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

Ji Qi
Chinese Academy of Environmental Science

Xin Tang*
Wuhan University
International Monetary Fund

Xican Xi
Fudan University

November 19, 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 results in the main text. Section III shows how the results in the appendices can be replicated. The data and code can be downloaded from the AEA Data and Code Repository (OPENICPSR-112005).

I. Housekeeping Information

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. 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, it should go without saying that for readers without a licence to MATALB, GNU Octave could be used instead.

The specific versions we use are: STATA 12, `R` 3.5.3 (code name *Great Truth*) and MATLAB R2016a. The readers may need to do some very minor syntax change if using other versions. None of our results should be affected though. We did test our code in other versions of the softwares to the best as we could. Below are two examples:

- 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, readers with more recent version of 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`.

- The syntax for several routines was slightly changed in MATLAB R2014b. If vintage versions are used, the reader would get a complaint saying (here from R2014a for instance)

```
Error using optimoptions (line 105)
'OptimalityTolerance' is not an option for FSOLVE.
A list of options can be found on the FSOLVE documentation page.
```

*Please send all correspondence to: zjutangxin@gmail.com

The code would run through by simply fixing the syntax to R2014a version; namely, the keywords to two optimization options need to be changed to `TolFun` and `MaxIter`. The code would then run through seamlessly.

We use three datasets in our analysis: The 2007 National General Survey of Pollution Sources (NGSPS, 第一次全国污染源普查), The 2004 National Economic Census (CNEC, 第一次全国经济普查) and the 2004 Statistics of U.S. Businesses (SUSB).

- The NGSPS is a confidential dataset housed at the Ministry of Ecology and Environment (MEE henceforth). The dataset is subject to regulation by *The Law of the People's Republic of China on Guarding State Secrets* (中华人民共和国保守国家秘密法) because it contains sensitive information. Please refer to *Regulation on Archiving the National General Survey of Pollution Sources Data* (*State Environmental Protection Administration [2007] No.187*, 污染源普查档案管理办法[环发[2007]187号]). The data can only be accessed within the MEE as well as institutions under its direct supervision (中华人民共和国生态环境部及其直属单位). Researchers affiliated with these institutions should submit application for access following the internal procedure of the MEE. This is how we access the data.

  Individual researchers not affiliated with the MEE and institutions under its direct supervision need to submit the application through the official portal for requesting Information Access (政府信息公开) by the MEE at the following url: `http://www.mee.gov.cn/weihu/201706/t20170625_416641.shtml`. The url was retrieved on November 15th, 2019 and may vary in the future. All applications are subject to official approval by the General Office of Ministry of Ecology and Environment (中华人民共和国生态环境部办公厅).

  For an official introduction to the NGSPS and declassified information at the aggregated level, interested readers could refer to the book *Data Collection of the National General Survey of Pollution Sources* (《污染源普查数据集》, 中国环境科学出版社, 2011), which is publicly available. We provide a summary in English in Online Appendix A of our paper. For more information on the NGSPS, please visit the following websites: `http://env.people.com.cn/GB/8220/113963/index.html`, or `http://www.mee.gov.cn/home/ztbd/rdzl/wrypc/`.[1] The urls were retrieved on November 15th, 2019 and may vary in the future.

- The CNEC was conducted by the National Bureau of Statistics (NBS) of China in 2004, and was designed to cover the entire population of Chinese firms in that year. The firm-level data from the CNEC are not available to the general public and therefore not uploaded as part of our data files. For data access, please contact the NBS of China at `wgsjsys@stats.gov.cn`. For more information on the CNEC, please visit the dedicated page maintained by NBS at the following url: `http://www.stats.gov.cn/ztjc/zdtjgz/zgjjpc`. The url was retrieved on November 15th, 2019 and may vary in the future.

- The SUSB is publicly available from the U.S. Census Bureau's dedicated website. The url, retrieved on November 15th, 2019, and subject to possible change in the future, is `https://www.census.gov/programs-surveys/susb.html`.

