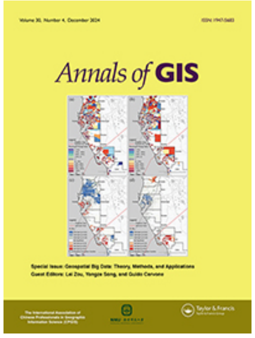

Annals of GIS

ISSN: 1947-5683 (Print) 1947-5691 (Online) Journal homepage: www.tandfonline.com/journals/tagi20

Geospatial big data: theory, methods, and applications

Lei Zou, Yongze Song & Guido Cervone

To cite this article: Lei Zou, Yongze Song & Guido Cervone (2024) Geospatial big data: theory, methods, and applications, Annals of GIS, 30:4, 411-415, DOI: 10.1080/19475683.2024.2419749

To link to this article: https://doi.org/10.1080/19475683.2024.2419749

Published online: 25 Oct 2024.

ANNALS OF GIS
2024, VOL. 30, NO. 4, 411–415
https://doi.org/10.1080/19475683.2024.2419749

EDITORIAL

OPEN ACCESS Check for updates

Geospatial big data: theory, methods, and applications

1. Introduction

The recent digital technological revolution has enabled the creation and collection of large, diverse geospatial data from satellite and drone images, social and news media, web applications, trajectories of GPS-equipped devices, onsite and portable smart sensors, surveillance vehicles, and crowdsourcing platforms. These data, referred to as geospatial big data, offer a unique lens to rapidly, timely, and multi-dimensionally observe the dynamics of human behaviours, urban development, environmental systems, and their interplay. Consequently, there is a growing interest from academia, government, organizations, and the public in leveraging geospatial big data to observe the social, urban, and environmental phenomena, understand the integrated systems, and support decision-making towards a smart, equal, and sustainable society and ecosystem.

Previous research and practices have developed novel theories and algorithms to obtain insights from geospatial big data and apply them in solving real-world challenges. However, existing literature also identifies limitations and emerging challenges in the theory, methods, and applications of geospatial big data. Initially, the definition of geospatial big data is conceptualized from the four Vs of big data (volume, velocity, variety, and veracity), but the definition is vague and inconsistent. In addition, although the use of geospatial big data has been ubiquitous, the methodologies of collecting, analysing, and visualizing geospatial big data, especially those newly emerged data, remain technically challenging and are usually subjectively determined by researchers or practitioners. This leads to reproducibility and replicability challenges. Meanwhile, biases, uncertainties, and ethical concerns should be considered in geospatial big data collection, analysis, and sharing but are paid less attention. Finally, more interdisciplinary explorations of geospatial big data applications in different fields are needed to fully unleash their potential.

The purpose of this special issue is to invite scholars to share investigations on obtaining novel geospatial big data sets, developing advanced models to analyse them, and using them for innovative applications. The editorial consists of three components. Section 2 introduces the concepts, analysis methods, and applications of geospatial big data. Section 3 overviews the articles included in this special issue and their key findings. Finally, some challenges and future directions of geospatial big data analysis are epitomized in Section 4. We believe the special issue will help identify achieved milestones and existing challenges, as well as shed light on research needs in the theory, methods, and applications of geospatial big data analytics.

2. Geospatial big data

Geospatial big data refers to massive datasets with spatial or geographical components, collected from diverse sources. These data capture dynamic spatial phenomena, representing the locations, movements, and activities of people, objects, or natural systems over time. The emergence of new technologies, such as smart mobile devices, various satellites, drones, and Internet of Things (IoT) sensors, has contributed significantly to the growth of geospatial big data. These datasets are often multidimensional, containing spatial, temporal, thematic, and relational information, which is essential for understanding complex geospatial patterns and interactions across various domains.

