Reproducibility Package for Informing Net Zero Industrial Pathways[1]

This file explains how to replicate the analysis for the paper "Informing Net Zero Industrial Pathways"

Overview

The data reproducibility package provides a systematic approach taken to ensure transparency and replicability in the analysis process. All raw data utilized for the analysis is meticulously organized within the designated Data folder, serving as the foundational repository for the study. The data processing is executed through an R script, which not only facilitates the transformation and analysis of the data but also delineates key outputs, ensuring clarity and accessibility for users.

The R script is designed to enhance reproducibility by incorporating explicit references to published data, along with corresponding figure numbers, labeled as "Data for Figure XX." This feature allows users to easily trace the data back to its original source.

Furthermore, all data employed in the creation of figures is systematically compiled in an Excel file named Charts. Within this file, the dark blue shaded tab specifically highlights the dataset extracted from the R script, providing a clear visual representation of the processed data.

Data Availability

This section outlines where and how the data supporting the findings of the study can be accessed and used. All data used in this analysis are publicly available.

Data Sources

We list the raw data sources here. We describe where one can download the data, and under which names they are called in. We use this raw data to extract the data needed for all the figures, tables and analysis reported in the paper.

Filename 1: NIC_2008 (as on the source site)

- **Source:** MOSPI, GoI
- **URL:** https://microdata.gov.in/NADA/index.php/catalog/205/study-description
- **Access year:** December 2024.

---

[1] Author(s): Meriem Ait Ali Slimane, Rishabh Choudhary.

- **Instructions:** This a pdf file and the relevant table has been converted to excel format using online tools. The resulting output is in the Data folder, "NIC.xlsx"

Filename 2: Codelist23 (as on the source site)

- **Source:** MOSPI, GoI
- **URL:** https://microdata.gov.in/NADA/index.php/catalog/205/study-description
- **Access year:** December 2024.
- **Instructions:** This a pdf file and the relevant table (with State name and codes) has been converted to excel format using online tools. The resulting output is in the Data folder, "Codelist18.xlsx". We have ignored the newly added states in the analysis. For instance, Ladakh is not included separately in the codelist.

Filename 3: Block A to J (Under "Get Microdata" on the source site)

- **Source:** MOSPI, GoI
- **URL:** https://microdata.gov.in/NADA/index.php/catalog/205/study-description
- **Access year:** December 2024.
- **Instructions:** ASI raw data is provided in .Nesstar format, with data structured by blocks corresponding to different sections of the schedule. Each block is exported to Stata format using Nesstar Explorer. This data can be found under the folder "Data" and sub-folder "202223".

Filename 4: ASUSE data: Level 1 to 16

- **Source:** MOSPI, GoI
- **URL:** https://microdata.gov.in/NADA/index.php/catalog/222
- **Access year:** December 2024.
- **Instructions:** Raw data is provided in .Nesstar format. Each LEVEL is exported to Stata format using Nesstar Explorer. This data can be found under the folder "Data">>"ASUSE">>"raw".

Data folder:

path
/Data/201920/A.dta
/Data/201920/B.dta
/Data/201920/C.dta

/Data/201920/D.dta
/Data/201920/E.dta
/Data/201920/F.dta
/Data/201920/G.dta
/Data/201920/H.dta
/Data/201920/I.dta
/Data/201920/J.dta
/Data/202223/A.dta
/Data/202223/B.dta
/Data/202223/C.dta
/Data/202223/D.dta
/Data/202223/E.dta
/Data/202223/F.dta
/Data/202223/G.dta
/Data/202223/H.dta
/Data/202223/I.dta
/Data/202223/J.dta
/Data/ASUSE/processed/ASI_NSS_figures_update.xlsx
/Data/ASUSE/processed/asi2022.dta
/Data/ASUSE/processed/multiple_data_22_merged.dta
/Data/ASUSE/processed/NSS_2022.dta
/Data/ASUSE/raw/LEVEL - 01(Block 1 & item 1503, 1508 of Block 15).dta
/Data/ASUSE/raw/LEVEL - 02(Block 2).dta
/Data/ASUSE/raw/LEVEL - 03 (Block 2.1, 2.2 & 2.dta
/Data/ASUSE/raw/LEVEL - 04 (Block 3 - section 3.1 to 3.dta
/Data/ASUSE/raw/LEVEL - 05 (Block 4).dta
/Data/ASUSE/raw/LEVEL - 06 (Block 5 - section 5.1 to 5.dta
/Data/ASUSE/raw/LEVEL - 07 (Block 6).dta
/Data/ASUSE/raw/LEVEL - 08 (Block 7.1 & 7.dta
/Data/ASUSE/raw/LEVEL - 09 (Block 8).dta
/Data/ASUSE/raw/LEVEL - 10 (Block 9).dta
/Data/ASUSE/raw/LEVEL - 11 (Block 10).dta
/Data/ASUSE/raw/LEVEL - 12 (Block 11).dta
/Data/ASUSE/raw/LEVEL - 13 (Block 11.1).dta
/Data/ASUSE/raw/LEVEL - 14 (Block 12).dta
/Data/ASUSE/raw/LEVEL - 15 (Block 13).dta
/Data/ASUSE/raw/LEVEL - 16 (Block 14).dta
/Data/ASUSE/raw/renamed/Block1-Identification of sample establishment.dta
/Data/ASUSE/raw/renamed/Block10-Land and fixed assets owned and hired.dta
/Data/ASUSE/raw/renamed/Block11-Loan outstanding.dta
/Data/ASUSE/raw/renamed/Block11pt1-Amount of loan advanced by financial establishment.dta
/Data/ASUSE/raw/renamed/Block12-Inventories on the last date.dta
/Data/ASUSE/raw/renamed/Block13-Particulars of use of information and communication technology.dta

