

Missing Evidence

Tracking Academic Data Use around the World

Brian Stacy

Lucas Kitzmüller

Xiaoyu Wang

Daniel Gerszon Mahler

Umar Serajuddin



WORLD BANK GROUP

Development Economics
Development Data Group
January 2024



Reproducible Research Repository

A verified reproducibility package for this paper is available at <http://reproducibility.worldbank.org>, click **here** for direct access.

Abstract

Data-driven research on a country is key to producing evidence-based public policies. Yet little is known about where data-driven research is lacking and how it could be expanded. This paper proposes a method for tracking academic data use by country of subject, applying natural language processing to open-access research papers. The model's predictions produce country estimates of the number of articles using data that are highly correlated with a human-coded approach, with a correlation of 0.99. Analyzing more than 1 million academic articles, the paper finds that the number of articles on a country is strongly

correlated with its gross domestic product per capita, population, and the quality of its national statistical system. The paper identifies data sources that are strongly associated with data-driven research and finds that availability of sub-national data appears to be particularly important. Finally, the paper classifies countries into groups based on whether they could most benefit from increasing their supply of or demand for data. The findings show that the former applies to many low- and lower-middle-income countries, while the latter applies to many upper-middle- and high-income countries.

This paper is a product of the Development Data Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://www.worldbank.org/prwp>. The authors may be contacted at bstacy@worldbank.org. A verified reproducibility package for this paper is available at <http://reproducibility.worldbank.org>, click [here](#) for direct access.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Missing Evidence: Tracking Academic Data Use around the World

Brian Stacy, Lucas Kitzmüller, Xiaoyu Wang, Daniel Gerszon Mahler, and Umar Serajuddin¹

Keywords: Data, academia, research, natural language processing

JEL codes: C45, C52, O30

¹ Stacy, Wang, Mahler, and Serajuddin are with the World Bank's Development Data Group. Lucas Kitzmüller completed the work while at the European Bank for Reconstruction and Development (EBRD). Corresponding author: Brian Stacy (bstacy@worldbank.org). We are grateful for comments from Dean Jolliffe, Jishnu Das, Olivier Dupriez, and Patrick Brock. We acknowledge financial support from a World Bank Research Support Grant (P178728). The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. Views presented are those of the authors and not necessarily of the EBRD.

1 Introduction

In recent decades, the amount of data produced has exploded, generating boundless opportunities for policies to improve people’s lives (World Bank 2021). Though data can be valuable in their raw form, the full value of data is only realized when they are analyzed to create insights, and these insights are converted to public policies or increased accountability.

Researchers have a vital role to play in this regard. Many researchers spend countless hours digesting data, using data to create new knowledge, and communicating this knowledge with the intent of impacting public discourse and public policies. There are numerous examples of data-driven analyses having real and important impacts on people’s lives (Jolliffe et al. 2023). One example from Brazil explicitly looks at researchers’ ability to influence policy outcomes. There, evidence from 2,150 municipalities found that informing municipal mayors of research findings on the effectiveness of a simple policy change increased the probability that their municipality implemented the policy by 10 percentage points (Hjort et al. 2021).

Without research, there is a risk that the return of data to society will be reduced, and policies to improve lives unrealized. Yet very little is known about where there is missing data-driven evidence and how governments can best stimulate an evidence base for local decision makers. This paper attempts to fill these gaps by addressing two questions: (1) Which countries are the subject of research papers using data? (2) How can countries increase their national evidence base? We focus on data-driven research due to the increasing importance of data for policy making and the specific policies that are needed to increase the supply and demand of data, such as boosting statistical capacity and improving data literacy.

To answer the first question, we introduce a new method for measuring data use in research articles based on 1 million English-language articles spanning 216 countries and various academic fields. These articles are made available by Semantic Scholar Open Research Corpus (s2orc), which has digitized millions of research papers worldwide and made its raw text accessible via APIs (Lo et al. 2020). With the aid of Amazon Mturk workers, we manually code 900 of these articles as using data or not, upon which we train a natural language model to predict the coding of the Mturk workers (Devlin et al. 2018). The model achieved an 87% out-of-sample accuracy rate, and when the articles are aggregated to the country level, the model had a correlation of 0.99 with the number of articles classified by Mturk workers. The model was then applied to 1 million academic articles from the s2orc database from 2000-2020. The model estimates the amount of data-driven research on a country regardless of where the researchers may be located, *not* the amount of data-driven research by the citizens of a country. We argue that the former is the relevant quantity to understand the evidence base available to national decision makers.

We find that data-driven research is strongly correlated with GDP per capita and population, which together account for around 75% of the variation across countries. High-income countries are the subject of nearly 50% of all papers using data despite representing only around 15% of the world’s population, while low-income countries, comprising approximately 10% of the world population, account for only about 5% of articles using data.

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/028062123077006033>