We cannot provide confidential micro-level data of the NGSPS and CNEC as part of the replication files. Instead, we post the source code that generates the results. However, to facilitate replication by the readers

[1] The latter website was initially dedicated to the first NGSPS starting in 2007. Since the launch of the second NGSPS in 2017, the content of the website has been changed accordingly. However, the two surveys share similar structures.

who have access to the datasets, we post the log files when running the analysis on our computer. More details on the log files are provided at the end of this note.

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 `files`. The parent folder is referred to as `./`.

- `./Empirical/`:
  - `./Data/`: contains the raw and cleaned data.
  - `./Results/`: contains all the figures and log-files.
  - `./cpsc_raw.do`: does initial processing of the NGSPS.
  - `./cec_clean.do`: does initial processing of the CNEC.
  - `./cpsc_convert.R`: converts the NGSPS to the binary format of `R`.
  - `./cec_convert.R`: converts the CNEC to the binary format of `R`.
  - `./Empirical_AEJ.R`: produces most empirical results in the main text.
  - `./Empirical_Appendix.R` produces most empirical results in the appendices.
  - `./Accounting.R`: executes the accounting exercise.
  - `./Descriptive.R`: computes some summary statistics of the NGSPS.
- `./Quantitative/`
  - `./Results/`: contains all the results and figures.
  - `./twosectors.m`: computes the equilibrium of the two-sector model with perfect substitution used in the main text and saves the results in `.mat` format.
  - `./fcn2.m`: computes the excess demand at a given vector of price; called by `twosectors.m`.
  - `./generate_tables_main.m`: prints Tables 4, 5 and 6 in the main text.
  - `./plot_figures.m`: plots Figures 5 and 6 in the main text.
  - `./twosectors_ces.m`: computes the equilibrium of the two-sector model with constant elasticity of substitution used in Appendix J and saves the results in `.mat` format.
  - `./fcnces.m`: computes the excess demand at a given vector of price; called by `twosectors_ces.m`.
  - `./generate_tables_appendix.m`: prints Table J.1 in Appendix J.
  - `./misc.m`: draws Figure D.2 and computes the decomposition in Sections IV.A and IV.B of the main text.

All the file reference in the source code is specified by relative position. Please set the working director of your software to the current folder that hosts the file. For example, if the package is saved locally at `C:/Users/AEJ/pollution_rep/`, then to replicate the empirical results, the working directory of STATA and `R` should all be set to `C:/Users/AEJ/pollution_rep/Empirical`, while the MATLAB working directory should be set to `C:/Users/AEJ/pollution_rep/Quantitative` when replicating the quantitative results.

C. *Data Cleaning*

The NGSPS comes by 5 STATA `.dta` files. `keynum.dta` contains the information of the key sources, while `reg1.dta` to `reg4.dta` host the information of the regular sources. There are 4 files for the regular sources because of the technical limitation back when we initially applied for the data. It is likely that the reader will get one file for the regular source at this moment. These files should be placed under `./Empirical/Data/`. Executing `cpsc_raw.do` will combine the four `regx.dta` to `regall.dta` and label all the variables in `keynum.dta` and `regall.dta`. It also generates another file `allfirms.dta` where the two files are combined. All three files are saved under `./Empirical/Data/`.

Our copy of the CNEC has two STATA `.dta` files. `cec2004_large_full.dta` holds those samples that overlap with the commonly used 2004 Annual Surveys of Industrial Firms (ASIF), while `cec2004_small.dta` contains small firms not surveyed in the annual ASIF. Again, they should be placed under `./Empirical/Data/`. The file `cec_clean.do` labels all the variables. Upon completion, the original un-labeled files will be replaced by files with the same name.

With these two steps completed, the reader needs to convert the data from STATA `.dta` format to the internal binary `R` format `.RData`. The file `cpsc_convert.R` converts `keynum.dta` to `KEYFIRM_R.RData`, `regall.dta` to `REGFIRM_R.RData` and `allfirms.dta` to `ALLFIRM_R.RData`. Similarly, `cec_convert.R` converts `cec2004_small.dta` to `DSMALL_R.RData`, `cec2004_large_full.dta` to `DLARGE_R.RData`. It then extracts only the variables used in the paper and combines the two `.RData` files to one file `CNEC_avgp.RData`. All the files are saved under `./Empirical/Data/` as before.