/Data/ASUSE/raw/renamed/Block14-Calculation of net surplus.dta
/Data/ASUSE/raw/renamed/Block2-Particulars of operation and background information.dta
/Data/ASUSE/raw/renamed/Block2pt1-Activities pursued by the establishment.dta
/Data/ASUSE/raw/renamed/Block3-Principal operating expenses during last 30 days.dta
/Data/ASUSE/raw/renamed/Block4-Other operating expenses during last 30 days.dta
/Data/ASUSE/raw/renamed/Block5-Principal receipts during last 30 days.dta
/Data/ASUSE/raw/renamed/Block6-Other receipts during last 30 days.dta
/Data/ASUSE/raw/renamed/Block7pt1-Calculation of gross value added.dta
/Data/ASUSE/raw/renamed/Block8-Employment particulars of the establishment.dta
/Data/ASUSE/raw/renamed/Block9-Compensation to workers during last 30 days.dta
/Data/Codelist18.xlsx
/Data/NIC.xlsx

Statement about Rights

The use of the raw data obtained is subject to specific limitations, as outlined below:

- **Restricted Access:** Access to the data is confined to the applicant and, if applicable, bona fide users within the applicant's organization. This restriction ensures that only authorized individuals can utilize the data.
- **Purpose Limitation:** The data must be used exclusively for the purposes stated in the application. It is prohibited to use the data for proprietary or law-enforcement purposes, ensuring ethical and intended use.
- **Security Measures:** The applicant is required to implement security measures to prevent unauthorized access to the data, safeguarding it against potential breaches or misuse.

These limitations are crucial for maintaining the integrity and confidentiality of the data.

Instructions for Replicators

- The user must make sure that all the datafiles listed above are available.
- Ensure all required software and dependencies are installed as listed in the Requirements section.
- Update the directory path accordingly in the R script "msme_emissions_analysis2022V2.R"
- Run the "msme_emissions_analysis2022V2.R" file.
- The "output" folder will have relevant files need for the figures as labeled in the R script. This can be matched with the processed data in the consolidated excel "Charts.xlsx"

- To replicate this analysis for 2019-20, the user can update the code directory for importing BLOCKS A to J dataset to folder "201920" (instead of sub-folder was "202223"). This is required as some figures provide comparison between two different versions of the data, 2022/23 and 2019/20.

Guidelines for replicating 2019/20 figures

(mostly referenced as part b of the corresponding figure)

1. Firstly, update the working directory in line 23 to import blocks data from the sub-folder "201920"

Guidelines for replicating figures 3, 4 and 5

(based on* Annual Survey of Unincorporated Sector Enterprises (ASUSE) 2022-2*3)

This section provides step-by-step instructions for replicating the data processing and analysis for the ASI-NSS manufacturing sector study. The process involves importing raw microdata, cleaning and processing both organized (ASI) and unorganized sector data, merging them, and generating final figures (3,4 and5).

