README: Replication Package

Title: Bridging divides with data: Open government data and ESG rating divergence

Authors:

Ximeng Liu (Inner Mongolia University)

Fengzheng Wang (Inner Mongolia University)

Youtao Xiang (Inner Mongolia University)

Date:

November 13, 2025

1. Overview

This zip file provides the data, code, and necessary instructions for reproducing the main empirical results (including tables and figures) of the paper titled “Bridging divides with data: Open government data and ESG rating divergence”. The purpose is to reproduce the key empirical findings of the aforementioned research paper. The package is designed to promote transparency and enable independent verification of all statistical results reported in the study. The data is sourced from third-party data platforms. Below is a detailed description of the datasets, code, and software used.

2. Content

The replication package includes the following folders and files:

2.1. Data Files

This folder contains seven datasets, covering the following areas: main regression data, channel analysis data, heterogeneity analysis data, and the data used for the robustness checks, etc. The corporate finance and governance data are mainly sourced from the

China Stock Market & Accounting Research (CSMAR)[1] and Wind[2] databases. ESG rating data are primarily sourced from six rating agencies: SusallWave, SynTao Green Finance, FTSE Russell, Wind, Huazheng, and Rankins. Researchers interested in the above datasets can contact or subscribe to the relevant databases. Below are the details of the five data files.

(1) OriginData.dta

This is a firm-year panel data basically used in our article, which contains the key variables that are used in regression analyses of our paper, e.g., ESGDiv: Dependent variable (ESG rating divergence) and Open: Independent variable (Open government data). The ESG score data used to calculate ESG divergence is sourced from the Bloomberg official website, Wind database, and CSMAR database.

(2) FirmControls.dta

It contains corporate-level data (e.g., financial data, corporate governance data) from CSMAR database. The following are the variables contained in this dataset.

(3) VarMechanism.dta

This data file contains the variables used to reproduce the regression results in Table 8 and Table 9, namely KV, DISP, Age, and Opacity. It is important to note that in Table 8 and Table 9, △ represents the mechanism variable. In Table 8, columns (1) and (2) use the variables firm age (Age) and analyst earnings forecast dispersion (DISP), while in Table 9, columns (1) and (2), △ represents the mechanism variables key value (KV) index and information disclosure quality (Opacity), respectively. This is explained in detail in the article. The data for these variables is sourced from the CSMAR database.

In this data, variable A represents the interaction between Open and Age, variable B represents the interaction between Open and DISP, variable C represents the interaction between Open and KV, and variable D represents the interaction between Open and

[1] https://data.csmar.com/
[2] https://www.wind.com.cn/portal/zh/EDB/index.html

Opacity.

(3) VarHeterogeneity.dta

This data primarily contains the heterogeneity variables used to obtain the regression results in Table 10, Table 11, and Table 12. To be specific, the variable (RIC) is a dummy variable for regional innovation capability, provided by the China Regional Science and Technology Innovation Evaluation Report. Interested researchers can refer to the website at http://www.casted.org.cn/.

Additionally, the variable (DT) represents the level of digital transformation of enterprises and comes from the CSMAR database, which requires researchers to subscribe in order to access it. Finally, the enterprise size variable (Size) in Table 12 is also sourced from the CSMAR database.

(4) VarRobust.dta

The file contains the key variable for robustness checks: ESG divergence (Table B1), where different ESG divergence variables are used to assess robustness. ESGDiv1 represents the ESG divergence index recalculated after excluding Rankins ratings. Additionally, we added Bloomberg's ESG score to the original six ESG rating agencies to recalculate ESG divergence (ESGDiv2), and the ESG data is sourced from Bloomberg's official website. The calculation method and data sources for ESG discrepancies are explained in detail in the text. Please be sure to review it.

(5) VarAlt.dta

This contains the data used for the regression results in Table 13. This data file contains four important variables: the number of analyst reports (AnalystReport), the quantity of news articles from newspapers (Media) and internet-based news (Internet), and whether the firm discloses an ESG report (ESGReport). Among them, the AnalystReport variable comes from the CSMAR database, Media and Internet are sourced from the Chinese Research Data Services Platform (CNRDS) database, and ESGReport is obtained from the Wind database.