The SUSB data are directly downloaded from the Census website as `susb04.csv`. We use the dataset in its original form.

In what follows, we will only use the files converted by `R`. A list of the essential files is provided below. All the files should be placed under `./Empirical/Data/`. For ease of reference, we use the two-letter acronyms in parenthesis when referring to the data files later.

- `KEYFIRM_R.RData (PK)`: Key sources of the NGSPS.
- `ALLFIRM_R.RData (PA)`: All sources of the NGSPS.
- `CNEC_avgp.RData (CA)`: All samples in the CNEC.
- `DLARGE_R.RData (DL)`: A subset of `CA` which overlaps with the 2004 ASIF.
- `susb04.csv (US)`: SUSB 2004.

Because we cannot post the NGSPS and CNEC publicly, the readers will not be able to run some of the files. Tables 1 and 2 list respectively which part of each file depends on which datasets and generates which result; the datasets are referenced using the above two-letter acronyms.

II. Replicate Results in the Main Text

We try to organize our source code by sections in the paper as much as we can. This is not always possible though (or creates unnecessary confusion). We do, however, break the code by sections within the file, where tasks that are reasonably connected logically are grouped together. The code is written such that each section is self-contained, so the readers do not have to run from the very beginning if she/he only wants to replicate

a certain result.[2] Table 1 summarizes the correspondence between the results in the paper (tables, figures, regression results, in-text numbers, etc.) and the section of code files that computes them, as well as data dependency. To further assist the readers to visually navigate through the code, important code snippets are always indicated by a 76-character line of =. This is the longest line in all the source code, and should be those most visually noticeable amongst all.

We now explain how to replicate the results in the main text step-by-step. The results are presented in the order of appearance in the main text.

1. Table 1 is produced by the first part of `Descriptive.R`. All the numbers are printed to the terminal. Row 1 comes first, with Rows 2 and 3 follow in order.

2. Regression (1) and Figure 1 are generated by Section 1 of `Empirical_AEJ.R`. The regression results are printed to the terminal. Figure 1 is saved at `./Empirical/Results/Figure1.pdf`.

3. Regressions (2), (3), and the un-numbered one in Section I.B part *The Role of End-of-pipe Treatment Technologies*, as well as Table 2 are produced by Section 2 of `Empirical_AEJ.R`. As one exception, Section 2 relies on Section 1 of `Empirical_AEJ.R`. The information in Table 2 is printed to the terminal in the order of Columns 2, 1, 3 and 4. The numbers in Columns 1 and 2 are directly taken from the output. Columns 3 and 4 are calculated by normalizing the physical equipment level to 100. Regression results are then printed to the terminal: first the unnumbered regression, with those of Regressions (2) and (3) follow in order.

4. Figure 2 is produced by Section 3 of `Empirical_AEJ.R`. The two panels are exported as `./Empirical/Results/Figure2_Left.pdf` and `Figure2_Right.pdf`.

5. The in-text number of the accounting exercise can be calculated from the results when executing `Accounting.R`. The 33% reduction in the last paragraph of Section I.C part *An Accounting Exercise* can then be calculated as the weighted average of the variable `ppar` with Row 1 of Table 1 being the weight. Notice that the variable `ppar` is exactly Row 3 of Table F.1. Because according to Row 1 of Table 1, the five industries combined contribute to 77% of total COD emission, assuming that industries emitting the other 23% COD remain intact, the 25% reduction in average intensity is simply calculated as $1 - (67\% \times 77\% + 100\% \times 23\%) \triangleq 25\%$. The calculation of the two numbers are automated in the code.

6. Figure 3 is produced by the second part of Section 4 of `Empirical_AEJ.R`. The two panels are exported as `./Empirical/Results/Figure3_Left.pdf` and `Figure3_Right.pdf`.

