Measuring the State of Open Science in Transportation Using Large Language Models

Junyi Ji[◇1], Ruth Lu[◇2], Linda Belkessa[‡3], Liming Wang[‡4], Silvia Varotto[‡*5], Yongqi Dong[6], Nicolas Saunier[7], Mostafa Ameli[3], Gregory S. Macfarlane[8], Bahman Madadi[5], and Cathy Wu[1]

[1]Vanderbilt University, Nashville, TN, United States

[2]Massachusetts Institute of Technology, Boston, MA, United States

[3]Université Gustave Eiffel, Paris,France
[4]Portland State University, Portland, OR, United States

[5]École Nationale des Travaux Publics de l'État (ENTPE), Vaulx-en-Velin, France
[6]Delft University of Technology, Delft, The Netherlands

[7]Polytechnique Montréal, Montréal, Canada

[8]Brigham Young University, Provo, UT, United States

Short summary

Open science initiatives have strengthened scientific integrity across many fields, yet their adoption in transportation research remains under-investigated. Key features of open science—data and code availability—are difficult to extract at scale due to the complexity of full-text publications. We introduce a scalable, LLM-based pipeline to measure data and code availability and validate its performance through inter-rater agreement analysis against a manually curated dataset of 96 papers. Applied to 10,724 research articles published in Transportation Research journals (2019–2024), we find that only 5% of quantitative papers share a code repository and 4% share a data repository, with patterns varying across journals, topics, and geographic regions. No significant link is found between open science practices and traditional academic incentives, suggesting that structural interventions from publishers and funding agencies are needed. An interactive explorer is available at `https://www.rerite.org/MOST`.
Keywords: Large Language Models, Open Science, Open Science Monitoring, Transportation Research

1 Introduction

Open science initiatives have accelerated research progress across many disciplines, including computational science (Peng, 2011), psychology (Open Science Collaboration, 2015), and statistics (Stodden, 2014). Recently, open science practices have gained increasing recognition in the transportation research community, as reflected by the introduction of data availability statements in Transportation Research journals. Notably, *Transportation Research Part C* explicitly emphasizes and promotes open science in its aims and scope, stating that "Special emphasis is given in open science initiatives and promoting the opening of large-scale datasets that can support transferability and benchmarking of different approaches" (*Aims & Scope - Transportation Research Part C: Emerging Technologies*, 2025). However, despite these policy advancements, the extent to which open science practices have been adopted and implemented in transportation research remains largely unexamined.

*The author who will present the paper if accepted
◇These authors contributed equally to this work.
‡These authors contributed equally to this work.

Open science monitoring (OSM) has emerged as a key research direction endorsed by UNESCO (UNESCO, 2022), since transparent measurement is a practical mechanism for guiding and accelerating the transformation of research practices. In 2025, core principles for OSM were established (Bobrov et al., 2025), emphasizing relevance, transparency, and responsible use. Making data and code available is the first essential step toward reproducibility and replicability (National Academies of Sciences et al., 2019; Crüwell et al., 2019), allowing other researchers to verify and build upon original work.

Measuring open science practices is, however, resource intensive. Traditional approaches rely on manual review (Stagge et al., 2019), which does not scale to large corpora. Regular expression-based approaches (Riehl et al., 2025) improve scalability but generate false positives by lacking contextual understanding, for example, flagging benchmark repositories or citations to third-party tools as evidence of artifact sharing. Meanwhile, the nature of transportation research poses additional challenges: the boundary between using publicly accessible datasets and making new research artifacts available is frequently blurred, requiring clear operational definitions before any measurement can be conducted.

To address these challenges, this study (i) formalizes definitions and decision rules for data and code availability, (ii) introduces a scalable LLM-based pipeline that balances accuracy with scale and validates it against a manually annotated dataset, and (iii) provides the first field-wide analysis of open science practices across 10,724 Transportation Research articles, examining temporal, topical, and geographic factors. An interactive explorer is available at `https://www.rerite.org/MOST`.

