Towards Accelerating Transportation Research: Measuring the Practice of Open Science

Junyi Ji, Ruth Lu, Yongqin Dong, Liming Wang, Bahman Madadi, Silvia Varotto, Nicolas Saunier, Gregory S. Macfarlane[a], Mostafa Ameli, Cathy Wu[b,*]

[a]*Brigham Young University, Civil and Construction Engineering,*

[b]*Massachusetts Institute of Technology, Civil and Environmental Engineering,*

Abstract

This is the abstract. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum augue turpis, dictum non malesuada a, volutpat eget velit. Nam placerat turpis purus, eu tristique ex tincidunt et. Mauris sed augue eget turpis ultrices tincidunt. Sed et mi in leo porta egestas. Aliquam non laoreet velit. Nunc quis ex vitae eros aliquet auctor nec ac libero. Duis laoreet sapien eu mi luctus, in bibendum leo molestie. Sed hendrerit diam diam, ac dapibus nisl volutpat vitae. Aliquam bibendum varius libero, eu efficitur justo rutrum at. Sed at tempus elit.

Keywords: Open Science, Large Language Models

1. Introduction

I need a citation to keep from getting a latex error, so Nosek et al. (2015) is a good one.

2. Analysis

2.1. Descriptive statistics and bivariate tests

The analysis dataset contains 10 480 papers, processed from all full-length articles that were identified as `is_quantitative_study` papers). Table 1 shows descriptive statistics for the dataset, including organized by whether the papers made code available. The dataset contains 2 810 papers with code available, about 5% of the papers. In total, about 28% of the papers cite publicly available data, but only 4% of the papers include a data repository; the remaining 69% neither cite data nor include a repository.

*Corresponding author
Email address: `cathywu@mit.edu` (Cathy Wu)

Table 1: Descriptive Statistics of Data Set

		Code is available (N=528)		Code is not available (N=9952)	
		Mean	Std. Dev.	Mean	Std. Dev.
Open science score		3.4	1.1	2.2	0.6
Age of paper [years]		3.1	1.7	3.4	1.7
N. tables in paper		6.3	4.6	6.4	4.5
N. figures in paper		11.0	6.5	8.6	6.1
N. references		57.3	21.0	58.2	26.9
N. authors		3.8	1.7	3.8	1.7
Times cited		27.2	42.9	28.4	43.2
Review time [days]		268.6	141.4	254.3	148.9
Page count		19.8	6.6	18.4	6.3
		N	Pct.	N	Pct.
Open Access	Not open access	324	61.4	7346	73.8
	Open access	204	38.6	2606	26.2
Data availability	Data is cited or linked	96	18.2	2810	28.2
	Data repository and citation	94	17.8	14	0.1
	Data repository is available	199	37.7	65	0.7
	No repository or citation	139	26.3	7063	71.0
Journal	Interdisciplinary Perspectives	36	6.8	945	9.5
	Part A: Policy and Practice	64	12.1	1555	15.6
	Part B: Methodological	67	12.7	771	7.7
	Part C: Emerging Technologies	212	40.2	1797	18.1
	Part D: Transport and Environment	72	13.6	1911	19.2
	Part E: Logistics and Transportation Review	41	7.8	1534	15.4
	Part F: Traffic Psychology and Behaviour	36	6.8	1439	14.5
Region of corr. author	Africa	5	0.9	69	0.7
	Asia	118	22.3	4398	44.2
	Europe	206	39.0	2699	27.1
	North America	156	29.5	2074	20.8
	Oceania	27	5.1	521	5.2
	South America	16	3.0	191	1.9
Availability statement	Availability statement is not present	139	26.3	7062	71.0
	Availability statement is present	389	73.7	2890	29.0
Paper topic	Air & Freight Logistics	2	0.4	77	0.8
	Automated Driving & Human Factors	34	6.4	957	9.6
	COVID-19 Impact on Travel & Activities	17	3.2	434	4.4
	Data-Driven Modeling & Prediction	142	26.9	772	7.8
	Driver Behavior & Safety Risk	10	1.9	631	6.3
	Electric Vehicles & Ride-Sharing	29	5.5	372	3.7
	Optimization & Routing Algorithms	114	21.6	1513	15.2
	Public Transit Service & Demand	29	5.5	489	4.9
	Road Infrastructure & Emergency Management	1	0.2	49	0.5
	Social & Policy Aspects of Mobility	26	4.9	862	8.7
	Supply Chain & Market Strategies	6	1.1	775	7.8
	Traffic Flow & Network Control	45	8.5	710	7.1
	Transportation Emissions & Policy	20	3.8	848	8.5
	Travel & Mode Choice Behavior	28	5.3	883	8.9
	Urban Environment & Active Transport	25	4.7	580	5.8
Links to code repositories	Link to GitHub	329	62.3	4	0.0
	Link to other service	118	22.3	5	0.1
	No links	81	15.3	9943	99.9

