Envisioning Unified Spatial Benchmarks

Paul Walther
TUM, Munich, Germany
paul.walther@tum.de

Fabian Deuser
UniBW, Neubiberg, Germany
fabian.deuser@unibw.de

Chiara Pugliese
IIT-CNR, Pisa, Italy
chiara.pugliese@iit.cnr.it

Wejdene Mansour
TUM, Munich, Germany
wejdene.mansour@tum.de

Xuanshu Luo
TUM, Munich, Germany
xuanshu.luo@tum.de

Johann Maximilian Zollner
TUM, Munich, Germany
maximilian.zollner@tum.de

Katharina Anders
TUM, Munich, Germany
k.anders@tum.de

Eleni Tzirita Zacharatou
HPI, UP, Potsdam, Germany
eleni.tziritazacharatou@hpi.de

Chiara Renso
ISTI-CNR, Pisa, Italy
chiara.renso@isti.cnr.it

Vagelis Hristidis
UCR, Riverside, California, USA
vagelis@cs.ucr.edu

Ahmed Eldawy
UCR, Riverside, California, USA
eldawy@ucr.edu

Martin Werner
TUM, Munich, Germany
martin.werner@tum.de

Abstract

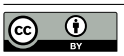
The rapid growth of data from domains such as remote sensing and mobility has led to a wide range of algorithms being proposed to analyze, evaluate, structure, and manage spatial data. However, due to the data's inherent multi-modal nature and the diversity of the research community and applications, the field of spatial computing remains highly fragmented. Unlike related domains, such as data management, with institutionalized benchmarks (e.g., Transaction Processing Performance Council (TPC)) and machine learning, with open community-driven comparisons (e.g., Hugging Face), in spatial computing neither a widely accepted semantic nor a performance benchmarking framework exists. Existing spatial benchmarks are usually designed for single tasks or specific applications, which limits comparability, reproducibility, and cumulative progress. In this vision paper, we argue that the lack of a unified task framework defining benchmarking areas constitutes a major bottleneck in spatial computing research. To address this, we analyze existing use cases and advocate for a task-based, generalized taxonomy that enables comprehensive comparisons across approaches spanning diverse data modalities and domains. Finally, we outline concrete next steps towards a unified SIGSPATIAL community approach to benchmarking of spatial computing applications, including reference implementations and governance mechanisms.

CCS Concepts

• **General and reference → Computing standards, RFCs and guidelines**; *Surveys and overviews*; • **Information systems → Spatial-temporal systems**.

Keywords

Spatial Computing, Evaluation, Benchmark, Taxonomy, Standard



SIGSPATIAL '26, Riverside, CA, USA

ACM ISBN 979-8-4007-2950-8/2026/11
https://doi.org/10.1145/3841645.3843415

ACM Reference Format:
Paul Walther, Fabian Deuser, Chiara Pugliese, Wejdene Mansour, Xuanshu Luo, Johann Maximilian Zollner, Katharina Anders, Eleni Tzirita Zacharatou, Chiara Renso, Vagelis Hristidis, Ahmed Eldawy, and Martin Werner. 2026. Envisioning Unified Spatial Benchmarks. In *The 34th ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '26), November 03–06, 2026, Riverside, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3841645.3843415

1 Introduction

Spatial data has become ubiquitous and increasingly multimodal. Rather than being described solely by coordinates, a single location can now be described by imagery, text, and sensor time series, and used in various applications. Although fundamental tasks such as clustering, classification, indexing, storage, or prediction are relevant across all those modalities and applications, spatial computing paradigms are often developed by researchers in a single modality and for a specific domain. This makes transferring them to other domains or modalities risky, since the outcome cannot be predicted upfront, and existing evaluations are often insufficient to judge whether an algorithm performs adequately across use cases.

Inspired by benchmarking traditions in related domains, we envision a unified, community-governed, cross-domain framework that reshapes how spatial computing measures progress and ensures fairness, transparency, and reproducibility. Specifically, we

- give an illustrative overview of existing computer science (CS) benchmarks with a special focus on spatial benchmarks
- outline our vision and an initial taxonomy for *task-based spatial benchmarking*, and
- call for concrete next steps toward joint, community-driven benchmarking procedures in SIGSPATIAL.

