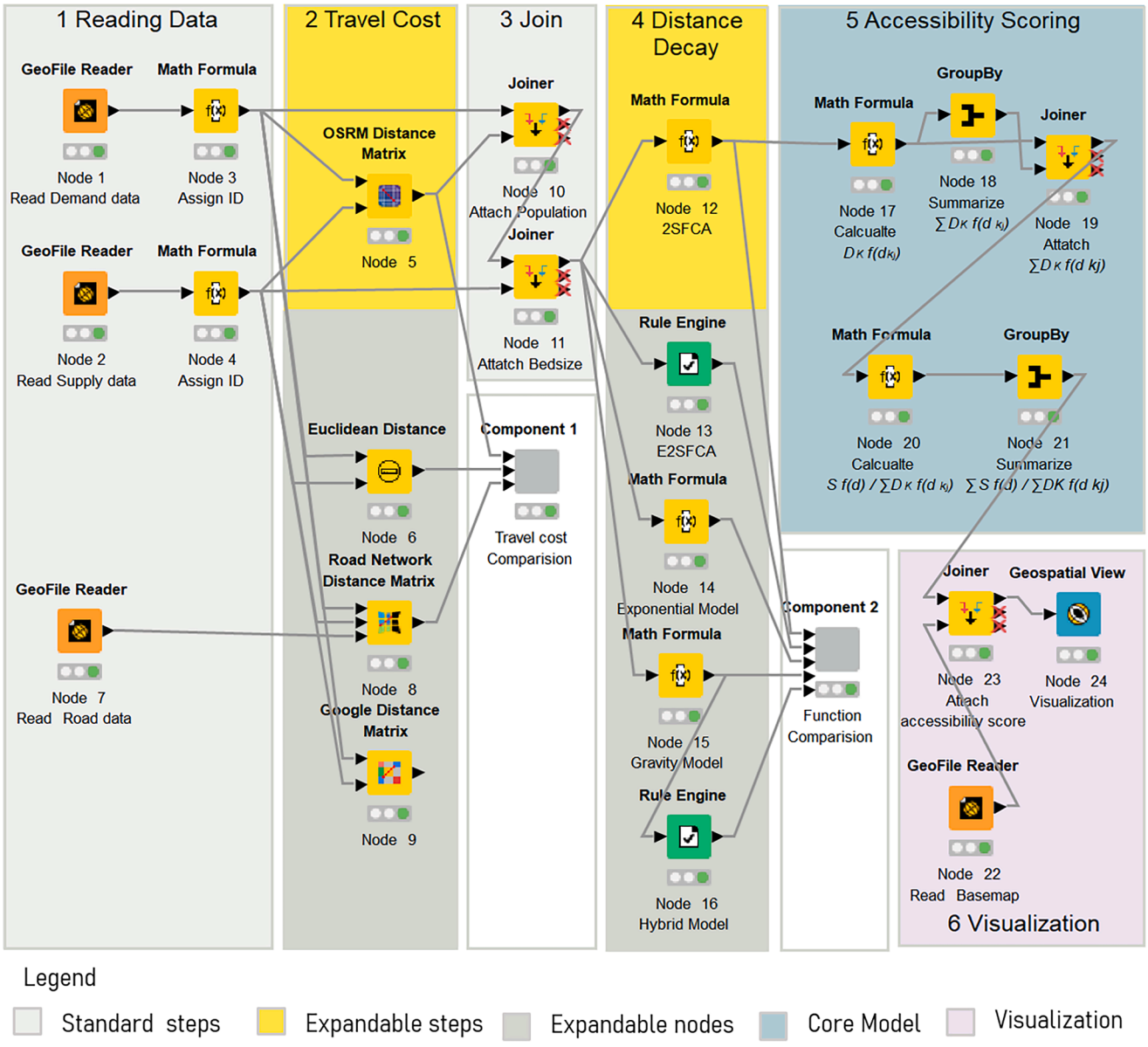


Fig. 7. KNIME workflow for G2SFCA model.

Table 1
Phases and Nodes in the G2SFCA Implementation.

Phase	Description	Nodes
1	Reading Data	Nodes 1–4
2	Calculating Travel Cost (OSRM Distance Matrix, Euclidean distance, Google Distance Matrix, Road Network)	Nodes 5–9
3	Data Joining	Nodes 10–11
4	Distance Decay Effect (2SFCA, E2SFA, Gravity, Exponential, Hybrid)	Nodes 12–16 (Figure S1 and S2)
5	Accessibility Scoring	Nodes 17–21
6	Visualization	Nodes 22–24

from the 2SFCA and G2SFCA models developed in the previous workflow.

Fig. 8(a) illustrates this workflow, which reads the OD matrix data, population, and hospital bed sizes, inherited from Node 11 in the preceding workflow. This workflow efficiently operationalizes the 2SFCA and G2SFCA methodologies as components, drawn from nodes delineated in earlier workflows. To bolster reproducibility, loop nodes are adeptly integrated to enable detailed parameter sensitivity analyses. Enhanced with custom parameter capabilities, this component harmoniously aligns with the behavior and scalability virtues of standard KNIME nodes.