2 Method

Definitions

We narrow *open science practices* to data and code availability connected with a paper, following Elsevier's Research Data Guidelines (Elsevier, 2025). For data, we distinguish two features: (i) *data repository*—whether the authors deposited a dataset in an accessible repository; and (ii) *data citation*—whether the paper cites or links to a publicly available dataset. These capture distinct behaviors: depositing data *contributes* to open science, while citing existing datasets *benefits* from it. Four data availability categories result: papers that neither cite nor deposit (NC, NR), deposit only (NC, R), cite only (C, NR), or both (C, R). For code, availability is defined as the presence of a public repository or archival deposit containing source files. Figure 1 illustrates the decision logic applied consistently throughout annotation, extraction, and analysis.

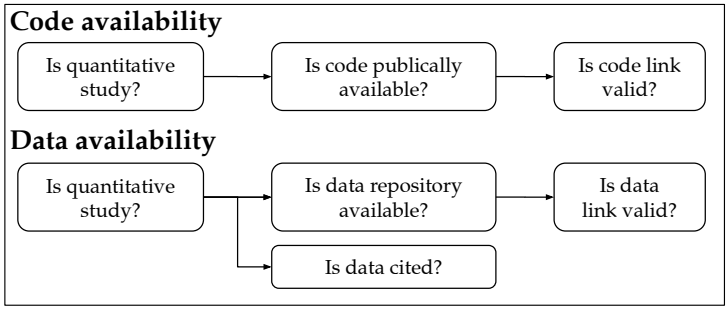


Figure 1: Decision logic for determining data and code availability.

Dataset

We curated a dataset of research articles published between 2019 and 2024 in Elsevier's Transportation Research journals (Parts A–F and Interdisciplinary Perspectives, TR-IP).

Starting from 10,990 publications retrieved via the Elsevier Text-and-Data-Mining API in structured XML format, we retained only full-length research articles after excluding editorials and reviews, yielding 10,724 papers. Figure 2 shows the distribution by journal and year.

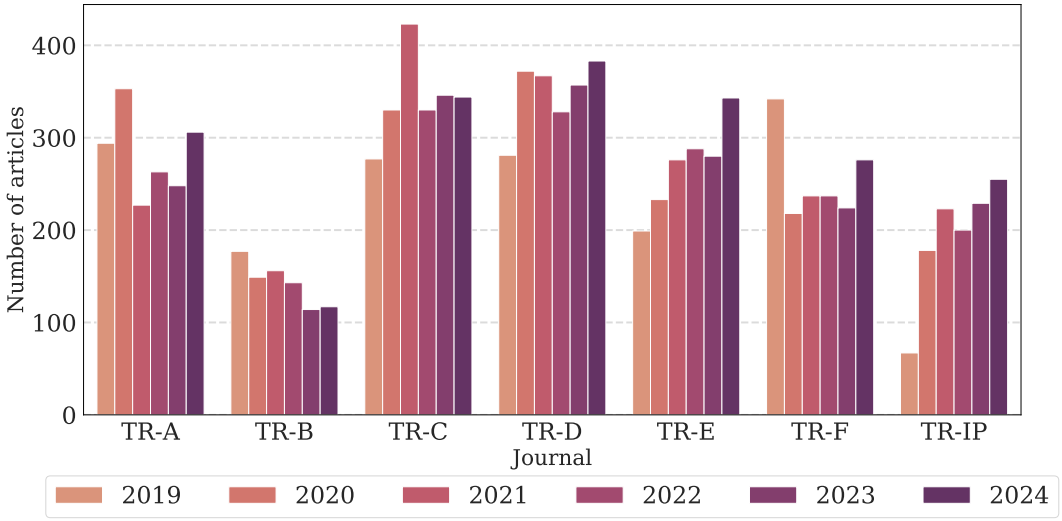


Figure 2: Articles published in Transportation Research journals (TR-A through TR-F and TR-IP) by year, 2019–2024 ($N = 10{,}724$).