7. To compute the quantitative results, the reader would need to execute `twosectors.m` several times. There are four cases to be computed: the benchmark case, the no distortion case, the regulation case and the flat tax case. The reader needs to manually set the parameters in `twosectors.m` for each cases, namely for the distortions `tauzd, tauzc`, the regulation intensity `xi` and the file name at the very end. The code is tuned to compute and save the benchmark results initially. The code contains all the relevant statements. Hence the reader only needs to comment and un-comment the relevant sections, which should be very easy to locate by searching the variable name.

   (a) Execute `twosectors.m`. Some information would be printed to the terminal. Please refer to the log file `quantitative_main_log.pdf` for details. The results are saved in `./Quantitative/Results/benchmark_new.mat`.

[2] That said, the readers should always check to ensure that all the functional packages of R are loaded.

(b) Set `tauzd` and `tauzc` to zero, and the export destination to `./Results/notax_new.mat`. The results will be saved in this file after execution. This case is labeled as Case (i) in the paper.

(c) Set `tauzd` and `tauzc` back to the benchmark level, change `xi` to 0.355, and the export destination to `./Results/regulation_new.mat`. The results will be saved in this file after execution. This case is labeled as Case (ii) in the paper.

(d) Set `tauzd` and `tauzc` to the flat tax level, reverse `xi` back to 0.23, and set the export destination to `./Results/flattax_new.mat`. The results will be saved in this file after execution. This case is labeled as Case (i') in the paper.

8. With all the `*.mat` files computed, execute `plot_figures.m` will produce Figures 5 and 6. The corresponding files will be saved under `./Quantitative/Results/`. The list of files include `Figure5_Left.pdf`, `Figure5_Right.pdf`, `Figure6_TopLeft.pdf`, `Figure6_TopRight.pdf`, `Figure6_BotLeft.pdf` and `Figure6_BotRight.pdf`.

9. Similarly, file `generate_tables_main.m` reads the four `.mat` files and prints Tables 4, 5 and 6 to MATLAB terminal.

   The in-text decomposition of total reduction to the contribution of size distribution and technology adoption can be performed using `misc.m`.[3] The reader needs to change the second input file to either `notax_new.mat` or `regulation_new.mat`, depending on the case to compute. The role of technology adoption is printed to the terminal, with the residual being that from the size distribution.

10. Several calibration targets are computed from the data as well. This last bulletin point explains where they are computed.

    - The share of firms using clean technology 57% is calculated in Section 2 of `Empirical_AEJ.R`. The variable `clean_share` hosts the number.
    - The average adoption cost of clean technology as a ratio of output for clean firms 2.5% is calculated at the end of Section 2 of `Empirical_AEJ.R`. The number is printed to the terminal directly.
    - The value of ϕ_1 is computed in the first part of Section 4 of `Empirical_AEJ.R`. The variable `phi_quant` carries the information.
    - The firm size and employment distributions [the light (yellow) bars in Figure 5] are computed at the end of Section 3 of `Empirical_AEJ.R`. The variable `distchn` is first used to save the employment distribution and then the firm size distribution.

That concludes our instructions of replicating all the results in the main text.

III. Replicate Results in the Appendices

We now proceed with replicating the results in the appendices step-by-step. The results are presented in the order of appearance in the appendices. The correspondence between the results in the appendices and the section of code files is summarized in Table 2.

[3] The code snippet *must be* executed by using the MATLAB shortcut of 'Select → F9' unless version 2019b is used. Executing `misc.m` by using the *Run* command in MATLAB toolstrip would yield an error when calling the function `plot`. This is confirmed to be an issue by staff from MathWorks. The supporting team informed us that the issue has been fixed in MATLAB R2019b. We were not able to test the code under the environment.

1. Table A.1 is produced by the second part of `Descriptive.R`. All numbers are printed to the terminal.

2. Figure B.1 is generated by Section 1 of `Empirical_Appendix.R`. The four panels are saved respectively as `./Empirical/Results/FigureB1_TopLeft.pdf`, `FigureB1_TopRight.pdf`, `FigureB1_BotLeft.pdf` and `FigureB1_BotRight.pdf`.

