Read me file

Studying Firm Growth Distributions with a Large Administrative Employment Database
By Dixon, Petrunia, and Rollin

Details on the database used

The Longitudinal Employment Analysis Program (LEAP) dataset used in the paper entitled "Studying Firm Growth Distributions with a Large Administrative Employment Database" is housed and maintained by Statistics Canada. Since LEAP is a confidential business micro dataset, access to the LEAP is restricted.

Statistics Canada established the Canadian Centre for Data Development and Economic Research (CDER) in 2012 to facilitate access to researchers and provide secure access to business microdata for analytical purposes. The LEAP database used in this paper is available at CDER. For a list of all available datasets, or to learn more about data access, visit the CDER website, http://www.statcan.gc.ca/eng/cder/index (website accessed on March 15, 2018). Note that in order to preserve the confidentiality of the data, analysts from Statistics Canada vet all results prepared by researchers prior to release.

Details on SAS and STATA files

0_ExtractLEAPdata.sas

This SAS code extracts the LEAP data, applies some restrictions and create some variables.

- Extract the LEAP data from various vintages (vintages 2000 to 2009 are used in the paper)
- Only keep businesses :
    - With payroll in the final year or the second to last year of the vintage (the current or previous year).
    - Not classified in Health, Education or Government industries (exclude NAICS 61, 62 and 91)
- For vintage 2000, recode the industry code from the SIC to the NAICS classification.
- Create analytical variables
    - incumbent/entry/exit indicator
    - sector
    - knowledge industries flag (not available for year 2000)
    - size class
    - age class
    - DHS growth rate
    - DHS growth rate class (0.05 bins used in the charts)
    - Main province and region

1_FinalizeData.sas

This SAS code finalizes the dataset used in all remaining steps.

- Stack the data from vintages 2000 to 2009 in a single dataset.

- Apply restrictions. Keep only:
    - Incumbent firms, with employment in both the current and previous years.
    - Firms with an average size less than 1 ALU (average over current and previous years)
- Create more analytical variables
    - ICT industries flag (not available for year 2000)
    - Size/age classes (small and young; small and old; large and young; large and old)
    - LowGrowth10 and HighGrowth10 flags (growth rate above or below 10%; not presented)
    - LowGrowth20 and HighGrowth20 flags (growth rate above or below 20%; presented)
    - LowGrowth30 and HighGrowth30 flags (growth rate above or below 30%; not presented)
    - LowGrowth40 and HighGrowth40 flags (growth rate above or below 40%; not presented)
    - LowGrowth50 and HighGrowth50 flags (growth rate above or below 50%; not presented)

2_DescriptiveStats.sas

This SAS code generates the summary statistics presented in the various tables of the paper. Statistics are obtained for the overall distribution, and across all dimensions studied in the paper (years, industries, regions, size classes, age classes, and size/age classes).

- Mean
- Standard deviation
- P10, P25, P50, P75, P90, Interquartile range

3_Shares_Low_High_Growth.sas

This SAS code calculates the proportion of firms above and below certain growth thresholds. Shares are obtained for the overall distribution, and across all dimensions studied in the paper (years, industries, regions, size classes, age classes, and size/age classes).

In the various tables of the paper, we only present GR(-20) and GR(20) since we define High-Growth and Rapidly Shrinking Firms as having growth rates above 20% and below 20% respectively. However, alternative growth rate thresholds were also used to assess where the 'tick tails' of the distributions start. Shares are calculated using the variables created previously:

- LowGrowth10 and HighGrowth10 flags (growth rate above or below 10%; not presented)
- LowGrowth20 and HighGrowth20 flags (growth rate above or below 20%; presented)
- LowGrowth30 and HighGrowth30 flags (growth rate above or below 30%; not presented)
- LowGrowth40 and HighGrowth40 flags (growth rate above or below 40%; not presented)
- LowGrowth50 and HighGrowth50 flags (growth rate above or below 50%; not presented)

4_Statistical_Tests.do

This STATA code conducts the various statistics tests presented in Table 9 of the paper. Tests are performed for all dimensions studied in the paper (years, industries, regions, size classes, age classes). The different tests performed are:

- Normality tests (check for skewness and kurtosis)

- Equality of distribution tests
    - K-sample tests (check for equality across all categories)
        - Kruskal-Wallis equality of distributions tests
        - Levene's equality of variance tests
        - Nonparametric K-sample equality of medians tests
    - Pairwaise tests (check for equality between a pair of categories)
        - Kolmogorov-Smirnov equality of distributions tests
        - Wilcoxon rank-sum distribution tests
        - Equality of means tests
        - Equality of variances tests