Feature Extraction Pipeline

Our pipeline extracts three categories of features from the XML files. *Direct features* are parsed from structured XML tags and include paper characteristics such as author count, submission date, page count, and journal. *Derived features* include paper topic, obtained via Latent Dirichlet Allocation (LDA) (Blei et al., 2003) applied to abstracts, titles, and keywords, yielding 15 coherent research topics (e.g., Data-Driven Modeling & Prediction, Optimization & Routing Algorithms, Automated Driving & Human Factors). *LLM-extracted features*—data and code availability—require contextual understanding that keyword search cannot reliably provide. We use Google Gemini `gemini-2.5-flash-lite` (Gemini Team, 2025), selected for its large context window (over 1 million tokens) and strong performance on information extraction tasks as of 2025. Each paper's XML is converted to Markdown and processed via the Gemini API; separate prompts are issued for code and data features to preserve accuracy (Son et al., 2024; Gozzi & Di Maio, 2024). Extracted artifact links are then validated for liveness and content. Citation counts are obtained via the Scopus Search API as an external feature.

Human-LLM Validation

To evaluate extraction quality, we constructed a Manual Validation Dataset (MVD) of 96 randomly selected papers, each independently annotated by two human experts using predefined guidelines. A three-rater inter-rater agreement (IRA) analysis (McHugh, 2012) was conducted among the two human raters (H1, H2) and the LLM pipeline, using Fleiss's Kappa (Fleiss, 1971) and Cohen's Kappa (Cohen, 1960). Table 1 summarizes the results, interpreted using the Landis & Koch (1977) scale.
For all the features validated, `is_code_available` achieved Almost Perfect agreement. Data features showed moderate to fair agreement, matching the human-human baseline, indicating definitional ambiguity rather than LLM-specific limitations. Only features meeting established thresholds are retained for downstream modeling.

Table 1: Inter-rater agreement on MVD ($n = 96$). Fleiss's κ interpretation: 0.21–0.40 = Fair ; 0.41–0.60 = Moderate ; 0.81–1.00 = Almost Perfect .

Feature	**Cohen's** κ (H1-H2)	**Cohen's** κ (H1-LLM)	**Fleiss's** κ (all)
`is_code_available`	0.75	0.93	0.84
`is_data_cited`	0.67	0.41	0.52
`is_data_repo_available`	0.33	0.33	0.40
`is_quantitative_study`	0.48	0.40	0.46

3 Results

Open Science Adoption Trends

Among the 10,480 quantitative papers in the analysis dataset, only 528 (5%) share a code repository and approximately 4% share a data repository. About 29% cite or link to an existing dataset, but merely 3% share both a data repository and a citation. The majority (67%) neither cite data nor include a repository.
Clear temporal trends emerge: newer publications more frequently share code and data directly, while older papers tend to cite existing datasets without releasing new repositories. Geographic patterns are also pronounced: papers with authors based in Europe and North America show substantially higher sharing rates than those from Asia, even as Asian researchers contribute a large absolute number of papers.