Common types of geospatial big data include satellite and aerial imagery, street view images, mobile phone data, social media posts, sensor networks, and transportation data. Satellite imagery provides continuous global coverage at different resolutions, enabling environmental and urban monitoring, land use and land cover observations, and resource and event management. Street view images, collected through platforms such as Google Street View, offer a ground-level perspective of urban and rural environments, which can be used for analysing built-environment conditions, accessibility, street design, and landscape features. These images are also increasingly utilized for training computer vision models to automatically detect street-level information like building types, vegetation, or sidewalk accessibility. Mobile phone data, such as Call Detail Records (CDR) and GPS trajectory of smartphones, offer valuable insights into human mobility and interaction patterns, including commuting behaviours and population movements. Additional mobility sources include GPS trajectories from vehicles, Bluetooth travel data, and public transportation records, all of which enable detailed analysis of traffic flows, urban mobility, and transportation planning. Social media data capture location-based real-time social interactions, awareness, and sentiment, providing insights into people's behaviours and experiences across different communities. Geotagged and geo-enriched posts from platforms like Twitter/X, Instagram, and Facebook help identify patterns in public interests, responses to events, and social movements. Platforms such as Yelp and Google Reviews contribute to understanding sense of place by crowdsourcing opinions about businesses, infrastructure, and urban services. Other examples of geospatial big data include environmental sensor data (e.g. air quality, noise levels, temperature), which are essential for monitoring environmental conditions, managing resources, and informing policies on land development.

The processing and analysis of geospatial big data require a diverse range of techniques to extract meaningful insights from complex, noisy, and sometimes biased datasets, such as (1) natural language processing (NLP), (2) image processing & computer vision, (3) machine learning (ML) & artificial intelligence (AI), (4) cyber infrastructure, (5) data fusion & visualization, (6) network analysis & modelling, (7) spatial analysis & modelling, (8) temporal analysis & modelling, and (9) interactions, association, & causality. These nine categories of techniques provide a robust framework for analysing geospatial big data.

NLP, including pre-trained large language models (LLMs), allows the extraction of spatial information from unstructured text data such as social media posts, news reports, and textual databases (e.g. Wikipedia). Image Processing and Computer Vision analyse satellite, aerial, street view, and user-posted images to detect patterns, classify objects (e.g. land cover, trees, and humans), and monitor their changes (e.g. flooded areas after a storm). The advances and incorporation of ML and AI techniques, including neural networks, significantly improves the extraction of deep, diverse information from text and image datasets. ML/AI also support the encoding of geographical objects and prediction of spatial patterns, e.g. coastal wetland losses. The recent advances in generative AI further facilitate geographical knowledge discovery and spatial prediction using geospatial big data. Cyberinfrastructure, encompassing High-Performance Computing (HPC) systems, cloud computing, and collaborative platforms, ensures that massive datasets can be stored, processed, and shared efficiently. For example, Google Earth Engine (GEE) provides cloud-based access to massive remote sensing images, scalable computing resources, and powerful APIs, enabling users to perform large-scale spatial-temporal analysis of geospatial processes and phenomena without computational hurdles. Data Fusion integrates information from multiple sources with varying formats and resolutions. Geo-visualizations present the results through interactive maps, dashboards, virtual reality (VR), and augmented reality (AR). Network Analysis and Modelling explores connections and interactions within spatial systems, such as transportation, ecological, or social networks. Spatial Analysis and Modeling employ geostatistical techniques and spatial-explicit models to explore relationships, detect hot spots, and model spatial patterns and variations across scales. Temporal Analysis and Modeling track changes over time, enabling dynamic monitoring and forecasting of trends, such as urban growth, disease spread, or environmental shifts. Finally, Interactions, Association, and Causality methods investigate complex cause-effect relationships between variables, supporting data-driven policy-making and decision-making processes.