STEP 1: Download and Prepare Raw Data

1.1. Download NSS microdata

- Visit the MOSPI ASUSE portal
- Download all data files for 2022-23 (Levels 1 through 16)
- Save all files in: Data/ASUSE/raw/

1.2. Download ASI microdata

- Download ASI 2022-23 data files (Blocks A through J)
- Save all files in: Data/202223/

STEP 2: Rename NSS Files

Script: Data/ASUSE/rename_files.R

Input: Raw files in Data/ASUSE/raw/

Output: Renamed files in Data/ASUSE/renamed/

Instructions:

- Open R or RStudio

- Set working directory to the Data/ASUSE/ folder

- Run: source("rename_files.R")

- Verify that renamed files appear in the renamed/ subfolder

Purpose: This script standardizes file names to ensure consistent processing in subsequent steps.

STEP 3: Process ASUSE Data

Script: Data/ASUSE/NSS.do

Input: Renamed files from Data/ASUSE/renamed/

Output: Data/ASUSE/processed/NSS_2022.dta

Instructions:

- Open Stata

- Run: do "Data/ASUSE/NSS.do"

- Verify that NSS_2022.dta is created in the processed/ folder

Purpose: This script imports the renamed NSS files, cleans the data, creates key variables (employment, output, GVA, receipts), and saves the processed unorganized sector dataset.

STEP 4: Process ASI Data

Script: Data/ASUSE/ASI_extraction.do

Input: Raw ASI files from Data/202223/

Output: Data/ASUSE/processed/asi2022.dta

Instructions:

- Open Stata

- Run: do "Data/ASUSE/ASI_extraction.do"

- Verify that asi2022.dta is created in the processed/ folder

Purpose: This script imports raw ASI data files (Blocks A-J), processes employment, capital, output, and input variables, creates industry classifications, and saves the processed organized sector dataset.

STEP 5: Merge ASI and NSS Data

Script: Data/ASUSE/ASI_and_NSS.do

Input: • Data/ASUSE/processed/NSS_2022.dta

• Data/ASUSE/processed/asi2022.dta

Output: Data/ASUSE/processed/multiple_data_22_merged.dta

Instructions:

- Open Stata

- Run: do "Data/ASUSE/ASI_and_NSS.do"

- Verify that multiple_data_22_merged.dta is created in the processed/ folder

Purpose: This script combines the organized (ASI) and unorganized (NSS) sector data into a single unified dataset with consistent variable naming conventions and industry classifications.

STEP 6: Generate Final Figures

Script: Data/Figure3_4_5.do

Input: Data/ASUSE/processed/multiple_data_22_merged.dta

Output: output/ASI_NSS_figures_update.xlsx

Instructions:

- Open Stata

- Run: do "Data/Figure3_4_5.do"

- Open ASI_NSS_figures_update.xlsx in Excel

Purpose: This script generates three key figures showing the distribution of formal and informal manufacturing enterprises.

List of Exhibits

The provided code reproduces selected tables and figures in the paper, as explained below.

Exhibit Name	Output File Name	Code Line	Data Description
Figure 1	-	-	Figure 1. Evolution of CO_2 emissions by sector on the economy in India since 2000 Source: IEA 2022. "How Much CO_2 Does India Emit?" 2022. https://www.iea.org/countries/india/emissions.
Figure 2			Directly imported from the source. Not generated using the code.
Figure 3,4 and 5	ASI_NSS_figures_update.xlsx	-	Use Figure3_4_5.do file for replication. See the guidelines for replicating figures 3, 4, and 5 for generating from raw data. Note: Not all sectors' data are reported in Figure 5. Only the ones relevant for analysis are shown.
Figure 6a	co2sec.csv	503	CO2 emissions by sector (both with and without process emissions)
Figure 6b			Same as 6a but based on 2019/20 data.
Figure 7a	asi_msmecat.csv	1304	Numbers reported in the figure are derived using this table output.
Figure 7b			Same as 7a but based on 2019/20 data.
Figure 8a	asi_msmecat.csv	1304	Comprehensive summary of emissions, employment, sales, and energy data aggregated by MSME category (micro, small, medium, large)
Figure 8b			Same as 8a but based on 2019/20 data.
Figure 9	energymix.csv	1400	Energy consumption breakdown by fuel source (coal, gas, diesel, electricity) for each MSME category
Figure 10	emissionmix_msme.csv	1387	CO2 emissions breakdown by fuel source (coal, gas, diesel, electricity) for each MSME category
Figure 11a	topco2wN_withprocess.csv	491	Total CO2 emissions (including process emissions) by sector and firm size (Large vs MSMEs)
Figure 11b			Same as 11a but based on 2019/20 data.