(6) VarPolicy.dta

This contains the data used for the regression results in Table 5, which contains two important variables: BigData, and IntGov. The data for these two indicators were constructed in this paper, and the detailed data have been provided in the supplementary materials (i.e., BigData. xlsx, and IntGov. xlsx).

2.2. Code Files

This STATA do file contains the code for conducting the regression analysis. The code is organized and documented to make it easy to understand and modify as needed.

(1) MergeData.do

Since there are 7 data sources in this paper, the Stata command primarily performs the merging of these 7 datasets. The final result is a synthetic dataset, DATA.dta, which can be used to replicate the main empirical results presented in the article.

(2) CODE-Main.do

This file can be used to replicate the empirical results presented in the study. The document provides corresponding explanations for the replicated tables and figures at the beginning, such as **--Table 5 and **-- Fig. 2, etc. This makes it very easy for users to replicate the corresponding empirical results.

All do-files are self-contained and can be executed sequentially, provided the data files are placed in the correct directory. Users should set the appropriate working directory to ensure smooth replication.

Please review the code in the CODE-Main.do do file for detailed information about how the regression analysis is performed on the specified tables and columns.

(3) CODE-Appendix.do

CODE-Appendix.do: Runs additional analyses and robustness checks reported in the Appendix A and Appendix B, including Table B1, Table B2, Table B3, and Figure A.

2.3. Figures Files

This folder contains all the figures generated by the analysis, specifically including Figure 1, Figure 2, Figure 3, and Figure A of the Appendix A. The figures are output as high-resolution images, suitable for both publication and presentations.

2.4. Log Files

The "Log" folder contains comprehensive Stata log files along with a PDF created during the execution of each do-file. These log PDFs offer a full record of all executed commands, the order of operations, and the results at each stage of the analysis.

Here, we provide two formats of log files, namely PDF and SMCL formats, for the convenience of other researchers. The log files we provide allow them to trace the analysis process and verify the reproducibility of each empirical result reported in the study, thereby increasing transparency.

3. Usage

To replicate the regression analysis, follow these steps:

1. Make sure you have STATA 15 or later installed.

2. Open STATA and navigate to the directory where you saved the code and data files.

3. Execute the CODE-Main.do do file to perform the regression analysis.

4. No additional user-written packages are required for basic replication.

4. Data Access and Restrictions

Due to licensing, the following proprietary datasets cannot be shared but may be obtained independently:

(1) Original information on publicly listed firms in China can be obtained from the official website: https://data.csmar.com/

(2) The data for other self-constructed variables, such as BigData.xlsx and IntGov.xlsx,

has been included in the replicable package.

(3) Below, we provide the specific guidelines for CSMAR IDs to facilitate the retrieval of the corresponding variable data.

Variable name	CSMAR IDs
code	Stkcd [股票代码] -
year	accper [会计年度] -
nnindcd	nnindcd [行业代码（证监会 2012 版）] - 上市公司最新的 2012 年证监会分类代码。
nnindnme	nnindnme [行业名称（证监会 2012 版）] - 上市公司最新的 2012 年证监会分类名称。
Top1	Shrcr1 [股权集中度 1] - 公司第一大股东持股比例。
Size	A001000000b [企业规模] - 取资产各项目之总计的自然对数。
Lev	F011201A [资产负债率] - 计算公式为：负债合计 / 资产总计分子为空，零值代替；分母为空或是零值，结果以 NULL 表示。
ROA	F050202B [总资产净利润率（ROA）B] - 计算公式为：净利润 / 总资产平均余额。
Growth	F081602C [营业收入增长率 B] - 计算公式为：（营业收入本年本期金额-营业收入上年同期金额）/（营业收入上年同期金额）；当分母未公布或为零或小于零时，以 NULL 表示。
BM	F101001A [账面市值比 A] - 计算公式为：资产总计/市值。
BoardSize	Boardsize3 [董事会规模 C] - 董事人数加 1 的自然对数。
AnaAttention	AnaAttention [分析师关注度] - 分析师跟踪人数加 1 的自然对数。
IndDirectorRatio	IndDirectorRatio [独立董事比例] - 计算公式为：独立董事数量与董事规模之比。
InsInvestorProp	InsInvestorProp [机构投资者持股比例] - 机构投资者持有股份总数量占上市公司总股份比例。

5. License