

**Read Me File for “Establishment Size Dynamics in the aggregate Economy”**  
**By Esteban Rossi-Hansberg and Mark L. J. Wright**

1. Complete list of the data sets used in the paper

1.1. ‘Data by industry and size (t).xls’

5 files corresponding to the following years  $t$ : 1990, 1992, 1994, 1995 and 2000.

A file contains 3 sheets. The first sheet records all possible industries which firms belong to. The industries are classified according to the SIC system (1990-1997) or the NAICS system (2000). Each industry is represented by a code. The second sheet reports the number of firms per industry and employment size category. The third sheet reports the number of establishments per industry and employment size category.

1.2. ‘Data by industry and size (1995-1996).xls’

The file contains 3 sheets. The first sheet records all possible industries which firms belong to. The second sheet reports the number of establishment deaths from  $t$  to  $t+1$  tabulated by *year* industry and size category in  $t - i$ , for each of the three years corresponding to  $i = 0, 1$  and  $2$ . The third sheet reports the number of establishment births from  $t$  to  $t+1$  tabulated by  $t$  industry and size category in  $t + i$ , for each of the three years corresponding to  $i = 1, 2$  and  $3$ .

1.3. ‘Growth data.xls’

This file contains 5 sheets. The first two sheets contain the description of the SIC codes and the NAICS codes. The three other sheets contain the number of continuing establishments and corresponding percent change in employment tabulated by industry and establishment employment size in 1990-1991, 1990-2000 and 1999-2000.

The industries are classified according to SIC for the years 1990-1991 and 1990-2000 and according to NAICS for the years 1999-2000.

1.4. ‘NADataSIC.xls’

This file reports the gross domestic product and the compensation of employees in every SIC industry category for every year between 1987 and 2001.

### 1.5. 'Adjustment.xls'

This file reports the coefficients of adjustment (of the capital share) for every SIC industry.

### 1.6. 'Naics2sic.xls'

This file contains the correspondence between the classification NAICS and SIC2 (which is SIC with some aggregation)

## 2. Derivation of the results presented in the paper

### 2.1. Preliminary remarks

The SIC systems have various levels of precision that correspond to the number of digits representing an industry. In all the empirical exercises described below, we pool the data into the following categories: 07 to 09, 10, 12, 13, 14, 15-17, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 42, 44, 45, 46, 47, 48, 49, 50 to 51, 52 to 59, 60, 61, 62, 63, 64, 65, 67, 70, 72, 73, 75, 76, 78, 79, 80, 81, 82, 83, 86, 84-87-89, and 99 so that they fit the less precise industry classification we have which is the data on GDP and Compensation of employees. These numbers correspond to the first 2 digits of the SIC, except the 84-87-89 category which is constructed from the 84, 87 and 89 industries, either by summing the relevant variables (for 'Data by industry and size' (*year*), 'Data by industry and size' (*year* – *year*+1), 'Growth data' - variable 'continuing firms') or by computing the average of the growth (equal to 1+growth rate) of firms in the previous categories weighted by the number of continuing firms ('Growth data' – variable growth rate).

The sheet 'NAICS2SIC' in the file 'NAICS2SIC' links SIC bins to NAICS bins. Note that we have to merge some bins. In order to harmonize the classification system, we will use this (partially merged) SIC system for the data using NAICS.

The size category is represented by a number that corresponds to the average of the sizes represented in a category (e.g. '2.5' contains all firms whose size is strictly positive and strictly smaller than 5), except for the highest size category. There are 2 sets of size categories (one is more precise than the other).

For the files related to "growth", the size categories (SZ1) are:

2.5, 7, 14.5, 24.5, 34.5, 44.5, 54.5, 64.5, 74.5, 84.5, 94.5, 149.5, 249.5, 349.5, 449.5, 549.5, 649.5, 749.5, 849.5, 949.5, 1249.5, 1749.5, 2249.5, 2749.5, 3249.5, 3749.5, 4249.5, 4749.5 and 8000.