As shown in Fig. 8(b), the sub-workflow inside the 2SFCA Model component amalgamates nodes from the antecedent workflow, notably 2SFCA distance decay parameter (Node 12) and Accessibility Scoring (Nodes 17–21). A double configuration node (Node 2) is added to create an interface to fine-tune the 2SFCA threshold parameter, as shown in Fig. 9(a).

In parallel, Fig. 8(c) presents the G2SFCA component's architectural blueprint. This component is a mosaic of five distinct models, each of which is harmonized with specific parameters. With deft integrations, it consolidates all distance decay models (Nodes 12–16) and Nodes 17–21 from the antecedent workflow. To amplify adaptability, four innovative configuration nodes (Nodes 6–8) have been incorporated to ensure a user-friendly interface for parameter adjustments, which is further depicted in Fig. 9(b).

Upon completing the workflow construction, researchers can utilize

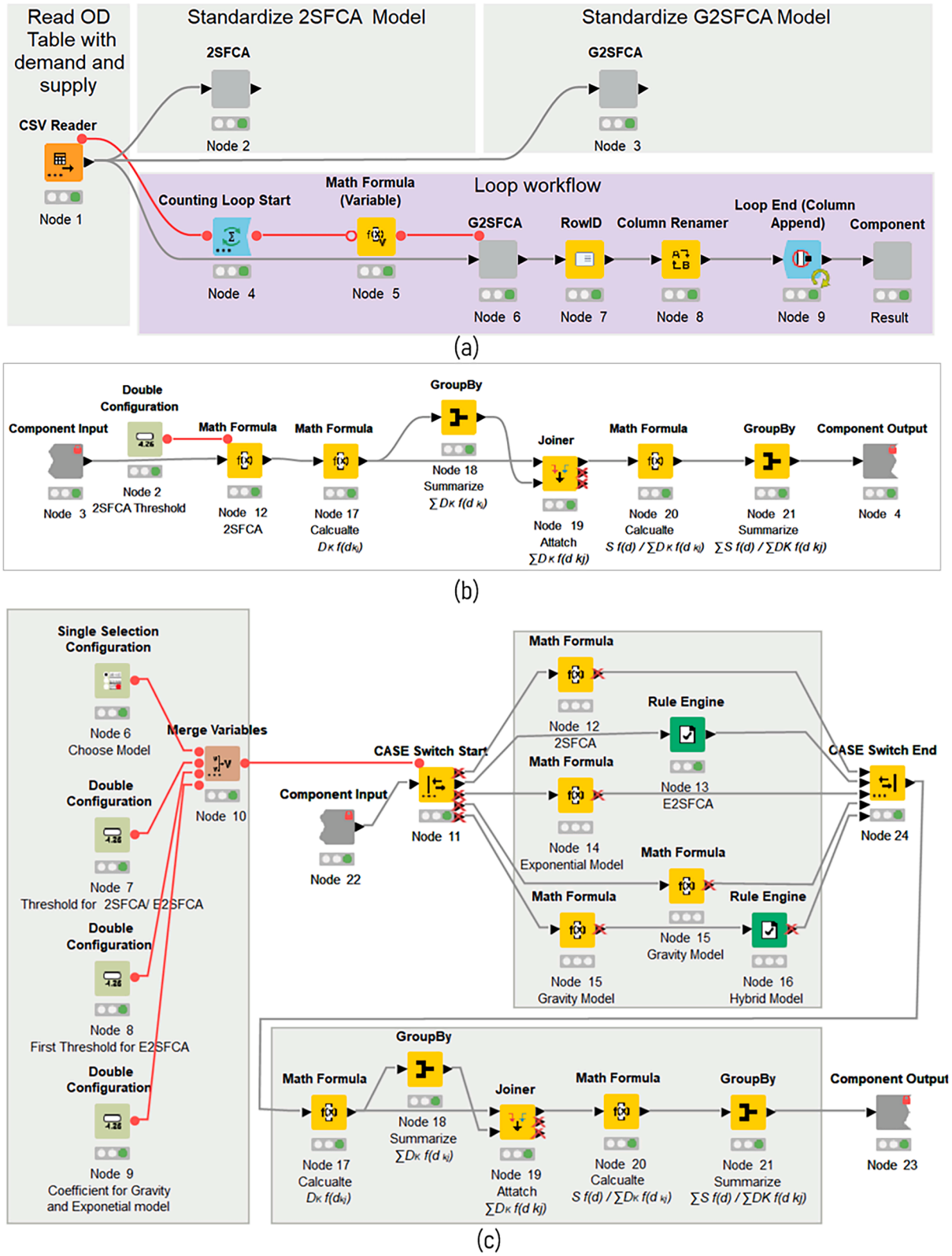


Fig. 8. (a) A refined workflow for using component of 2SFCA and G2SFCA, and its sub-workflows inside the components of (b) 2SFCA Model and (c) G2SFCA.