The growing availability of geospatial big data has introduced new applications that were previously impossible or impractical. For example, geospatial big data enable real-time monitoring of urban mobility by tracking traffic flows, public transportation usage, and pedestrian movement, helping cities optimize public services. These data also facilitate efficient disaster management through advancing real-time hazard detection, evacuation tracking, early warning systems, and emergency rescue operations, improving response and recovery strategies. Continuous environmental monitoring is made possible through long-term, large-scale satellite data tracking deforestation, air and water quality, and climate shifts. By analysing large amounts of traffic patterns, transportation networks can be optimized to enhance transit safety and efficiency. Health analytics benefit from geospatial big data by mapping disease outbreaks through crowdsourcing, assessing healthcare accessibility considering real-world visitations, and examining the impact of environmental factors such as pollution and green spaces on mental and physical well-being through long-term analysis. Besides, agriculture leverages big remote sensing and field sensing data for crop monitoring to optimize irrigation and detect pest to ensure food sustainability. Additionally, socio-economic analysis utilizes population mobility patterns together with socioeconomic status to reveal social inequity. These are just a few popular application domains of geospatial big data. Recently, the use of geospatial big data has opened new opportunities in interdisciplinary fields, such as tourism, security, energy, and sports.

3. Overview of this special issue

This special issue contains seven articles developing novel methods to collect, process, and analyse geospatial big data to delineate human activities (Hu et al. 2024; Liao, Kwan, and Liu 2024; Liu and Yuan 2024), model social interactions with the environment (Liu et al. 2024), disasters (Li, Qiang, and Cervone 2024), and health crises (Lin et al. 2024), and predict environmental characteristics (Zhao et al. 2024).

The first paper analyzes human mobility big data to investigate human activity fragmentation across cyber and physical spaces (Liao, Kwan, and Liu 2024). It leverages massive Call Detail Records (CDR) and Uniform Resource Locator (URL) data from millions of mobile phone users in Jilin Province, China. The results discover the inverse relationship between physical human activity fragmentation (PHAF) and cyber human activity fragmentation (CHAF). The investigation also reveals that physical fragmentation is influenced by the built environment, while cyber activities are less bound by spatial constraints. Meanwhile, Cyber activity fragmentations are more influenced by physical fragmentations and such relationship is stable across different cities. A unique aspect of the study is the introduction of mobile phone location and web browsing data, making it possible to explore how human activities fragment and distribute in both cyber and physical spaces.

The second article explores urban mobility patterns using Bluetooth traffic flow big data to identify communities (Liu and Yuan 2024). This is achieved through a customized community detection algorithm based on Dynamic Time Warping (DTW). The use of Bluetooth data provides a fine-grained, time-sensitive perspective on how individuals move through urban spaces. Using the city of Austin in Texas, US as an example, this research identifies three types of communities on each day of a week. The novelty of this study lies in its ability to detect non-geographic communities, or regions that share similar mobility patterns but are not necessarily connected spatially. This expands the concept of community beyond traditional spatial interactions.

The third study focuses on Arctic accessibility using a special type of mobility big data (Hu et al. 2024). It compares the Arctic Transport Accessibility Model (ATAM) with real-world ship trajectory data from 2013 to 2020 collected by the Protection of the Arctic Marine Environment Arctic Ship Traffic Data (ASTD) project. It provides valuable insights into evolving maritime traffic patterns. A unique feature of this analysis is its focus on validating the ATAM model, which highlights discrepancies between modelled and observed ship speeds and routes. The results find that ATAM underestimates accessibility due to outdated model parameters, particularly regarding sea ice thickness and vessel travel speeds. This study also underscores the challenges of modelling rapidly changing environments like the Arctic, where climate change is transforming sea ice coverage.

The fourth investigation utilizes mobility big data to reveal the relationship between human activities and environmental conditions (Liu et al. 2024). Specifically, it examines the relationship between nitrogen dioxide (NO_2) levels and human mobility in Southeast Asia during the COVID-19 pandemic, utilizing mobility data from Facebook and Apple and satellite observed NO_2 data from Sentinel-5/TROPOMI (TROPOspheric Monitoring Instrument). The Multi-Layer Perceptron (MLP) was employed to predict NO_2 levels and explore influencing factors like travel modes and meteorological conditions using SHapley Additive exPlanations (SHAP) values. A major limitation noted is the complexity of isolating the effects of mobility on NO_2, considering that other variables like weather also play significant roles. It is worth mentioning that this is one of the few studies conducting long-term (two years) analyses of human mobility changes during the COVID-19 pandemic and their impacts on the environment.

