The Size Distribution of Firms and Industrial Water Pollution: A Quantitative Analysis of China
Replication Instruction

XIN TANG
INTERNATIONAL MONETARY FUND

September 17, 2019

This note explains step-by-step how to replicate all the results in *The Size Distribution of Firms and Industrial Water Pollution: A Quantitative Analysis of China*. We begin with some housekeeping stuff in Section I. Section II explains in detail how to replicate all the empirical results. Section III shows how the quantitative results can be replicated.

I. Housekeeping Stuff

A. *Overview*

To replicate all the results in the paper, three softwares are used. Specifically, STATA is used for initial data processing. All the empirical results are then produced using `R`. MATLAB is used to compute all the quantitative results. There is no particular scientific reason that we use one statistical software (STATA) for data cleaning and the other (`R`) for empirical analysis.[1] We do this simply because it is easier to handle Chinese characters with STATA while `R` switches back and forth between different datasets swiftly. Veterans of STATA (or `R`) should be able to replicate all the empirical results with just one software using our source code as reference with ease. In addition, which in fact should go without saying, for readers without a licence to MATALB, GNU Octave could be used instead.

We use three datasets in our analysis: The National General Survey of Pollution Sources (NGSPS, 第一次全国污染源普查), The National Economic Census (CNEC, 第一次全国经济普查) and the Statistics of U.S. Businesses (SUSB).

- The NGSPS is a confidential dataset housed at the Ministry of Ecology and Environment. The data can only be accessed within the Ministry of Ecology and Environment as well as institutions under its direct supervision. Individual researchers requesting access should submit the application through the official portal for Information Access (政府信息公开) by the Ministry of Ecology and Environment at the following url: `http://www.mee.gov.cn/weihu/201706/t20170625_416641.shtml`. The url is retrieved on September 15th, 2019 and may vary in the future. All applications are subject to official approval by the Ministry of Ecology and Environment.

[1] Due to limited backward compatibility of STATA and that the `foreign` package of `R` supports only `.dta` files saved by STATA up to Version 12, to run the code without any modification, the reader would need STATA 12. Readers with more recent version of the STATA would need to either modify in the `.do` file everywhere `.dta` files are read and written, or use alternative `R` package (for instance `readstata13`) to convert the data in `cpsc_convert.R` and `cec_convert.R`. Compatibility of our code is not tested under such environments though. None of our results would be affected though.

- The CNEC is also a confidential dataset belonging to National Bureau of Statistics of China. Many universities and institutions have legal copy of the dataset. Our copy was obtained from Wuhan University. One portal to request access to the data for individual researchers is through the Micro-data Lab of the National Bureau of Statistics (国家统计局微观数据实验室，国家统计局-清华大学数据开发中心). The url, retrieved on September 15th, 2019, is http://microdata.stats.gov.cn/; and again it may vary in the future. All applications are subject to official approval by the National Bureau of Statistics.
- The SUSB is publicly available from the Census Bureau's dedicated website. The url, retrieved on September 15th, 2019, and subject to possible change in the future, is https://www.census.gov/programs-surveys/susb.html.

In the replication files, we cannot provide micro-level data of the NGSPS, due to the Non-Disclosure Agreement with the Ministry of Ecology and Environment. Therefore, we only post the source code that generates the results. The NGSPS is used to compute the following results:

- Tables: 1, 2, A.1, C.1, D.1 and F.1.
- Figures: 1, B.1 (upper panels), C.1, C.2, D.3.
- Regressions: 1 and the un-numbered one in Section I.B, 2, 3, the 6 un-numbered regression in Appendix C.1 and the un-numbered regression in Appendix D.2.
- Some miscellaneous information scattered across the paper.

The rest of the results can all be replicated by following this guide.

B. *File Structures*

The readers will get a number of files for replicating the results. To ease digestion, we briefly introduce the structure of the folders and files. As typographical convention, we use `blue computer modern typewriter` to indicate `folders`, and `black computer modern typewriter` to label `files`. The current folder is referred to as `./`.