the integrated Explorer feature within the KNIME Analytics Platform to upload it to the KNIME Hub. This not only facilitates the sharing of data and workflows with other users but also synchronizes the associated metadata and annotations. Furthermore, with minor adjustments, this workflow can be uploaded to a cloud server licensed with the free educational edition via the Explorer feature. Such a configuration empowers users in any internet-enabled location to harness the robust computational resources of the server through the KNIME platform and adjust various parameters as needed. This feature of cloud-based sharing and synchronized computational access offers geospatial researchers significant convenience and advantages to explore large-scale models, advanced algorithms, and complex operational environments, and yet

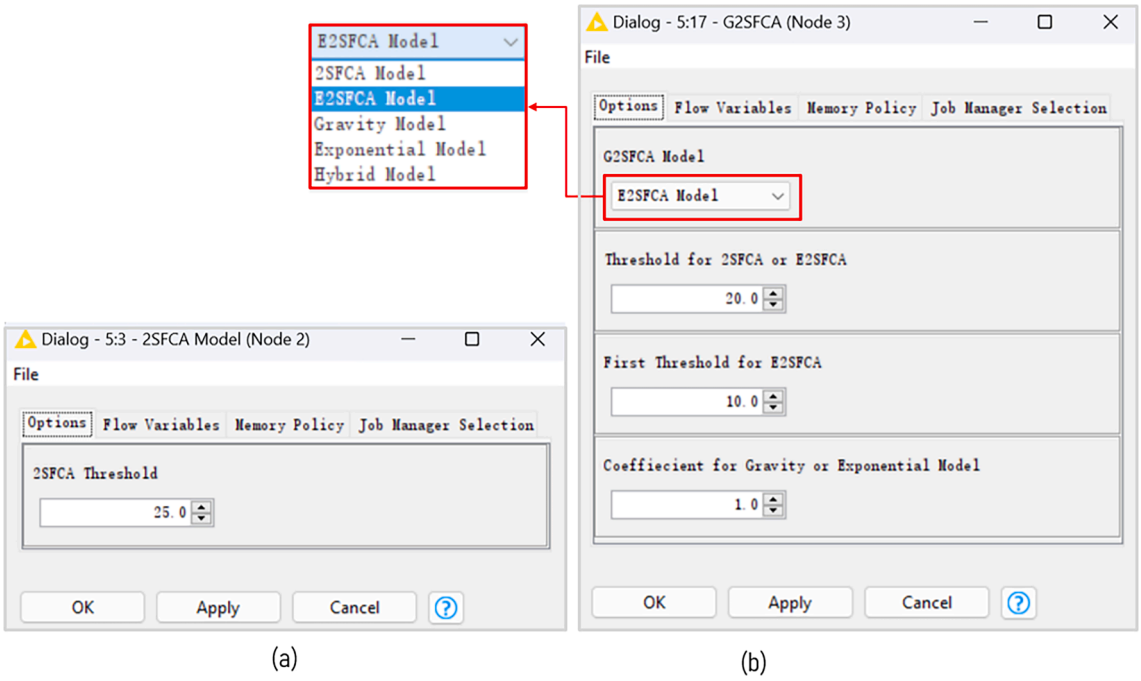


Fig. 9. Interface of the components (a) 2SFCA and (b) G2SFCA.

requires no extensive programming background.

4. Result

Component 1 in the workflow shown in Fig. 7 examines the correlation between different travel times and distances. It incorporates an OSRM node, which utilizes the functions of the Open-Source Routing Machine API (Huber and Rust, 2016), a Road Network Distance Matrix node that operates based on a user-defined driving network and speed information, and a Euclidean distance node for calculating distances using coordinate data. Overall, the different types of travel distance and time demonstrate a high degree of correlation, (Table 2). As shown in Fig. 10(a), the driving time computed by OSRM Distance Matrix is, on average, 4.7 min longer than that derived from road network calculations. Such a result is consistent with a previous study between Google Map API time and road network time based on ArcGIS (Wang and Xu, 2011). Additionally, the travel distance was approximately 144 m more, as shown in Fig. 10(b).

Fig. 11 showcases a comparative analysis of the distance decay curves for five distinct models and their respective parameters, derived from Component 2 in Fig. 6. Notably, different models and their parameters significantly influenced accessibility calculations. Future research can further integrate more distance decay models to expand the current choices available in the G2SFCA model and select the best-fitting distance decay model and its parameters based on actual travel data (Shin and Lee, 2018; Jing, 2023).