The fifth study uses human mobility data to analyse evacuation patterns during Hurricane Ian, which made landfall in Florida, US in 2022 (Li, Qiang, and Cervone 2024). Leveraging SafeGraph data on population flows and evacuation orders, the study tracks how individuals responded to government-mandated evacuations. The findings reveal that most populations in evacuation zones complied with the orders. The study also highlights challenges in monitoring real-time compliance across large geographic areas and suggests incorporating additional data sources, such as social media check-ins or public transportation usage. This study offers a large-scale, high-resolution analysis of evacuation behaviours that are difficult to capture with traditional data sources. The analytical workflows offer actionable tools for policymakers to assess the effectiveness of evacuation plans, identify communities that failed to evacuate on time, and improve future disaster response and evacuation strategies.

The sixth paper offers a systematic review of the use of geospatial big data to understand human dynamics and interactions during a health crisis – the COVID-19 pandemic (Lin et al. 2024). This review synthesizes various publications to identify commonly used geospatial big data in COVID-19 studies, including mobility from different sources, social media, policy evaluations, and health records. This review provides an overarching view of how geospatial big data can be analysed to quantify human behaviours at multiple spatial and temporal scales. It also compares the geographical disparities of detected relationships between human behaviours and COVID-19 spread. The findings can inform localized public health responses. In the meantime, the paper identifies several challenges in analysing geospatial big data, including data privacy concerns, the need for standardized data-sharing protocols, and the need for long-term spatial-temporal analysis and modelling. Unique to this study is its comparison of multiple geospatial big datasets and research findings across different global regions.

The seventh article focuses on developing tools to process and analyse geospatial big data. This study introduces the intelligent Soil Land Inference Model (iSoLIM), a novel spatial prediction tool designed to handle large-scale geospatial data (Zhao et al. 2024). iSoLIM employs a similarity-based prediction approach. Two challenges in spatial prediction are (1) integrating diverse datasets from different sources and (2) the substantial computation time. To address these challenges, iSoLIM incorporates an expandable knowledge representation schema that allows for more seamless integration of multi-source data and adopts a block division-based parallelization strategy. Using the Raffelson watershed in Wisconsin, US as the first case study, iSoLIM successfully predicts the spatial distribution of soil types by utilizing expert descriptions, field samples, and historical thematic maps, achieving an accuracy of 67.5%, which is higher than using single-source knowledge. Another case study uses iSoLIM to predict groundwater levels in the Wisconsin River watershed in the U.S. using large-volume covariate data. It demonstrates that iSoLIM can significantly reduce computation time compared to existing software such as ArcGIS.

4. Challenges and looking forward

Studies included in this special issue illustrate the vast potential of geospatial big data for understanding human activities, interactions with the environment and disasters, and spatial prediction. Concurrently, these analyses acknowledge the challenges and propose forward-thinking solutions. This editorial will mainly discuss three of them, including geo-privacy and ethics concerns, multimodal GeoAI, and scale issues in geospatial big data collection, analysis, sharing, and applications.

First, the growing use of geospatial big data introduces significant concerns related to geoprivacy and ethics, as it often involves the collection and processing of sensitive location data from individuals. Issues such as unauthorized tracking, re-identification of anonymized data, and the potential misuse of personal location information raise privacy risks. Furthermore, the lack of clear frameworks and ethical guidelines for managing geospatial data presents challenges for organizations and researchers. Future research could focus on developing geoprivacy-preserving algorithms, such as differential privacy or federated learning, that allow insights to be extracted without disclosing individual locational information. Additionally, transparent consent mechanisms and ethical data governance models are essential to ensure that users understand how their location data will be used. Future research can also explore bias detection and mitigation techniques to ensure that geospatial models are used fairly and equitably, especially in sensitive areas.