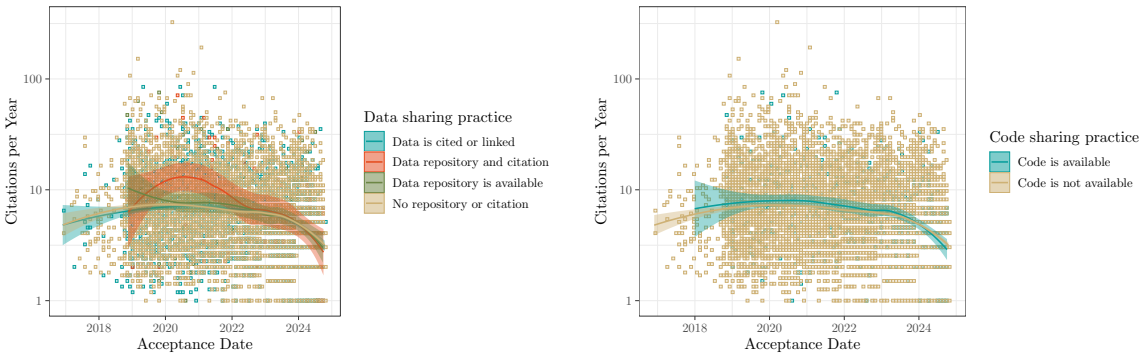
Factors Associated with Availability

To examine drivers of open science behavior, we estimated a binary logit model for code availability and a multinomial logit model for data availability (Bierlaire, 2023), incorporating journal, geographic region of the first author, research topic, and paper characteristics as explanatory variables. All reported effects are statistically significant at the 5% level.
Journal. Papers in TR-B and TR-C are significantly more likely to share a code repository than those in TR-D (reference). TR-C papers are also more likely to provide a data repository, consistent with TR-C's explicit commitment to open science. The other journals (TR-A, TR-E, TR-F, TR-IP) do not differ significantly from TR-D in code sharing.
Geography. Papers with first authors based in Europe, North America, and Oceania are substantially more likely to share both code and data than those from Asia. The effect is largest for Europe, South America, and Africa for code sharing, and for Europe and North America for data repositories.
Research topic. Data-Driven Modeling & Prediction papers show the highest likelihood of sharing both code and data. Driver Behavior, Logistics, Road Infrastructure, and Supply Chain topics are the least likely to share artifacts, and topics such as Automated Driving, COVID-19, Transport Policy, and Traffic Flow are also significantly less likely to share code than the Optimization reference category.
Paper characteristics. Papers with more figures are consistently more likely to share code and data, likely reflecting computationally intensive workflows. Newer papers are more likely to share code but less likely to have a data repository. Longer papers and larger author teams are negatively associated with data repository sharing, potentially reflecting coordination challenges in preparing complex datasets for release.

The Incentive Gap

A central question is whether open science practices are associated with measurable academic rewards. Figure 3 shows citations per year organized by data and code sharing

practice. In neither case are citation rates meaningfully different based on availability—no statistically significant difference was found between papers that shared artifacts and those that did not (aside from an anomalous spike in 2020–2021 driven by two heavily-cited outlier papers with shared data).



(a) Citations per year by data sharing practice.

(b) Citations per year by code sharing practice.

Figure 3: Citations per year over time by sharing practice, with a LOESS moving average for mean comparison.

Sharing data is not associated with shorter review times. However, sharing code is associated with a *longer* review duration: papers with code available averaged 269 days compared to 254 days for those without (95% CI: +2 to +27 days). While modest in absolute terms, this suggests that providing code incurs a review burden without a compensating reward.

This persistent misalignment—where sharing code or data neither accelerates citations nor shortens review time—implies that current incentive structures offer limited professional returns for open science effort.

4 Conclusion

This study provides the first large-scale, field-wide analysis of open science practices in transportation research. With only 5% of quantitative papers sharing a code repository and 4% sharing a data repository, the transportation research community lags well behind adjacent fields such as machine learning, where structured submission policies have pushed code availability rates above 50% (Pineau et al., 2021). The absence of significant incentive effects confirms that voluntary cultural change alone will not close this gap.

Publishers represent the highest-leverage intervention point. We recommend mandating structured data and code availability statements, creating open science badge schemes for verified repositories, hosting special issues dedicated to open datasets and benchmarks, and incorporating open science checklists into peer-review workflows. Conferences can amplify adoption by dedicating keynotes, sessions, and awards to open science. Funding agencies such as the NSF and EU Horizon can accelerate change by requiring open sharing as a grant condition—evidence suggests that funder requirements are among the most consistently effective motivators for data sharing (Hahnel et al., 2024).

The automated pipeline developed in this study can be readily scaled to other journals and fields, representing a critical step toward sustained open science monitoring. We are committed to periodically tracking new publications and releasing updated statistics. Data, code, and the interactive explorer are available at `https://www.rerite.org/MOST`.