The result of the looping workflow for sensitivity analysis further reveals that while different parameters influence accessibility scores, the inherent model trends vary with parameter adjustments. As illustrated in Fig. 12, both the Gravity model and the Hybrid gravity model generally yield consistent ranking trends in terms of accessibility scores. However, as the decay coefficient increases, the disparity in accessibility scores becomes more pronounced. For the 2SFCA model, as the threshold increases, the accessibility scores converge towards the mean, but the rankings exhibit irregular fluctuations. This highlights the sensitivity of the 2SFCA model to spatial interactions with various threshold catchment area sizes. With such model adjustments, the workflow for G2SFCA offers a robust avenue for exploring the characteristics of accessibility models.

Ultimately, all models in the G2SFCA workflow depicted in Fig. 13 (b) can be consolidated into a knowledge tree for spatial accessibility measures for healthcare, as illustrated in Fig. 13(a), set against a broader context (Liu, 2022). As new methodologies emerge, this knowledge tree will undergo continuous expansion. For instance, new models, such as the *2-step virtual catchment area (2SVCA) method* that is tailored to measure the accessibility via a virtual space (internet) such as telehealth access can be integrated on the left side of the diagram (Liu, 2023).

Table 2
Correlations between spatial costs by different methods.

	OSRM distance	Euclidean distance	Road network time	Road network distance
OSRM time	0.977	0.963	0.949	0.974
OSRM distance		0.985	0.95	0.983
Euclidean distance			0.951	0.980
Road network time				0.977

Note: OSRM Time (Distance) and Road Network Time (Distance) represent the travel time (distance) calculated by driving by OSRM node and Road Network Distance Matrix nodes.

5. Concluding comments

This paper reports our effort to delve into the potential of integrating visual programming platforms, specifically KNIME, with GIS functionalities to fortify the RRE framework and GIS knowledge tree. This exploration highlights the tangible advantages of this integration, especially in the context of increasing challenges faced by the GIS community.

The Geospatial Analytics Extension for KNIME uses a workflow-based platform to mitigate technical challenges often faced in GIS. Its modular and visual programming capabilities pave the way for a more intuitive and accessible geospatial analysis experience. The 4E approach provides a systematic methodology for integrating new geospatial innovations into established platforms. As the GIS field rapidly evolves,

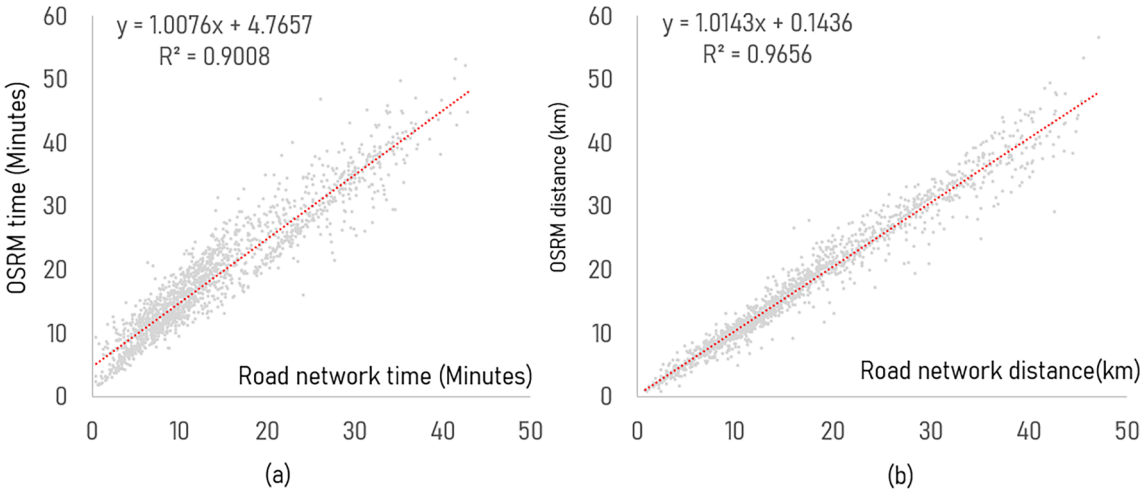


Fig. 10. Scatter plot of (a) travel time and (b) distance between Road network and OSRM node.

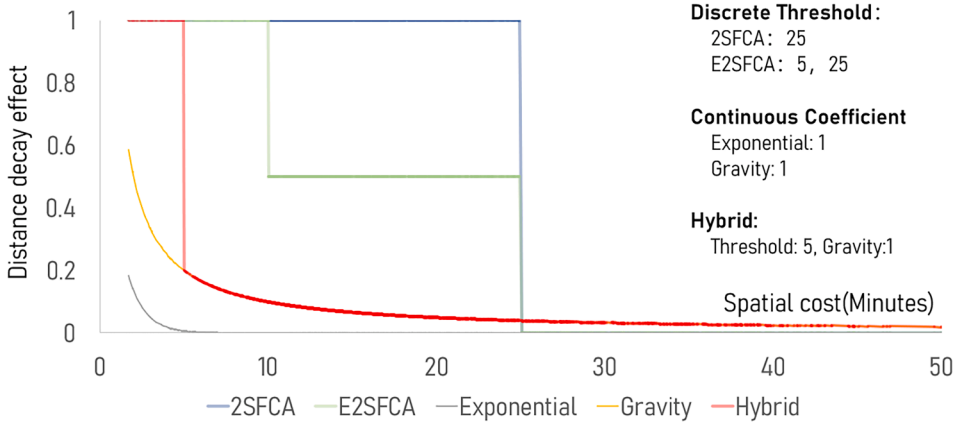


Fig. 11. Distance decay effect by different models and corresponding parameters.

such structured strategies are crucial for assimilating innovations and ensuring adaptability in the GIS community.