For the files related to "entries" and "exits", the size categories (SZ2) are: 2.5, 7, 12, 17, 22, 27, 32, 37, 42, 47, 54.5, 64.5, 74.5, 84.5, 94.5, 112, 137, 162, 187, 212, 237, 274.5, 324.5, 374.5, 424.5, 474.5, 549.5, 649.5, 749.5, 849.5, 949.5, 1124.5, 1374.5, 1624.5, 1874.5, 2124.5, 2374.5, 2749.5, 3249.5, 3749.5, 4249.5, 4749.5, 7499.5 and 12500.

Consequently, we can define an observation as an n-uplet of the following form (industry code, size category, other relevant variables ...). Note that we dropped all the

observations which have at least one missing value in all the following regressions, as well as the observations concerning the industry represented by 99. Finally, note that we drop in every regression the firms with 0 employees.

## 2.2. Construction of the capital shares

As mentioned in the paper, the physical capital share (KS) of a given industry is one minus the labor share of this industry.

For the data partitioned in SIC, the labor share is constructed for every year between 1987 and 2001 as the ratio of the compensation of employees over gross domestic product from the file “NADDataSIC”. The KS for every year are stocked in the file ks.dta.

The adjusted physical capital share (adjks) is the product of the physical capital share in a given industry by its coefficient of adjustment. We also construct the average capital share over 1990-2000, denoted ksav, and the adjusted average capital share over 1990-2000, equal to the average capital share times the adjustment coefficient, denoted ksavadj (= ksav x adjustment coeff.)

Similarly, for the data partitioned in NAICS, we use the file “NADDataNAICS” which contains the same variables as “NADDataSIC” for the (partially merged) SIC system.

## 2.3. Figures

All the figures of the paper are constructed directly from the data. ‘Manufacturing’ comprises SIC=20-- or NAICS=31-33 and ‘Educational Services’ comprises SIC= 8200 and NAICS=61.

## 2.4. Conditional Growth Rates Between 1990 and 2000 - Tables 1 and 2

We transfer the sheets ‘Growth (1990-2000)’ from Growth Data.xls and ‘Table 2’ from Data by Industry and Size 1990.xls into the files growth9000.dta and est1990.dta. Then, the program Growth.do is run. Note that some growth rate are not provided for reasons of confidentiality (denoted (D) in the .xls file). We arbitrarily suppress these observations from the data.

The program Growth.do consists of the following steps:

Construct KS (ks.dta)

Construct Growth (intgrowth9000.dta)

Use growth9000.dta

Aggregate data into the relevant industry bins.

The growth rate (Growth9000) of a given bin is equal to the average growth (g9000) from the initial industry bins, weighted by the number of surviving firms (ctg9000), minus 1.

The number of surviving firms (ctg9000) of a given bin is equal to the sum of the number of surviving firms from the initial bins

Define size bins by the lowest size (minsize) or the average size (avsize) they contain.

Construct Establishments (intest1990.dta)

Aggregate data into the relevant industry bins.

Harmonize size bins with size bins from growth data.

Define Variables for Regressions

Define ksav, ksavadj, weigths, etc.

Run regression  $\ln(1+g_j) = a_j + b \ln(x_j) + e \alpha_j \ln(x_j) + \epsilon_{j,t}$  with different weights and KS's → Table 1

Manufacturing Industries are defined by SIC  $\in [2000; 3900]$  and Non-Manufacturing by SIC  $\in [700; 2000) \cup (3900; 8900]$

Run regression  $\ln(1+g_j) = a_j + [b^m \ln(x_j) + e^m \alpha_j \ln(x_j)] \cdot \mathbf{1}_{\{j \text{ is a manufacturing industry}\}} + [b^{nm} \ln(x_j) + e^{nm} \alpha_j \ln(x_j)] \cdot \mathbf{1}_{\{j \text{ is a non manufacturing industry}\}} + \epsilon_{j,t}$  with different weights and KS's → Table 2

## 2.5. Net Exit Rate – Table 3

We transfer the sheets 'Table 2' from the files 'Data by Industry and Size (t).xls' for  $t=1994, 1995, 1997$  and 'Table 3A' and 'Table 3B' from the files 'Data by Industry and Size (1995-1996).xls' into est1994.dta, est1995.dta, est1997.dta, Exits9596.dta and Entries9596.dta. Then, the program Net Exit Rate.do is run.