Another promising direction is developing multimodal GeoAI to analyse geospatial big data with diverse sources, formats, availability, and resolutions. Multimodal GeoAI refers to the integration of multiple data modalities – such as satellite imagery, sensor data, social media posts, and text-based information – using artificial intelligence techniques to generate location-based insights and solve geospatial problems. The primary challenges include data heterogeneity, alignment across modalities, and the development of models capable of handling diverse data formats. Recent advances in AI has started to tackle this problem. For instance, the Contrastive Language-Image Pretraining (CLIP) model developed by OpenAI is designed to understand and process images and texts simultaneously. Future research can focus on developing advanced fusion techniques that can seamlessly combine disparate geospatial data sources to generate insights. Additionally, the rise of deep learning techniques calls for the design of scalable and interpretable models that can process multimodal geospatial data efficiently. Future efforts can target on addressing the computational challenges posed by multimodal geospatial datasets.

Finally, scale remains a critical issue in geospatial big data, as data collected at different spatial and temporal resolutions can affect analysis outcomes and introduce uncertainty. Spatial scale mismatches, such as combining fine-grained urban mobility data with coarse-scale environmental data, can result in misleading conclusions. Temporal scale inconsistencies, such as mismatched time intervals between yearly health outcome reports and daily air pollution measurements, further complicate analyses. Future research could focus on multiscale modelling approaches that accommodate different spatial and temporal granularities, such as hierarchical models or scalable geospatial algorithms. Additionally, it is beneficial to develop spatial downscaling and upscaling techniques for harmonizing big geospatial datasets collected at different scales. Researchers could also explore the impacts of scale on pattern detection and inference to ensure that insights are valid across multiple spatial resolutions.

We hope that this special issue inspires further discussions, encourages the adoption of geospatial big data analytics, and drives innovations across diverse fields. We look forward to future studies that address the challenges outlined and advance the frontiers of geospatial big data research.

Acknowledgments

This study is based on work supported by a grant funded by the U.S. National Science Foundation: Collaborative Research: HNDS-I: Cyberinfrastructure for Human Dynamics and Resilience Research (Award No.: 2318206). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

Disclosure statement

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

References

Hu, W., G. Cervone, L. Trusel, and M. Yu. 2024. "Arctic Accessibility: Recent Trends in Observed Ship Tracks and Validation of Arctic Transport Accessibility Model." *Annals of GIS*: 1–19. https://doi.org/10.1080/19475683.2024.2380678.

Li, X., Y. Qiang, and G. Cervone. 2024. "Using Human Mobility Data to Detect Evacuation Patterns in Hurricane Ian." *Annals of GIS* 1–19. https://doi.org/10.1080/19475683.2024.2341703.

Liao, M., M. P. Kwan, and X. Liu. 2024. "Exploration of Human Activity Fragmentation in Cyber and Physical Spaces Using Massive Mobile Phone Data." *Annals of GIS* 1–18. https://doi.org/10.1080/19475683.2024.2341704.

Lin, B., L. Zou, M. Yang, B. Zhou, D. Mandal, J. Abedin, H. Cai, and N. Ning. 2024. "Progress in Understanding Human-COVID-19 Dynamics Using Geospatial Big Data: A Systematic Review." *Annals of GIS* 1–21. https://doi.org/10.1080/19475683.2024.2418584.

Liu, J., and Y. Yuan. 2024. "Exploring Dynamic Urban Mobility Patterns from Traffic Flow Data Using Community Detection." *Annals of GIS* 1–20. https://doi.org/10.1080/19475683.2024.2324393.

Liu, Z., Y. Li, A. Law, J. Y. K. Tan, W. H. Chua, Y. Zhu, and W. Luo. 2024. "Association Between NO_2 and Human Mobility: A Two-Year Spatiotemporal Study During the COVID-19 Pandemic in Southeast Asia." *Annals of GIS*: 1–18. https://doi.org/10.1080/19475683.2024.2325577.

Zhao, F. H., A. X. Zhu, L. J. Zhu, and C. Z. Qin. 2024. "iSoLIM: A Similarity-Based Spatial Prediction Software for the Big Data Era." *Annals of GIS* 1–15. https://doi.org/10.1080/19475683.2024.2324381.

Lei Zou
Department of Geography, Texas A&M University
lzou@tamu.edu

Yongze Song
School of Design and the Built Environment, Curtin University

Guido Cervone
Department of Geography, The Pennsylvania State University