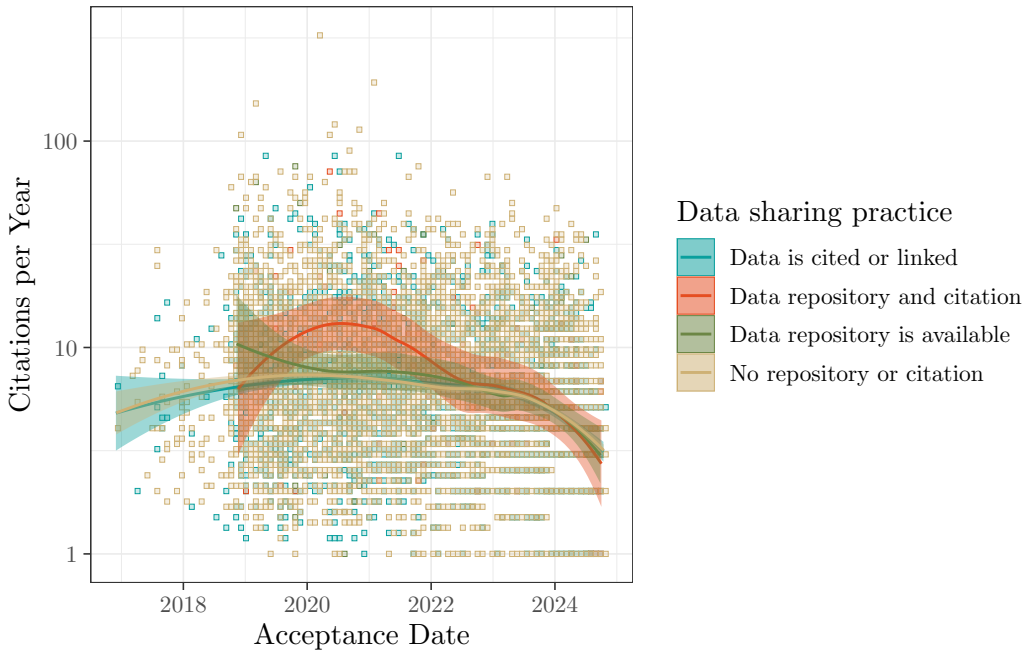
We conducted a series of initial bivariate statistical tests on the dependent variables — specifically, whether the papers provide code or data — using t-tests on differences of means when the independent variable is numeric, and Pearson χ^2 tests of independence when the independent variable is categorical. The results of these tests revealed several statistically significant relationships that guided the analysis that follows below. To begin, papers that make code available are also more likely to share data (χ^2 test statistic of 4 411.71). Newer papers make code available more frequently (t-test 5.02), and papers that share data directly tend to be newer than papers simply citing existing data (t-test between data repository and links 3.15) Papers with corresponding authors in Europe and North America are more likely to make data available than papers with authors in Asia; the same trend holds for sharing code, with South American authors joining the group more likely to share than Asian authors. In absolute terms, authors based in Asia still contribute a substantial number ($N = 118$) of code repositories, but their share appears lower when expressed as a percentage of the total number of publications from the region ($N = 4516$).

These findings are informative, but they may be the result of correlated omitted variables. In a section below we provide a choice model of data and code availability that can accommodate the simultaneous influence of multiple variables. Additionally, the analyses assume that the data is accurate, which the above agreement analysis indicated is not entirely true.