3. Section 2 of `Empirical_Appendix.R` produces the results in Appendix C.1. The beginning of that section first runs the regression of intensity to size by industry. The regression results are printed to the terminal, which are exactly the results reported in Table C.1. The five panels of Figure C.1 are then exported to `./Empirical/Results/` with the names `FigureC1_TopLeft.pdf`, `FigureC1_TopRight.pdf`, `FigureC1_MidLeft.pdf`, `FigureC1_MidRight.pdf` and `FigureC1_BotLeft.pdf`. Notice that the bottom-right panel is simply Figure 1 in the main text. The code then repeats the exercise for the manufacturing sector as a whole. Results of the un-numbered regressions on Pages 7, 9 and 10 (the last one) are then printed to the terminal, with Figure C.2 exported to `./Empirical/Results/FigureC2.pdf`.

   The three un-numbered regressions on the intensity-size relationship by the type of equipment are estimated in Section 2 of `Empirical_AEJ.R`.

4. The three panels of Figure D.1 are produced by Section 4 of `Empirical_Appendix.R`. The figures are exported as `./Empirical/Results/FigureD1_TopLeft.pdf`, `FigureD1_TopRight` and `FigureD1_Bot.pdf`.

5. Figure D.2 is generated by the last part of `misc.m`. This requires the results from the benchmark case `./Quantitative/Results/benchmark_new.mat`. The file that hosts this figure would be `./Quantitative/Results/FigureD2.eps`.

6. Results in Appendix D.2 are computed by Section 3 of `Empirical_Appendix.R`. Specifically, the 4 panels of Figures D.3 are first plotted and saved as `./Empirical/Results/FigureD3_TopLeft.pdf`, `FigureD3_TopRight.pdf`, `FigureD3_BotLeft.pdf` and `FigureD3_BotRight.pdf`. A series of results are then printed to the terminal: first are the in-text numbers of the mean of the ratio of adoption cost to output for firms with output in each quintile (22%, 11%, 6.7%, 5.0% and 2.7%), second the coefficients of the un-numbered regression, with the results of Table D.3 come at last.

7. Panels in Figure E.1 are produced by Section 5 of `Empirical_Appendix.R`. The five panels are saved as `./Empirical/Results/FigureE1_TopLeft.pdf`, `FigureE1_TopRight.pdf`, `FigureE1_MidLeft.pdf`, `FigureE1_MidRight.pdf` and `FigureE1_BotLeft.pdf`. Similarly as for Figure C.1, the bottom-right panel is simply the left panel of Figure 2 in the main text.

8. The accounting exercises in Appendix F are performed by `Accounting.R`. Variables `pmed`, `preg` and `ppar` are respectively for Rows 1, 2 and 3 in Table F.1.

9. Finally, to reproduce Table J.1, one essentially follows the same logic as in the main text: namely execute `twosectors_ces.m` many times with different parameter values, print the raw numbers for the corresponding table entries and then manually compute the percentage changes in Table J.1. There are two cases in general, where CES $= 1.5$ and CES $= 3.0$. There are three parameters that are "CES-specific:" `sigmaces`, `ke` and `varces`. They are conveniently grouped together in the code, under the snippet `% CES = 1.5` and `% CES = 3.0`. The reader would also need to modify `tauzd` and `tauzc` for different scenarios. As before, all the statements have been coded; and the reader only needs to

comment and un-comment the relevant snippets of the code. Likewise, the instruction to save the results at the end of the file also needs to be changed. One complication is that the reader needs to set the initial guess for the market clearing prices. For each case, the initial guesses are conveniently stored in the code snippets at the beginning of Section `Solve the equilibrium wage`. The program is reasonably robust to different initial guesses. However, if the readers do not use the exact initial guess we use, there may be very minor numerical difference for some of the variables. Similarly, if the reader uses other versions of MATLAB (we use 2016a) or Octave, such numerical errors would probably show up as well.