The case studies implemented in KNIME in a recent textbook, especially one focusing on spatial accessibility measures by the G2SFCA model, demonstrate the versatility and efficacy of the Geospatial Analytics Extension for KNIME. A particularly intriguing revelation is the intersection of geospatial analysis operations with foundational data science tasks. This intersection underscores the dual capabilities of platforms like KNIME. On one hand, it facilitates specialized geospatial functionalities. On the other hand, it integrates seamlessly with core data science competencies. Such synergy is pivotal as it not only augments the breadth of geospatial studies but also promotes interdisciplinary collaborations.

Furthermore, this investigation has illuminated some of the inherent challenges facing the GIS community, particularly in the realm of open-source GIS. As the GIS domain burgeons with a plethora of tools, datasets, and models, it calls for a structured synthesis. The challenge is not just about collating this vast reservoir of knowledge but organizing it in a manner that is intuitive, accessible, and conducive to further research. Our in-depth illustration of creating a knowledge tree for healthcare accessibility integrated with nodes for workflows in visual programing software is a step in this direction, and provides a structured framework that can be continually expanded and refined as new methodologies and insights emerge.

Our work on the Geospatial Analytics Extension for KNIME shows the transformative potential of integrating visual programming platforms and standardized GIS functionalities, enhances interdisciplinary cooperation within the RRE framework. While this integration marks a significant step in advancing the RRE in geospatial analyses, it requires further development. Expanding GIS functionalities beyond textbook tools, incorporating nodes for remote sensing imagery, and conducting more empirical studies to validate KNIME's effectiveness in geospatial analysis are key areas for future focus. This work sets the stage for advancing geospatial analytics through improved integration of GIS and visual programming.

CRediT authorship contribution statement

Lingbo Liu: Writing – original draft, Conceptualization. **Fahui Wang:** Writing – review & editing, Data curation, Conceptualization. **Xiaokang Fu:** Methodology. **Tobias Kötter:** Methodology. **Kevin Sturm:** Methodology. **Weihe Wendy Guan:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Shuming Bao:** Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and workflow are publicly shared in a folder named