2 Existing Unified Benchmarks in CS

We can differentiate benchmarks in computer science by their objectives (from semantic correctness to performance) and governance (from community-driven efforts to organizationally maintained standards). These dimensions are shown in Figure 1 together with

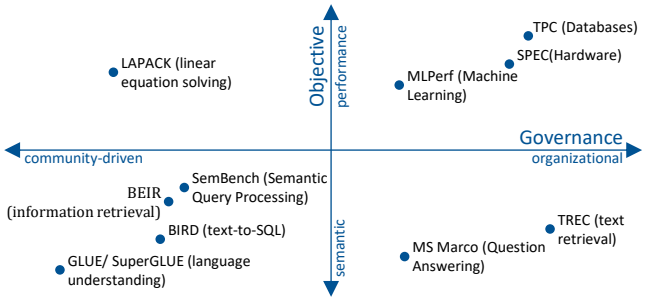


Figure 1: Exemplary benchmarking frameworks in CS qualitatively ordered by their objective and governance.

some exemplary benchmarks used in computer science (CS) and can give us a valuable idea of "good benchmarks".

Benchmarks that have achieved broad and lasting adoption, such as TPC [20] in databases, have clear objectives and strong governance. TPC provides a comprehensive framework for creating reproducible benchmarks with four essential components: a *dataset* or *data generator*, a *workload* of operations to execute, *metrics* that quantify performance, and *execution protocols* that fix the testing conditions. Additionally, TPC has a *governance* body that audits results, versions the benchmark, and manages its evolution.

The data can be synthetic for controlled scalability testing or real for fidelity to a target application. A common problem, particularly in machine learning, is treating a dataset alone as a benchmark: ImageNet [4], for instance, defines the data but leaves the remaining components to the user, making comparisons difficult. Spatial computing is no exception: as we show next, most of its benchmarks are datasets without standardized workloads, metrics, and protocols.

3 Existing Spatial Benchmarks

In the following, we introduce some existing benchmarks and benchmark datasets across various spatial data modalities and application domains. This is not intended to be a comprehensive survey, but should serve as an argument for unified spatial benchmarks. Notably,, apart from the broad but still point cloud-focused International Society of Photogrammetry and Remote Sensing (ISPRS) benchmarks [9], no known benchmark provider covers more than one domain or application. While all considered spatial benchmarks provide datasets or data generators, many of those lack defined workload, metrics or execution protocols and almost none provides a proper governance model (compare Table 1). This may be because these benchmarks are mostly community-driven or introduced with single-solution papers, not datasets.

Shape. Basic shapes are mainly used to benchmark spatial indexes, e.g., Jackpine builds on TIGER files [21], covers database operations like range query and spatial joins and defines not only the data but also workloads, metrics and execution protocols [17]. Because this benchmark is too small and has no global coverage researchers switched to OpenStreetMap [2] although there defined benchmark procedures are missing apart from the data.

Image. Images are the most benchmark-rich spatial modality and benchmarks are based on, e.g., high-resolution satellite imagery, such as the ISPRS object detection benchmark [9] , or drone-based

Table 1: Overview of exemplary spatial datasets and benchmarks. ● indicates a defined, ◑ a partly defined and ○ a missing component. For the governance ○ means absence, ◑ community-driven and ● institutional governance. The main tested objective is performance (P) or semantic (S).

	Benchmark	Source	Workload	Metrics	Protocol	Governance	Objective
Shape	Jackpine	[17]	●	●	●	◑	P
	OpenStreetMap	[2]	○	○	○	◕	P
Image	ISPRS Object Detection	[9]	●	●	●	●	S
	UAVid	[14]	●	●	●	◑	S
	GeoSR-Bench	[12]	●	●	●	○	S
	University-1652	[27]	●	●	●	○	S
	Cross-City-Matters	[8]	●	◑	●	◑	S
Mobility	GeoLife	[26]	○	○	○	○	P/S
	T-Drive	[25]	○	○	○	○	P/S
	Foursquare	[24]	○	○	○	○	P/S
	BerlinMOD	[6]	●	●	●	◑	P/S
	HuMob	[23]	●	●	●	◑	P/S
Point Cloud	ISPRS Benchmarks	[9]	●	●	●	●	P/S
	S3DIS	[3]	◑	◑	●	○	S
	DALES	[22]	◑	◑	◑	○	S
	KITTI	[7]	●	●	●	◑	S
Text	MapQA	[13]	●	◑	◑	◑	S
	MapEval	[5]	●	◑	●	◑	S
	GS-QA	[19]	●	●	◑	○	S
	SPARTQA	[16]	●	●	●	○	S
	ProximityQA	[10]	●	◑	◑	◑	S
Multi-Modal	PANGAEA	[15]	●	●	●	○	S
	VRSBench	[11]	●	●	◑	○	S
	MLPerf	[18]	●	●	●	●	P/S