(a) By default, `twosectors_ces.m` is tuned to produce the benchmark results with CES $= 1.5$. Executing the file would print some information to the terminal, again please refer to the log file `quantitative_ces_log.pdf` for detail. The equilibrium results are then saved under `./Quantitative/Results/benchmark_ces.mat`. This case is called benchmark.

(b) Set `tauzd` to zero and leave `tauzc` untouched, change the initial guesses and the export destination. The results for eliminating only the distortions in the polluting sector would be saved under `./Quantitative/Results/nodirty_ces.mat`. This is Case (ii) in Appendix J.

(c) Set both `tauzd` and `tauzc` to zero, change the initial guesses and the export destination. The results for the no distortion case would be saved under `./Quantitative/Results/notax_ces.mat`. This case is labeled as Case (i) in Appendix J.

10. Repeat Step 9 with the parameters set to CES $= 3.0$ will generate three corresponding `.mat` files:

    - `./Quantitative/Results/benchmark_ces3.mat`
    - `./Quantitative/Results/nodirty_ces3.mat`
    - `./Quantitative/Results/notax_ces3.mat`

11. With all the `*.mat` files computed, as before, `generate_tables_appendix` prints Table J.1 to MATLAB terminal. This completes the replication of our paper.

Right before our final submission, we did a final test run on all the programs. We saved all the log files of this particular run in `.pdf` format. The correspondence between the log files and the source code is as follows. Once again, to *exactly* replicate what is in these files, the reader needs to use the exact same version of the softwares as we do. The original log files are saved in text files of `.log` extension with the same names. For instance, executing the source code `cpsc_raw.do` will generate a text log file `cpsc_raw.log`, which we converted to `cpsc_raw_log.pdf` manually.

- `cpsc_raw.do` $\longleftrightarrow$ `cpsc_raw_log.pdf`
- `cec_clean.do` $\longleftrightarrow$ `cec_clean_log.pdf`
- `cpsc_convert.R` $\longleftrightarrow$ `cpsc_convert_log.pdf`
- `cec_convert.R` $\longleftrightarrow$ `cec_convert_log.pdf`
- `Empirical_AEJ.R` $\longleftrightarrow$ `empirical_main_log.pdf`
- `Empirical_Appendix.R` $\longleftrightarrow$ `empirical_appendix_log.pdf`
- `Descriptive.R` $\longleftrightarrow$ `descriptive_log.pdf`

- `Accounting.R` ⟷ `accounting_log.pdf`

The log files for the quantitative exercises are not automatically generated. We manually copied and pasted the intermediate results sent to the terminal to two text files: `quantitative_main.log` for the perfect substitute case in the main text and `quantitative_ces.log` for the constant elasticity of substitution case in the appendix. Then we convert the two files to `.pdf` format.

- `quantitative_main_log.pdf` ⟷ `quantitative_main.log`
- `quantitaitve_ces_log.pdf` ⟷ `quantitative_ces.log`

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
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	

† Acronyms for datasets:

- KEYFIRM_R.RData (PK): Key sources of the NGSPS.
- ALLFIRM_R.RData (PA): All sources of the NGSPS.
- CNEC_avgp.RData (CA): All samples in the CNEC.
- susb04.csv (US): SUSB 2004.

‡ All the *.mat files will be produced by twosectors.m.

◇ Notation (!#) stands for "un-numbered."

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	*.mat	
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	

† Acronyms for datasets:

- KEYFIRM_R.RData (PK): Key sources of the NGSPS.
- ALLFIRM_R.RData (PA): All sources of the NGSPS.
- CNEC_avgp.RData (CA): All samples in the CNEC.
- DLARGE_R.RData (DL): A subset of CA which overlaps with the 2004 Annual Surveys of Industrial Firm.
- susb04.csv (US): SUSB 2004.

‡ All the *.mat files will be produced by twosectors_ces.m.

◇ Notation (!#) stands for "un-numbered."