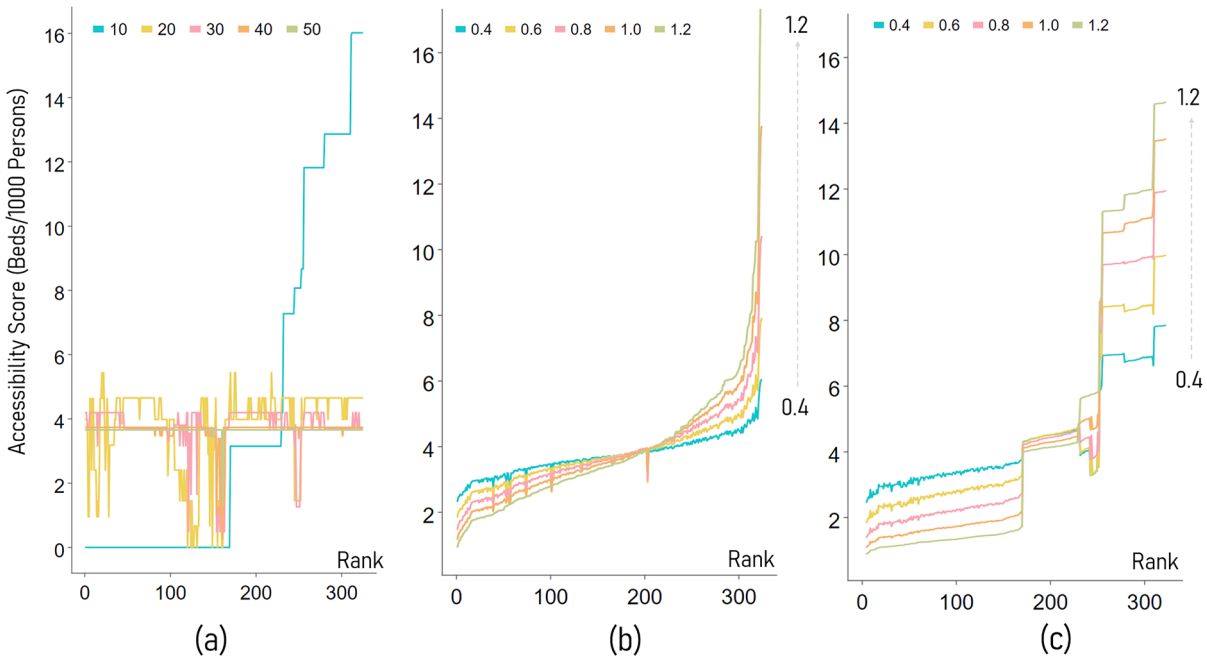


Fig. 12. Sensitive analysis for (a) 2SFCA with varying threshold from 10 to 45 min ranking by the ascending score while threshold = 10 min, (b) Gravity model with varying coefficient from 0.4 to 1.2, and (c) Hybrid model with a initial fixed threshold of 10 min and Gravity model with varying coefficient from 0.4 to 1.2, ranking by the ascending score while coefficient = 1.2.

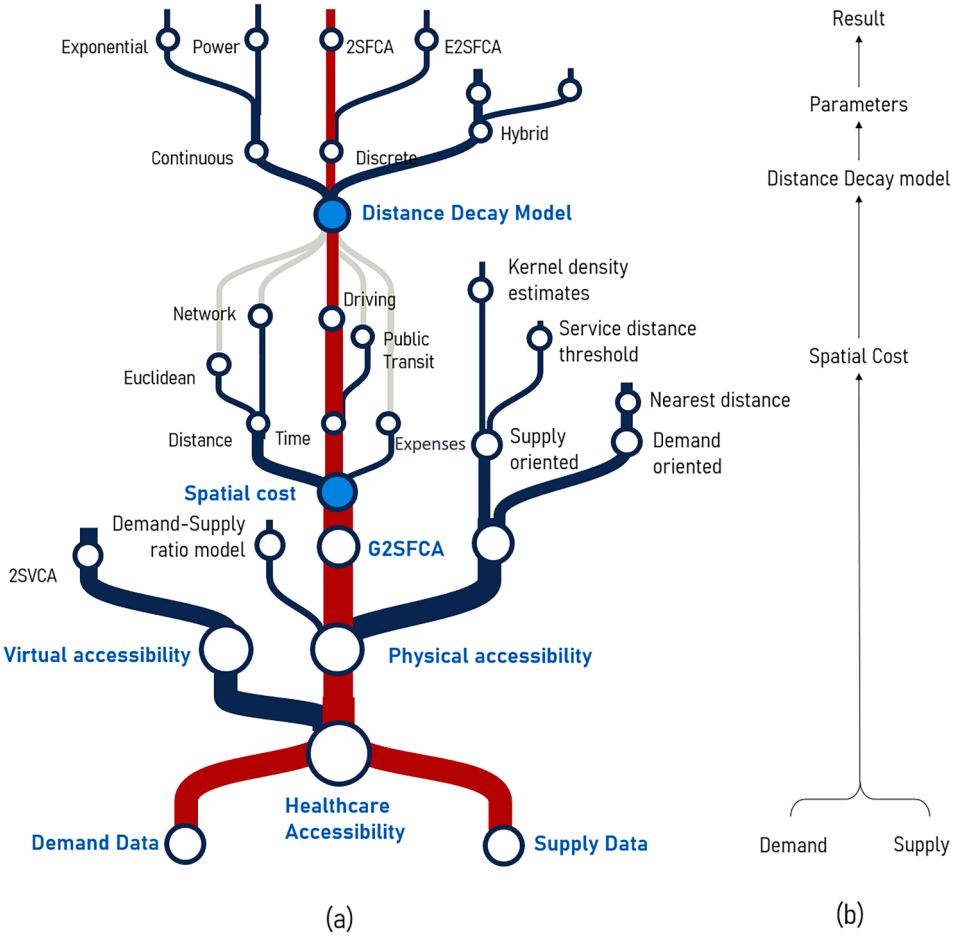


Fig. 13. (a) Knowledge tree of healthcare accessibility, and (b) conceptual workflow for G2SFCA model.

Healthcare Accessibility by G2SFCA Model under KNIME Hub. https://hub.knime.com/center for geographic analysis at Harvard university/spaces/Geospatial Analytics Examples/latest/

Acknowledgement

This work is partially funded by NSF grant #1841403.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2024.103948.

References

Berthold, M.R., et al., 2007. KNIME: The Konstanz Information Miner. Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007). Springer.

Brunsdon, C., Comber, A., 2021. Opening practice: supporting reproducibility and critical spatial data science. J. Geogr. Syst. 23 (4), 477–496.

Bush, R., et al., 2020. Perspectives on Data Reproducibility and Replicability in Paleoclimate and Climate Science. Harvard Data Sci. Rev. 2 (4).

Chauhan, C. and S. Sehgal. *Sentiment Classification for Mobile Reviews using KNIME.* in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*. 2018.