UAVid segmentation dataset [14] and University-1652 [27], which combines drone, ground, and satellite views for cross-view localization. GeoSR-Bench [12] consolidates diverse Earth observation tasks into unified evaluation suites. Cross-City-Matters [8] additionally targets cross-region transfer across cities. These provide data, workloads, metrics, and execution protocols, but the procedures are inconsistently followed and reported. Further, the coverage is uneven geographically, and random tile splits leak information between neighboring tiles. Therefore, cross-region generalization remains the main open problem rather than data shortage.

Mobility. In the mobility domain, the availability of public benchmarks remains limited and is based solely on published data. Trajectories are available for vessel, animal, and hurricane movements, but individual human traces are rarely released due to their sensitivity, and many large-scale datasets remain proprietary. Existing datasets can be broadly divided into *dense*, being mostly GPS traces (e.g., GeoLife [26] and T-drive [25]), and *sparse*, being mostly discrete check-in datasets from location-based services (e.g., Foursquare [24]). For both types, workloads for real-world data are not standardized, and metrics and evaluation procedures are strongly tailored to the specific application. To overcome limitations in these raw real datasets, datasets are *aggregated* [23] and *synthetic trajectory generators* are widely employed, ranging from statistical

trajectory synthesis to more complex mobility simulation, such as BerlinMOD [6]. In tendency, these processed datasets include better benchmarking procedures. A good example is HuMob[23], which established a challenge around the dataset with defined workloads, metrics, and execution protocols [1].

Point Cloud. The ISPRS benchmark suite [9] provides some of the most established reference datasets, covering tasks such as semantic labeling, 3D reconstruction, and point cloud registration across urban and indoor environments. Complementary datasets extend this landscape across different acquisition settings and scales. Indoor benchmarks (e.g., S3DIS [3]) focus on scene understanding and reconstruction at building scale, while outdoor datasets target large-scale mapping and digital twin creation [22], as well as robotics and autonomous driving [7] using airborne, mobile, and terrestrial LiDAR. These benchmarks typically provide high-quality datasets and, in many cases, clearly defined workloads and metrics. Still, workloads are often domain-specific and not standardized across datasets, and execution protocols are not always thoroughly specified. Moreover, governance aspects such as structured update and maintenance processes are largely absent.

Text. For the textual modality, spatial benchmarks are framed predominantly as question answering (QA), organized primarily by whether the space is geospatially grounded, which shapes both data representation and the questions asked. Geolocation-based benchmarks such as MapQA [13], MapEval [5], and GS-QA [19] are grounded in real-world coordinates, representing data in structured (vector or raster) geospatial databases, and posing queries regarding geospatial facts such as locations, distances, and routes. In contrast, closed-domain spatial reasoning dispenses with geospatial grounding, requiring a model to infer spatial relations from a textual scene description, such as relative positions [16] and proximity [10]. When spatial context is conveyed beyond text alone, e.g., additionally with images, the task becomes visual spatial reasoning, which is examined next in the multimodal setting.

Multi-Modal. Multi-modal datasets form less a category of their own than an axis cutting across the data discussed above. Several benchmarks already mentioned are inherently multi-modal, e.g., with image and point cloud data [7] or raster maps with textual data [13, 19]. The term is best reserved for datasets whose primary challenge is fusing two or more heterogeneous modalities, as in PANGAEA [15] for optical-SAR fusion and VRSBench [11] for vision-language tasks, while MLPerf [18] provides a task-agnostic performance harness rather than a dataset. The modalities that fuse tend to be those co-registered by the sensing platform, while data needing cross-source alignment, such as trajectories, vector geometries and tabular attributes, stay largely absent, exactly the case where alignment is hard rather than automatic.

In summary, despite the richness of available data, current spatial benchmarks remain fragmented across application domains and data modalities, limiting comparability and hindering the establishment of a unified benchmarking framework.