This program comprises the following steps:

Construct KS (ks.dta)

Construct estall.dta, containing all data besides KS

Merge all the files

Aggregate data into the relevant industry bins

Define size bins by the lowest size (minsize) or the average size (avsize) they contain.

Merge estall.dta and ks.dta

Define Variables for Regressions<sup>1</sup>

Define ksav, ksavadj, weigths, etc.

Run regression  $\ln[1+2*(l1995-l1996)/(est1995+est1996)] = a_j + b \ln(x_j) + e \alpha_j \ln(x_j) + \epsilon_{j,t}$

## 2.6. Size Distributions of firms - Table 4

The program Size Distribution 1990.do is run. It comprises the following steps.

Construct KS for 1990 (ks90.dta)

Construct size distribution

---

<sup>1</sup> In particular, the number of establishments in 1996 is computed by adding the number of entries between 1995 and 1996 (within the set of establishments existing in 1996 – denoted l1996) to the number of establishments in 1995 (est1995) and subtracting the number of exits between 1995 and 1996 (within the set of establishments existing in 1995 – denoted l1995). By doing this we implicitly neglect the number of firms which move from a size or industry bin to another

Aggregate data into the relevant industry bins  
 Define size bins by the lowest size (minsize).  
 Compute the total number of firms in every industry bin  
 Compute  $P_j$  = number of firms with size large than minsize in the same industry bin  
 Merge with ks90.dta.  
 Run regression  $\ln(P_j) = a_j + b_j \ln(x_j) + d (\ln(x_j))^2 + e \alpha_j (\ln(x_j))^2 + \epsilon_j$

We transfer the sheet 'Table 2' from the files 'Data by Industry and Size (2000).xls' into est2000.dta and the upper half of the sheet of the file 'NADDataNAICS.xls' into gdpnaics, dta and the lower half of the sheet of the file 'NADDataNAICS.xls' into cenaics, and naics2sic.xls in the latter into gdpnaics.dta, cenaics.dta, and naics.dta. Given the fact that these classifications overlap, we need to aggregate both some industries from SIC and NAICS to get a classification within which we can adapt both types of data. This classification system is named SIC2. The file naics2sic.dta (from the sheet naics2sic in New Naics Data.xls) contains the correspondence from naics to SIC2. The correspondence between SIC and SIC2 can be clearly deduced from the codes of SIC2. To construct KS, we have 2 files with compensation of employees (cenaics) and gdp (gdpnaics) which are already classified according to SIC2. A file of adjustment coefficients for SIC2 is constructed by computing, when needed, the average of the coefficients of the file with the SIC classification.

Then, the program Size Distribution 2000.do is run. It comprises the following steps.

Construct KS for 2000 (ks00.dta)  
 Construct size distribution from est2000.dta  
     Use naics2sic.dta to transfer data in SIC2  
     Aggregate data into the relevant industry bins  
     Define size bins by the lowest size (minsize).  
     Compute the total number of firms in every industry bin  
     Compute the proportion of firms with size large than minsize in the same industry bin  $P_j = \sum_{k \text{ s.t. size } k \geq \text{size } j} N_k / \sum_k N_k$  where  $N_k$  is the number of establishments in the size bin  $k$ , for a given industry.  
 Merge with ks00.dta  
 Run regression  $\ln(P_j) = a_j + b_j \ln(x_j) + d (\ln(x_j))^2 + e \alpha_j (\ln(x_j))^2 + \epsilon_j$