Dangermond, J., Goodchild, M.F., 2019. Building geospatial infrastructure. Geo-Spatial Information Science 23 (1), 1–9.

Di Martino, S., et al., 2024. A visual-based toolkit to support mobility data analytics. Expert Syst. Appl. 238, 121949.

Dietz, C., et al., 2020. Integration of the ImageJ Ecosystem in KNIME Analytics Platform. Front. Comput. Sci. 2.

Dillon, E., M. Anderson-Herzog, and M. Brown, *Studying the Novice's Perception of Visual Vs. Command Line Programming Tools in CS1.* Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2012. **56**(1): p. 605-609.

Egger, R., 2022. Software and Tools. In: Applied Data Science in Tourism. Springer, pp. 547–588.

Eronen, P.J., et al., 2002. Kid's club as an ICT-based learning laboratory. Inform. Educ. 1 (1), 61–72.

Evans, M.R., *Enabling spatial big data via CyberGIS: Challenges and opportunities.* 2019: p. 143-170.

Gahegan, M., *Reproducible Geocomputation: an open or shut case?* , in *GeoComputation 2019*. 2019: AUCKLAND.

Goeva, A., S. Stoudt, and A. Trisovic, *Toward reproducible and extensible research: from values to action.* 2020.

Goldberg, D.W., Bowlick, F.J., Stine, P.E., 2020. Virtualization in CyberGIS instruction: lessons learned constructing a private cloud to support development and delivery of a WebGIS course. J. Geogr. High. Educ. 45 (1), 128–154.

Goodchild, M.F., et al., 2021. Introduction: Forum on Reproducibility and Replicability in Geography. Ann. Am. Assoc. Geogr. 111 (5), 1271–1274.

S. Grieve, F.C.S.M., *Reproducible topographic analysis.* 2020. **23**: p. 339-367.

Halbert, M.D., 2022. Advancing Reproducibility at the NSF. Computer 55 (8), 31–39.

Hirudkar, A.M., Sherekar, S., 2013. Comparative analysis of data mining tools and techniques for evaluating performance of database system. Int. J. Comput. Sci. Appl. 6 (2), 232–237.

Huber, S., Rust, C., 2016. Calculate travel time and distance with OpenStreetMap data using the Open Source Routing Machine (OSRM). Stata J. 16 (2), 416–423.

Iosifescu Enescu, I., et al., 2019. Open Science, Knowledge Sharing and Reproducibility as Drivers for the Adoption of Foss4g in Environmental Research. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLII-4/W14, 107–110.

Jagla, B., Wiswedel, B., Coppee, J.Y., 2011. Extending KNIME for next-generation sequencing data analysis. Bioinformatics 27 (20), 2907–2909.

Jasny, B.R., et al., 2011. Again, and Again, and Again. Science 334 (6060), 1225.

Jing, C.B., et al., 2023. Trajectory big data reveals spatial disparity of healthcare accessibility at the residential neighborhood scale. Cities 133.

Kedron, P., et al., 2021. Reproducibility and replicability: opportunities and challenges for geospatial research. Int. J. Geogr. Inf. Sci. 35 (3), 427–445.

Kedron, P., et al., 2021. Reproducibility and Replicability in Geographical Analysis. Geogr. Anal. 53 (1), 135–147.

Kedron, P., Frazier, A.E., 2022. How to Improve the Reproducibility, Replicability, and Extensibility of Remote Sensing Research. Remote Sens. (Basel) 14. https://doi.org/10.3390/rs14215471.

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

KNIME Python API. 2023 [cited 2023 11-3]; Available from: https://knime-python.readthedocs.io.

KNIME. *KNIME WebPortal User Guide*. KNIME Server 4.16 2023 2023-07 [cited 2023 12-3]; Available from: https://docs.knime.com/2023-07/webportal_user_guide/index.html.

KNIME. *KNIME Components Guide*. KNIME Analytics Platform 5.1 2023 2023-07 [cited 2023 2023-12-3]; Available from: https://docs.knime.com/2023-07/analytics_platform_components_guide.

Konkol, M., Kray, C., Pfeiffer, M., 2019. Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study. Int. J. Geogr. Inf. Sci. 33 (2), 408–429.

Leek, J.T., Peng, R.D., 2015. Reproducible research can still be wrong: adopting a prevention approach. Proc. Natl. Acad. Sci. U S A 112 (6), 1645–1646.

Lin, X., 2020. Learning Lessons on Reproducibility and Replicability in Large Scale Genome-Wide Association Studies. Harvard Data Science Review 2 (4).

Liu, L.B., et al., 2022. Multiscale Effects of Multimodal Public Facilities Accessibility on Housing Prices Based on MGWR: A Case Study of Wuhan, China. ISPRS Int. J. Geo Inf. 11 (1).

Liu, L.B., et al., 2023. Refining 2SVCA method for measuring telehealth accessibility of primary care physicians in Baton Rouge, Louisiana. Cities 138.