4 A Spatial Benchmarking Taxonomy

Classically, computational systems are compared via complexity theory, emphasizing asymptotic worst-case behavior. This rarely reflects real performance, since algorithms often perform differently

Figure 2: A taxonomy of task areas for benchmarking in the spatial data domain in five hierarchical levels

by exploiting data characteristics, hardware effects (e.g., memory access patterns), and approximate or randomized methods. Consequently, systems are increasingly evaluated empirically on concrete *tasks* with realistic data on shared hardware, focusing on the trade-off between (semantic) correctness and performance. However, as described in Section 3, spatial benchmarks are organized by modality, so recurring tasks (such as indexing, segmentation, clustering, and reasoning) are benchmarked in isolation within each modality. We therefore suggest organizing spatial computing by *task* rather than modality, creating a consistent framework for defining workloads, metrics, and protocols often lacking in current benchmarks. This serves as a foundation for a common, community-driven application independent definitions of benchmarks structured in five hierarchical levels with increasingly complex tasks (Figure 2).

The group *1 Foundational Spatial Processing* defines fundamental task categories that support many advanced algorithms or applications, ranging from computation on primitives to basic indexing and storage principles. The group *2 Spatial Perception* mainly covers tasks in data gathering, including sensor management, sensor

(co-)registration, and localization. These are usually shared across domains like large-scale mapping of urban and natural landscapes, autonomous system development, and data acquisition for, e.g., remote sensing. *3 Spatial Pattern Discovery & Analytics* relies on the basic principles from *1* and *2* and defines tasks that apply additional structure to the underlying data. This includes spatial clustering, classification, and statistics. Based on this aggregated and structured data, *4 Spatial Reasoning & Semantics* tasks aim to define and pursue relationship analyses among the identified patterns and data from *2* and *3*, using *1 Foundational Spatial Processing*. At the highest level, tasks in the group of *5 Integrated Spatial Systems* combine all underlying tasks to solve complex real-world problems.

Based on such a hierarchical task structure, the community can define a common understanding of task dependencies to enable a thorough evaluation of spatial processes across domains. For example, people developing urban digital twins would then know that their mapping procedures rely on the same foundational tasks as developments for virtual reality applications. Publications on these basic tasks may then be directly cross-referenced and used across domains as a common benchmark is established. Therefore, we see high potential in classifying developments in our community using a task-based taxonomy.

5 Conclusion and Call to Action

As described, there are already many datasets denoted "benchmarks" for specialized subfields of spatial computing, though none of them are widely accepted by the general community. In our opinion, this can be due to the following three reasons: the **social factor** (groups tend to use their own datasets and develop their own unique tasks or use cases), the **domain-specific publication** (although tasks are shared, existing datasets and benchmarks are usually framed for single domains only), and the **insufficient benchmark quality** (denoted datasets either do not cover specific data characteristics needed for the domain, or the "benchmarks" miss a thorough description of workloads, metrics, and execution protocols).

To overcome these limitations and progress in the field of spatial computing, we advocate a joint effort of the SIGSPATIAL community. This includes

(1) the definition of a widely accepted cross-domain taxonomy of shared spatial tasks,
(2) the explicit denotation of complete benchmarks for these spatial tasks including either datasets or data generators from various domains, workloads, metrics and execution protocols, and
(3) the maintenance and governance of these benchmarks and a set of reference implementations.

To achieve this, let us join forces and promote the introduction of a SIGSPATIAL Benchmarking Council for all spatial computing modalities and domains to unify benchmarking across our field. This will simplify testing, accelerate iterations, and strengthen scientific progress, laying a strong foundation for future research.

Acknowledgments

This work is partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 563303373.