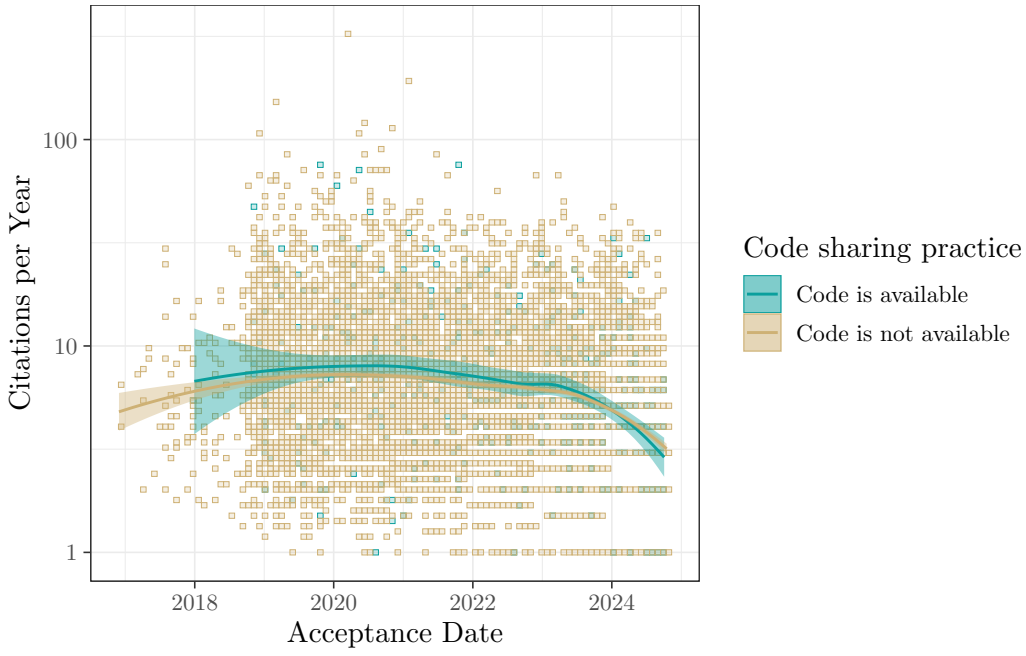
2.2. Incentives for availability

For researchers to take the extra effort to publish their data or their analysis code, there need to be incentives to do so. These incentives may be captured by a higher rate of citations or a reduced review time prior to publication. Unfortunately, neither incentive is observable in the papers we studied. Figure 1a shows the citations per year for each paper organized by paper acceptance date in SCOPUS and data availability, and Figure 1b the same information by code availability. In neither case are the average citation rates different based on the availability of data or code. The average citation rate for papers that provided both a data repository and a citation to existing data was elevated in 2020 and 2021; this effect was primarily driven by two heavily-cited papers accepted in 2020 that shared data and a citation to existing data. The overall citation rate for papers sharing both a data repository and a citation to existing data has returned to a similar rate as the rate for papers that do not share data in the intervening years.

In terms of review time, there was no difference in mean time for papers based on data sharing practice. There was, however, a statistically meaningful *increase* in the time to review papers that shared code repositories based on a t-test of the difference in mean review time. We observed a 95% confidence interval between 1 and 27 days longer for papers with a code repository than papers that did not share a code repository based on a t-test of the difference in mean review time.



(a) Citations per year by data sharing practice.