Acknowledgements

This research utilized data accessed through Elsevier's APIs in compliance with Elsevier's Text and Data Mining policy. We are grateful for the Reproducible Research in Transportation Engineering (RERITE) Working Group for providing us with the opportunity to build and strengthen this collaborative team. We appreciate the feedback and support from our colleagues, including Zuduo Zheng at University of Queensland, Christine Buisson at École Nationale des Travaux Publics de l'État (ENTPE), Bidisha Ghosh at Trinity College Dublin, Joan Walker at UC Berkeley, Xinyu Chen at MIT, Gergely Zachár, Daniel B. Work and Jonathan Sprinkle at Vanderbilt University, Micah Altman at MIT Libraries. We gratefully acknowledge the volunteers who contributed to the manual validation dataset, including Khadidja Kadem, Cong Son Duong, Ali Naaman, Jincheng Ke, Maissa Mati, and Tuan Nguyen from Gustave Eiffel University COSYS/GRETTIA; Cameron Hickert, Jung-Hoon Cho, Shreyaa Raghavan, Vindula Jayawardana, Han Zheng, Hank Lin, Tianyue Zhou, and Edgar Ramirez Sanchez from MIT; Yiru Jiao from TU Delft; and Zhexian Li from USC. Their efforts were essential to the accuracy and reliability of our feature extraction process. Early versions of this work were presented at Transportation Research Symposium (TRS) 2025, Transportation Research Board Conference on Data and AI for Transportation Advancement (TRB DATA) 2025, International Symposium on Transportation Data and Modeling (ISTDM) 2025, and Modeling Mobility Conference (MoMo) 2025. This research was partially supported by the National Science Foundation (NSF) CAREER award #2239566 (Wu), NSF award #2434399 (Wu), the Marie Sklodowska-Curie COFUND grant agreement no. 101034248 (Belkessa, Ameli), French ANR award #ANR-24-CE22-7264 (Ameli) and #ANR-24-CE22-2750 (Varotto). The views expressed herein are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Aims & scope - Transportation Research Part C: Emerging technologies. (2025). https://www.sciencedirect.com/journal/transportation-research-part-c-emerging-technologies/about/aims-and-scope. (Accessed: 2025-06-16)

Bierlaire, M. (2023). *A short introduction to biogeme* (Tech. Rep. No. TRANSP-OR 230620). Transport and Mobility Laboratory, École Polytechnique Fédérale de Lausanne. Retrieved from https://transp-or.epfl.ch/documents/technicalReports/Bier23.pdf (Version prepared using Biogeme 3.2.11 (April 20, 2023))

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.*, *3*(4-5), 993–1022. doi: 10.1162/jmlr.2003.3.4-5.993

Bobrov, E., Bracco, L., Dacos, M., Fressengeas, N., Hrynaszkiewicz, I., Iarkaeva, A., ... Sabo, R. (2025, July). *The principles of open science monitoring.* Open Science Monitoring Initiative. Retrieved from https://doi.org/10.5281/zenodo.17853227 doi: 10.5281/zenodo.17853227

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., ... Schulte-Mecklenbeck, M. (2019). Seven easy steps to open science. *Zeitschrift für Psychologie*.

Elsevier. (2025). *Research data guidelines.* Retrieved from https://www.elsevier.com/researcher/author/tools-and-resources/research-data/data-guidelines (Accessed: 2025-12-23)

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Gemini Team. (2025). *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities* (Tech. Rep.). Google. Retrieved from

https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf (Accessed: 2025-06-20)

Gozzi, M., & Di Maio, F. (2024). Comparative analysis of prompt strategies for large language models: Single-task vs. multitask prompts. *Electronics*, *13*(23), 4712.

Hahnel, M., Smith, G., & Campbell, A. (2024, 12). *The State of Open Data 2024: Special Report Bridging policy and practice in data sharing.* Retrieved from https://digitalscience.figshare.com/articles/report/_b_The_State_of_Open_Data_2024_Special_Report_b_Bridging_policy_and_practice_in_data_sharing/27337476 doi: 10.6084/m9.figshare.27337476.v2

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

National Academies of Sciences, Medicine and Policy, Global Affairs, Board on Research Data, Information and Division on Engineering and Physical Sciences, Committee on Applied and Theoretical Statistics, et al. (2019). *Reproducibility and replicability in science*. National Academies Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., ... Larochelle, H. (2021). Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, *22*(164), 1–20.

Riehl, K., Kouvelas, A., & Makridis, M. A. (2025). Revisiting reproducibility in transportation simulation studies. *European Transport Research Review*, *17*(1), 22.

Son, G., Baek, S., Nam, S., Jeong, I., & Kim, S. (2024). Multi-task inference: Can large language models follow multiple instructions at once? In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 5606–5627).

Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A., & James, R. (2019). Assessing data availability and research reproducibility in hydrology and water resources. *Scientific data*, *6*(1), 1–12.

Stodden, V. (2014). The reproducible research movement in statistics. *Statistical Journal of the IAOS*, *30*(2), 91–93.

UNESCO. (2022). *Understanding open science* (Vol. 35). https://doi.org/10.54677/UTCD9302. (Document code: SC-PBS-STIP/2022/OST/1, 6 pages)