References

[1] 2025. GIS Cup 2025. https://sigspatial2025.sigspatial.org/giscup/index.html. ACM SIGSPATIAL'25. Last accessed: 2026-06-10.
[2] 2026. Planet OSM. Online dataset. https://planet.openstreetmap.org Last accessed: 2026-06-04.
[3] Iro Armeni, Ozan Sener, Amir R. Zamir, et al. 2016. 3D Semantic Parsing of Large-Scale Indoor Spaces. In *CVPR*. 1534–1543. doi:10.1109/CVPR.2016.170
[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*. 248–255. doi:10.1109/CVPR.2009.5206848
[5] Mahir Labib Dihan, MD Tanvir Hassan, MD Tanvir Parvez, et al. 2025. MapEval: A Map-Based Evaluation of Geo-Spatial Reasoning in Foundation Models. In *ICML'25*.
[6] Christian Düntgen, Thomas Behr, and Ralf Hartmut Güting. 2009. BerlinMOD: A Benchmark for Moving Object Databases. *The VLDB Journal* 18, 6 (2009), 1335–1368. doi:10.1007/s00778-009-0142-5
[7] A Geiger, P Lenz, C Stiller, and R Urtasun. 2013. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* 32, 11 (2013), 1231–1237. doi:10.1177/0278364913491297
[8] Danfeng Hong, Bing Zhang, Hao Li, et al. 2023. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sensing of Environment* 299 (2023), 113856. doi:10.1016/j.rse.2023.113856
[9] International Society for Photogrammetry and Remote Sensing (ISPRS). n.d.. ISPRS Benchmark Datasets. https://www.isprs.org/resources/datasets/benchmarks/. Last accessed: 2026-06-09.
[10] Jianing Li, Xi Nan, Ming Lu, Li Du, and Shanghang Zhang. 2024. Proximity QA: Unleashing the Power of Multi-Modal Large Language Models for Spatial Proximity Analysis. arXiv:2401.17862 doi:10.48550/arXiv.2401.17862
[11] Xiang Li, Jian Ding, and Mohamed Elhoseiny. 2024. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *NeurIPS* 37 (2024), 3229–3242. doi:10.52202/079017-0106
[12] Zhili Li, Kangyang Chai, Zhihao Wang, et al. 2026. Beyond Visual Fidelity: Benchmarking Super-Resolution Models for Large-Scale Remote Sensing Imagery via Downstream Task Integration. (2026). arXiv:2605.00310 doi:10.48550/arXiv.2605.00310
[13] Zekun Li, Malcolm Grossman, Ehsan Qasemi, et al. 2025. Benchmarking Geospatial Question Answering with MapQA. In *SIGSPATIAL'25*. 1042–1045. doi:10.1145/3748636.3764174
[14] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. 2020. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm. Remote Sens.* 165 (2020), 108–119.
[15] Valerio Marsocci, Yuru Jia, Georges Le Bellier, et al. 2024. Pangaea: A global and inclusive benchmark for geospatial foundation models. (2024). arXiv:2412.04204 doi:10.48550/arXiv.2412.04204
[16] Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. In *NAACL-HLT*. 4582–4598. doi:10.18653/v1/2021.naacl-main.364
[17] Suprio Ray, Bogdan Simion, and Angela Demke Brown. 2011. Jackpine: A benchmark to evaluate spatial database performance. In *ICDE*. 1139–1150. doi:10.1109/ICDE.2011.5767929
[18] Vijay Janapa Reddi, Christine Cheng, David Kanter, et al. 2020. Mlperf inference benchmark. In *ISCA*. 446–459. doi:10.1109/ISCA45697.2020.00045
[19] Majid Saeedan, Muhammad Shihab Rashid, Ahmed Eldawy, and Vagelis Hristidis. 2026. GS-QA: A Benchmark for Geospatial Question Answering. arXiv:2605.22811 doi:10.48550/arXiv.2605.22811
[20] Transaction Processing Performance Council. 1988. (TPC). https://www.tpc.org. Last accessed: 2026-08-14.
[21] US Census Bureau. 2026. TIGER/Line Shapefiles. Online dataset. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html
[22] Nina Varney, Vijayan K. Asari, and Quinn Graehling. 2020. DALES: A Large-scale Aerial LiDAR Data Set for Semantic Segmentation. In *CVPRW*. 717–726. doi:10.1109/CVPRW50498.2020.00101
[23] Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, et al. 2023. Metropolitan Scale and Longitudinal Dataset of Anonymized Human Mobility Trajectories. arXiv:2307.03401 [cs.SI] https://arxiv.org/abs/2307.03401
[24] Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. Participatory Cultural Mapping Based on Collective Behavior Data in Location-Based Social Networks. *ACM Trans. Intell. Syst. Technol.* 7, 3, Article 30 (2016). doi:10.1145/2814575
[25] Jing Yuan, Yu Zheng, Chengyang Zhang, Xing Xie, and Guang-Zhong Sun. 2010. T-Drive: Driving Directions Based on Taxi Trajectories. In *SIGSPATIAL'10*. 99–108. doi:10.1145/1869790.1869807
[26] Yu Zheng, Longhao Wang, Ruochi Zhang, Xing Xie, and Wei-Ying Ma. 2008. GeoLife: Managing and Understanding Your Past Life over Maps. In *MDM*. 211–212. doi:10.1109/MDM.2008.20
[27] Zhedong Zheng, Yunchao Wei, and Yi Yang. 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *ACM-MM'20*. 1395–1403. doi:10.1145/3394171.3413896