Building a Granular Dataset of UK Companies

Alfred Holmes

September 2018

Abstract

The UK government, through the Office for National Statistics and Companies House release detailed data on UK companies which can be combined to generate. We show that, using publicly available data, it is possible to track the location and assets of companies through time and using few assumptions assign branches, employees and turnover to these companies in such a way as to match the Office for National Statistics annual reports.

1 Introduction

To use contemporary techniques to model the UK economy, a detailed granular dataset of UK companies is required. Ideally the dataset would contain the employment size and turnover evolution of each company, as well as the location and number of employees of individual branches through time. From this, one could test hypotheses about company location decisions, track the flow of employment and generate a detailed picture of the changes in the UK economy easily.

1.1 Definitions

These definitions are used constituently throughout this document, and provide clarity when describing subtly different entities.

- Company - an entity registered on Companies House. In June 2017 there were 3.1 million registered companies in the UK.
- Enterprise - a business that is reported by ONS. We assume that an enterprise is a collection of companies. In 2017 the ONS reported 2.7 million enterprises.
- Local Unit - a site where an enterprise operates, e.g. a branch of a large supermarket chain or the only location from which a small firm operates.
- Local Authority - a connected area of land the UK governed by a council. At the time of writing, local authorities have a population of 170 000 people and 7 000 enterprises on average. There are 391 local authorities in the UK.

2 Data

The UK government through Companies House and the Office for National statistics release detailed data. All data used in this study is provided through the Open Government License.

2.1 Companies House

Companies House is the public service responsible for incorporating and dissolving companies in the UK, as well as storing and releasing company data. They provide monthly snapshots of basic company information - names, registered office addresses, SIC codes - as well as all the online accounts that have been filed since 2008 and also provide an API where more detailed information - changes of address and people of significant control - can be accessed on a company by company basis.

2.1.1 Snapshots

The companies house snapshots are only available for the current month and only contain active companies and don't contain data on the history of each company. Luckily webarchive.org has archived approximately 1 snapshot per year, so this can be used to get low resolution historic data for companies. The

2.1.2 Accounts

If a company files its accounts online then that account filing is available to download through the companies house accounts data product. Some small businesses are exempt from filing accounts, but for 2012 xx % of companies filed accounts with 75% of those accounts containing the current assets of the company. This increases to xx % and xx % respectively for 2017. The files are given as one XML (or HTML) file per account file per company per year. The most effective way to read these files is to recursively see if there is a number between two tags and if there is pull the number, its title and date.

2.1.3 Companies House API

The Companies House API can be used to get detailed information about a company given its name or company number. The API has a request limit of 2 requests per second per API key, so to make one query per company for all companies registered since 2012 it would take 32 days. The service does allow multiple API keys to be registered which can speed up data acquisition. The API can be used to get all the filings for a particular company, with the exact dates of each.

2.2 Office for National Statistics

The Office for National Statistics releases yearly summery datasets

2.2.1 Business Activity, Size and Location

2.2.2 Other Useful Data

Population

Employment

3 The Log Normal Hypothesis

To make the ONS data usable some sort of size distribution needs to be assumed. Using Maximum Likelihood Estimation (MLE) and the size bins, it is possible to fit distributions to the data. This has two benefits in that it assigns the mass of the probability distribution in a band in a sensible way and also removes the irritating *250+* size band.

3.1 MLE and it's Bias

Assuming a log normal distribution where the size X of a company is such that $\log X \sim \mathcal{N}(\mu, \sigma^2)$, μ and σ can be calculated by maximising the log likelihood, given by

$$l(\mu,\sigma) = \sum_i n_i \log\left(\Phi\left(\frac{\log a_{i+1} - \mu}{\sigma}\right) - \Phi\left(\frac{\log a_i - \mu}{\sigma}\right)\right) \tag{1}$$

since $\mathbb{P}(a_i < X < a_{i+1}) = \mathbb{P}(\log a_i < \log X < \log a_{i+1}) = \Phi\left(\frac{\log a_{i+1}-\mu}{\sigma}\right) - \Phi\left(\frac{\log a_i-\mu}{\sigma}\right)$.
Typically when dealing with bias from maximum likelihood estimates the bias depends on the number of observations in the sample and perhaps the values of the parameters themselves. Since the ONS data is binned and the sample size is large for each local authority, we assume that the sample proportions for each size band are the actual proportions for the distribution, to remove the dependence of the bias on the number of companies in the local authority. This can be justified by running simulations. We can then run simulations to examine the effects of the distribution parameters on the difference between the expected value of the MLE and the actual parameters.

References

[1] Hart, P. E., Oulton N. (1997)
Zipf and the size distribution of firms, Applied Economics Letters, 4:4, 205-206
DOI: 10.1080/758518494