- `./Empirical/`:
  - `./Data/`
  - `./Results/`
  - `./cpsc_raw.do`
  - `./cec_clean.do`
  - `./cpsc_convert.R`
  - `./cec_convert.R`
  - `./Empirical_AEJ.R`
  - `./Empirical_Appendix.R`
  - `./Accounting.R`
- `./Quantitative/`

TABLE 1—NAVIGATING THE MAIN TEXT

Content	File Name	File Dependency	Location in Code
Section I.A			
Table 1	`Descriptive.R`	`PK, PA`	
Section I.B			
Figure 1	`Empirical_AEJ.R`	`PK`	Section 1
Table 2	`Empirical_AEJ.R`	`PK`	Section 2
Regression 1	`Empirical_AEJ.R`	`PK`	Section 1
Regressions 2, 3	`Empirical_AEJ.R`	`PK`	Section 2
Regression (!#)	`Empirical_AEJ.R`	`PK`	Section 2
Section I.C			
Figure 2	`Empirical_AEJ.R`	`CA, US`	Section 3
Accounting	`Accounting.R`	`PK, CA, US`	
Section II.A			
Figure 3	`Empirical_AEJ.R`	`CA`	Section 4
Figure 4	Illustrative figure.		
Section III			
Size Distribution	`Empirical_AEJ.R`	`CA, US`	Section 3
φ_1	`Empirical_AEJ.R`	`CA`	Section 4
Clean share	`Empirical_AEJ.R`	`PK`	Section 2
k_E/Y for clean firms	`Empirical_AEJ.R`	`PK`	Section 2
Figure 5	`twosectors.m` `plot_figures.m`	`*.mat`	
Section IV.A			
Table 4	`compute_tables.m`	`*.mat`	
Table 5	`compute_tables.m`	`*.mat`	
Figure 6	`plot_figures.m`	`*.mat`	
Decomposition	`misc.m`	`*.mat`	
Section IV.B			
Table 6	`compute_tables.m`	`*.mat`	
Decomposition	`misc.m`	`*.mat`	

- ./Results/
- ./Figures/
- ./twosectors.m
- ./fcn2.m
- ./compute_tables.m
- ./plot_figures.m
- ./twosectors_ces.m
- ./fcnces.m
- ./compute_tables_ces.m
- ./misc.m

We try to group results according to their order of appearance in the paper. But this is not possible all the time. To ease the navigation, Table summarizes the results in the paper and in which file they are replicated. Acronym: `KEYFIRM_R.RData (PK)`, `ALLFIRM_R.RData (PA)`, `DLARGE_R.RData (DL)` `CNEC_avgp.RData (CA)`, `susb04.csv (US)`.

TABLE 2—NAVIGATING THE APPENDIX

Content	File Name	File Dependency	Location in Code
Appendix A.1			
Table A.1	`Descriptive.R`	`PK, PA`	
Appendix B			
Figure B.1	`Empirical_Appendix.R`	`PK, PA, DL, CA`	Section 1
Appendix C.1			
Figure C.1	`Empirical_Appendix.R`	`PK`	Section 2
Table C.1	`Empirical_Appendix.R`	`PK`	Section 2
Figure C.2	`Empirical_Appendix.R`	`PK`	Section 2
Regressions (!#)	`Empirical_Appendix.R`	`PK`	Section 2
3-Regressions (!#)	`Empirical_AEJ.R`	`PK`	Section 2
Appendix D.1			
Figure D.1	`Empirical_Appendix.R`	`CA`	Section 4
Figure D.2	`misc.m`		
Appendix D.2			
Table D.1	`Empirical_Appendix.R`	`PK`	Section 3
Regression (!#)	`Empirical_Appendix.R`	`PK`	Section 3
Fixed Cost Ratio	`Empirical_Appendix.R`	`PK`	Section 3
Appendix E.1			
Figure E.1	`Empirical_Appendix.R`	`CA, US`	Section 5
Appendix F			
Table F.1	`Accounting.R`	`PK, CA, US`	
Appendix J			
Table J	`twosectors_ces.m` `compute_tables_ces.m`	`*.mat`	

C. *Data Cleaning*

II. Empirical Results

The instructions are organized by logic instead of the order of appearance in the paper.

III. Quantitative Results

1. set `twosectors.m` to benchmark and run
2. set `twosectors.m` to no tax and run
3. set `twosectors.m` to regulation and run
4. set `twosectors.m` to flat tax and run
5. run `plot_figures.m`
6. set `compute_tables.m` to benchmark and run
7. set `compute_tables.m` to no tax and run
8. set `compute_tables.m` to regulation and run
9. set `compute_tables.m` to flat tax and run

10. manually compute the numbers for the tables