Liu, L., et al., 2024. *Geospatial Analytics Extension for KNIME.* Softwarex 25, 101627.

Liu, L. and F. Wang, *Computational Methods and GIS Applications in Social Science-Lab Manual*. 2023: CRC Press.

Machicao, J., et al., 2022. Mitigation Strategies to Improve Reproducibility of Poverty Estimations From Remote Sensing Images Using Deep Learning. Earth Space Sci. 9 (8).

Mai, G., et al., 2022. Symbolic and subsymbolic GeoAI: Geospatial knowledge graphs and spatially explicit machine learning. Trans. GIS 26 (8), 3118–3124.

Moreau, D., Wiebels, K., Boettiger, C., 2023. Containers for computational reproducibility. Nat. Rev. Methods Primers 3 (1), 50.

National Academies of Sciences, E. and Medicine, *Reproducibility and replicability in science*. 2019: National Academies Press.

Neteler, M., Mitasova, H., 2008. Open Source software and GIS. In: Open Source Gis. Springer, pp. 1–6.

Nüst, D., Pebesma, E., 2021. Practical Reproducibility in Geography and Geosciences. Ann. Am. Assoc. Geogr. 111 (5), 1300–1310.

M. Olsson, P.M.J.C., *Visualisation and Gamification of e-Learning and Programming Education.* Electronic Journal of e-Learning, 2015. **13**: p. 441-454.

O'reilly, T., *What is web 2.0*. 2009: " O'Reilly Media, Inc.".

Ostermann, F.O., Granell, C., 2016. Advancing Science with VGI: Reproducibility and Replicability of Recent Studies using VGI. Trans. GIS 21 (2), 224–237.

Peng, R.D., 2011. Reproducible Research in Computational Science. Science 334 (6060), 1226–1227.

Pijanowski, B.C., et al., 2014. A big data urban growth simulation at a national scale: Configuring the GIS and neural network based Land Transformation Model to run in a High Performance Computing (HPC) environment. Environ. Model. Softw. 51, 250–268.

D. Saito, H.W.Y.F., *Influence of the Programming Environment on Programming Education.* Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education, 2016.

Sandve, G.K., et al., 2013. Ten simple rules for reproducible computational research. PLoS Comput Biol 9 (10), e1003285.

Shin, K., Lee, T., 2018. Improving the measurement of the Korean emergency medical System's spatial accessibility. Appl. Geogr. 100, 30–38.

Shook, E., et al., 2019. Cyber Literacy for GIScience: Toward Formalizing Geospatial Computing Education. Professional Geographer 71 (2), 221–238.

Steiniger, S., Bocher, E., 2009. An overview on current free and open source desktop GIS developments. Int. J. Geogr. Inf. Sci. 23 (10), 1345–1370.

Stevens, J.R., 2017. Replicability and Reproducibility in Comparative Psychology. Front Psychol 8, 862.

Sui, D., Kedron, P., 2021. Reproducibility and Replicability in the Context of the Contested Identities of Geography. Ann. Am. Assoc. Geogr. 111 (5), 1275–1283.

University, C.f.G.A.a.H. *Geospatial Analytics Examples*. 2023.

Wainwright, J., 2020. Is Critical Human Geography Research Replicable? Ann. Am. Assoc. Geogr. 111 (5), 1284–1290.

Wang, F., 2012. Measurement, Optimization, and Impact of Health Care Accessibility: A Methodological Review. Ann. Assoc. Am. Geogr. 102 (5), 1104–1112.

Wang, F., 2014. Quantitative methods and socio-economic applications in GIS. Crc Press.

Wang, F. and L. Liu, *Computational Methods and GIS Applications in Social Science*. 2023: CRC Press.

Wang, F., Xu, Y., 2011. Estimating O-D travel time matrix by Google Maps API: implementation, advantages, and implications. Ann. GIS 17 (4), 199–209.

Wang, S., *CyberGIS for geospatial discovery and innovation*. 2019.

Wilson, J.P., et al., 2020. A Five-Star Guide for Achieving Replicability and Reproducibility When Working with GIS Software and Algorithms. Ann. Am. Assoc. Geogr. 111 (5), 1311–1317.

S. Xinogalos, M.S.C.M., Microworlds, games, animations, mobile apps, puzzle editors and more: What is important for an introductory programming environment? Educ. Inform. Technol. 2015. **22**, 145–176.

Yin, D., et al., 2018. CyberGIS-Jupyter for reproducible and scalable geospatial analytics. Concurrency Comput. Pract. Experien. 31 (11), e5040.

Zaragozí, B.M., Trilles, S., Navarro-Carrión, J.T., 2020. Leveraging container technologies in a GIScience project: a perspective from open reproducible research. ISPRS Int. J. Geo Inf. 9 (3), 138.

Zhu, A.X., et al., 2020. Next generation of GIS: must be easy. Ann. GIS 27 (1), 71–86.