Figure 12a	top10dataLarge.csv	576	Top 10 sectors' CO2 emissions by fuel source for large firms
Figure 12b	top10dataMSME.csv	543	Top 10 sectors' CO2 emissions by fuel source for MSME firms
Figure 13a			Same as 12a but based on 2019/20 data.
Figure 13b			Same as 12b but based on 2019/20 data.
Figure 14	msmetopcoal3digN.csv	771	Coal consumption by MSMEs across different sectors
Figure 15	largetopcoal3digN.csv	808	Coal consumption by large firms across different sectors
Figure 16a	elec_msmes.csv	825	Electricity consumption (own generation vs purchased) by MSME category
Figure 16b	coal_msmes.csv	789	Coal consumption by MSME category (micro, small, medium, large)
Figure 16c			Same as 16a but based on 2019/20 data.
Figure 16d			Same as 16b but based on 2019/20 data.
Figure 17	soe_emissions.csv	1433	SOE emissions by firm size and sector.
Figure 18	-	-	Based on coefficients reported in Table A4.1.
Figure 19, 21	asi5dig_cem_alumLarge.csv	1261	5-digit industry level analysis of cement/ceramics and aluminum sectors for Large firms, including CO2 emissions by fuel type, employment, sales, and export data
Figure 20, 22	asi5dig_cem_alumMSMEs.csv	1204	5-digit industry level analysis of cement/ceramics and aluminum sectors for MSMEs, including CO2 emissions by fuel type, employment, sales, and export data
Figure 23	-	-	Share of state in national GDP (nominal terms), 2022/23 Source: MOSPI Ministry of Statistics and Program Implementation, India (Statement 17 under data tab: https://mospi.gov.in/data)
Figure 24a, 25a, 26a	state_msme_co2N.csv	958	State-wise CO2 emissions breakdown by firm size (MSMEs vs Large firms)
Figure 24b, 25b, 26b			Same as 24a,25a,26a but with 2019/20 data
Figure 27a	stateco2decomp.csv	982	State-wise CO2 emissions decomposition by fuel type, including employment, sales, and firm count data

Figure 27b			Same as 27a but with 2019/20 data
Figure 28	statesectorMSME_ALL.csv	1540	State-wise emissions by emitting sectors (MSMEs only)
Figure 29	statesectorLarge.csv	1569	State-wise emissions by emitting sectors (Large firms only)
Figure 30	-	-	Source: Atlas of Economic Complexity, https://atlas.hks.harvard.edu/
Figure 31.a	-	-	Source: Authors' calculation using WITS. https://wits.worldbank.org/
Figure 31.b	-	-	Source: Authors calculation using WITS. https://wits.worldbank.org/
Figure 32-Zer	-	-	Source: Hasanbeigi 2010. "Energy Efficiency Improvement Opportunities for the Textile Industry." Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
Figure 33			Simulations done in excel sheet "Figure 33" but based on data in Figure 12a and 12b (sum of MSMEs and Large firms emissions for textiles).
Figure 34	-	-	Authors' assumptions. Based on factors reported in Table A1.1 and CEA (2022). Not generated from the code.
Figure 35	-	-	Simulations of grid electricity greening. Authors estimates using data underlying figure 7.a.
Table A1.1			Authors' assumptions. Based on factors reported in CEA (2022). Not generated from the code.
Figure A2.1			Generated in excel but uses data from Figure 7a and 7b.
Table A3.1			Based on definition provided as in the source. Not generated from the code.
Table A4. 1	Run the script lines for publishing latex output	1782	Regression table

Note: In the paper, a 2019-20 version of the same figure could be presented as another subset of the same figure numbering. For instance, Figure 6a indicates 2022/23 data while Figure 6b represents 2019/20 data. In such cases, the code remains the same except for the initial process of importing the 2022/23 version or 2019/20 version of the data. Update line #23 for importing 2019/20 dataset.

Requirements

Computational Requirements

Software Requirements

• R (used version 4.4.2)

- Stata (used version 18)

Memory and Runtime and Storage Requirements

The code was last run on a Windows 11 laptop with 32GB of RAM and Intel(R) Core(TM) Ultra 7 165U 2.10 GHz processor. The R scripts take less than a minute to run.

Additional details

- All raw data used for the analysis is stored in the Data folder.
- The data processing is carried out using the provided R script.
- Key outputs are clearly specified within the R script.
- Where applicable, the R script includes explicit references to published data and corresponding figure numbers (as “Data for Figure XX”).
- All data used for generating figures is compiled in the Excel file titled Charts. The dark blue shaded tab denotes the dataset extracted from R script.