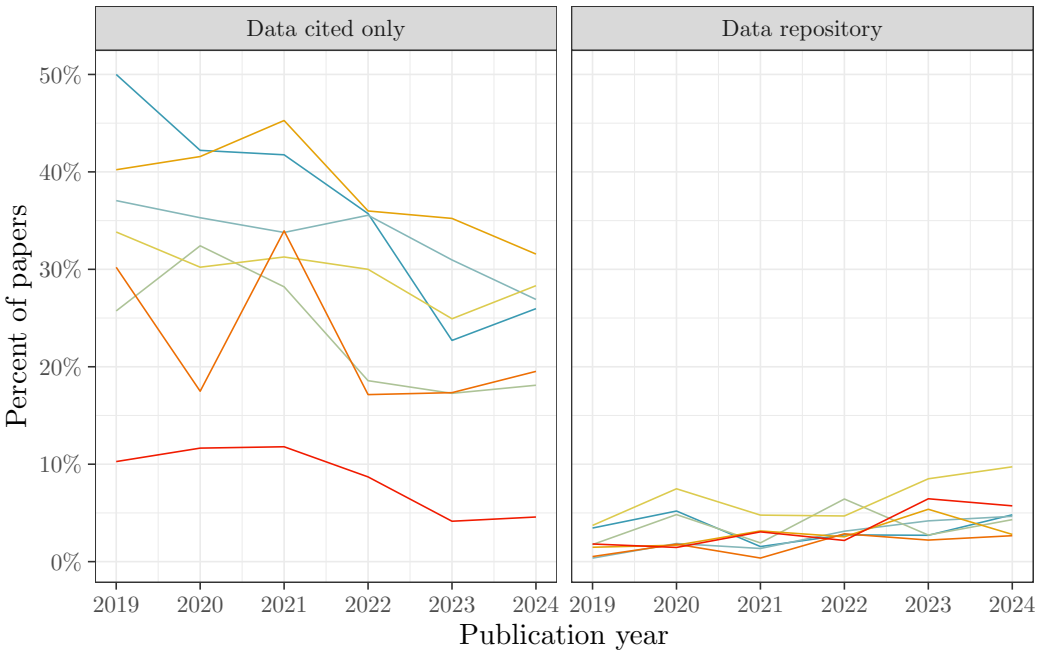


(b) Citations per year by code sharing practice.

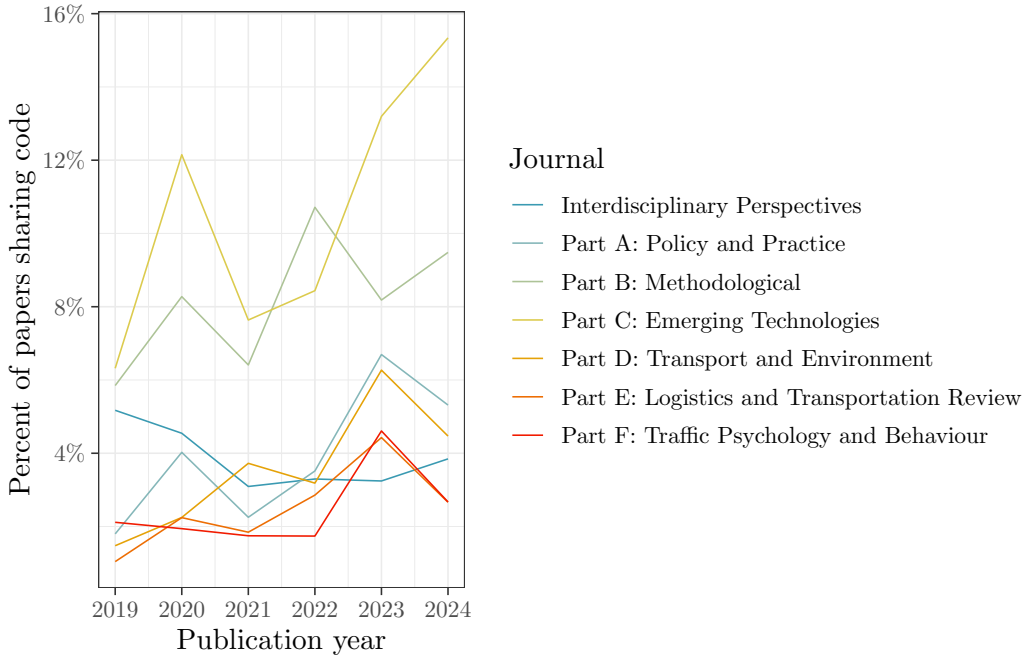
Figure 1: Citations per year of each paper over time by code and data-sharing practice, with a LOESS moving average line applied for mean comparison.

2.3. Temporal, topic, and journal trends of data and code availability

Figure 2 shows the temporal trend of data availability by journal in Figure 2a, and the temporal trend of code availability by journal in Figure 2b. As to data availability, there is a slight positive trend in the percent of papers that include a data repository, but there is a somewhat negative trend in the percent of papers using cited data.



(a) Data availability by journal over time.



(b) Code availability by journal over time.

Figure 2: Temporal trends of data and code availability.