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. Using this publicly available data, it is possible to track the location and assets of companies through time and using few assumptions assign employment sizes to these companies and reproduce ONS reported data. This report summaries the available data and details the use of the data to compile a granular dataset.

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 easily generate a detailed picture of the UK economy through time.

1.1 Definitions

These definitions are used consistently 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 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 Available 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 [1] 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 [2] has archived approximately 1 snapshot per year, so this can be used to get low resolution historic data for companies.

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 [3]. This means that for increasingly many companies detailed financial data is available. 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. The `data/accounts` folder in the GitHub repository [4] contains useful python scripts to process the accounts data. The following table summarises the available accounts data. The percentages proportion of accounts that contain the heading. There is a high correlation between having one of the heading and having the other. The proportion of filed accounts of active companies to active companies increases from 35% in 2008 to about 60% in 2017.

Year	Number of Accounts	netcurrentassetsliabilities	currentassets
2008	185163	59%	58%
2009	344403	64%	63%
2010	585143	71%	70%
2011	793218	74%	73%
2012	1033277	76%	74%
2013	1219357	77%	71%
2014	1573870	77%	72%
2015	1811446	77%	72%
2016	2071697	78%	72%
2017	2324272	79%	73%

2.1.3 Companies House API

The Companies House API [5] 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. See the folder `data/API` in [4] for a script that pulls the changes of address for companies.

2.2 Office for National Statistics

2.2.1 Business Activity, Size and Location

The Business Activity, Size and Location tables from the ONS [6], available for 2012 to 2017, report by local authority and SIC code the number of companies with size in a particular range. For example, for employment statistics the ONS typically give the number of companies with employees in the ranges *0-4, 5-9, 10-19, 20-49, 50-99, 100-249* and *250+*. The turnover is treated similarly.

2.2.2 Postcode Data

Since the data by location in 2.2.1 is by local authority, a useful resource is the postcode lookup tables [8]. These tables allow addresses of individual companies to be matched to local authorities. Included in the folder `useful` in [4] is a slimmed down postcode table with just the postcode and its associated local authority's id as well as `lad_17_geo_info.csv` which is a collection of the possible names on ONS documents and region information for each local authority[1].

2.2.3 Employment

The ONS also report employment [7] by location and also by SIC code broken down into public and private sector jobs. We assume that all the companies listed on companies house are private entities, and hence any employees are working in the private sector. The employment is useful as it allows the mean company size to be calculated.

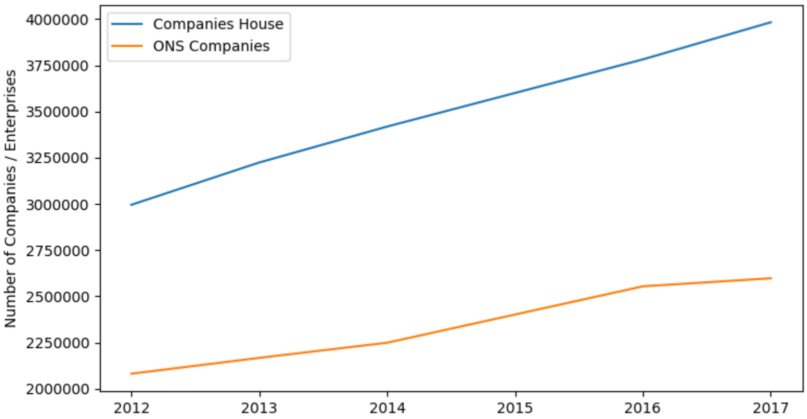
3 Issues Combining ONS and Companies House data

3.1 Number of Companies

Looking at figure 1 it is clear that an ONS enterprise isn't just a companies house company. This is because many enterprises are made up of multiple companies. Because of this issue, for this report we assume that if two companies share the same set of addresses then they are part of the same enterprise. This reduces the error

[1] Local authority names sometimes have different spellings / word orders in different ONS datasets

Figure 1: Total Number of Companies registered on Companies House and reported by ONS



between the companies, but there is still a large difference between the number of companies and enterprises. Due to this issue, we will use proportions in this report rather than actual numbers of companies and enterprises.

3.2 SIC Codes

Companies House companies can have multiple SIC codes and also have 5 digit SIC codes. The ONS release reports contain the number or enterprises with a particular 4 or 2 digit SIC code. It is not clear how they decide which SIC code to give to a particular enterprise, but it it likely that the enterprise give the SIC code in response to a survey so it would be difficult to infer the SIC code from an enterprise's companies as from the data it is difficult to work out the main business of an enterprise made up of many companies, each with multiple SIC codes. For this reason, the 4 digit SIC code data given by ONS does not match the companies house companies very well, but the 2 digit SIC codes match reasonably well. The 2 digit SIC codes are better because if a company reports two SIC codes, then it is likely that are in the same broad industry group and so they share the first two digits.

4 Fitting Lognormal Distributions

To make the ONS data usable some sort of size distribution needs to be assumed. This is not ideal since there are many different factors going in to company growth that can't be captured by simple growth models. Using Maximum Likelihood Estimation (MLE) and the ONS size bins as well as the mean company sizes, 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.

4.1 MLE

Assuming a lognormal distribution where the size X of a company is such that $\log X \sim \mathcal{N}(\mu, \sigma^2)$, μ and σ can be estimated 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 $\mathbb{P}(a_i < X < a_{i+1}) = \mathbb{P}(\log a_i < \log X < \log a_{i+1}) = \Phi\left(\frac{\log a_{i+1} - \mu}{\sigma}\right) - \Phi\left(\frac{\log a_i - \mu}{\sigma}\right)$.
This parameter estimation is limited because the data is binned, and running experiments with we find that there is an error or bias with the recovered parameters that varies with the distribution parameters. For local authority and country wide parameter fitting, the total private employment, e, is released by ONS. From this, as the number of companies, n, is known the mean of the distribution can be used to constrain the parameters such that μ and σ satisfy the equation $\exp(\mu + \sigma^2) = \frac{e}{n}$ in order to reduce the error of the predictions.

4.1.1 Simulations

In order to test whether the MLE can recover the distribution parameters from the size bins we ran simulations: picking random numbers from a known distribution and then trying to recover the distribution parameters. In this section we focus on recovering the employment parameters, so use parameters which generate distributions that fill the ONS company sizes size bins. For these simulations we assume that the number of companies is large enough that the number of companies in each bin is the expected number of companies. The python scripts for these simulations can be found in the `Lognormal Bias` folder in [4]. Running the script `analysis.py` gives an interactive version of the 3D graphs in figure 2.

Figure 2: Simulated Bias. The spikes are due to the binned data being ambiguous - at least two sets of parameters lead to the observed binned distribution.

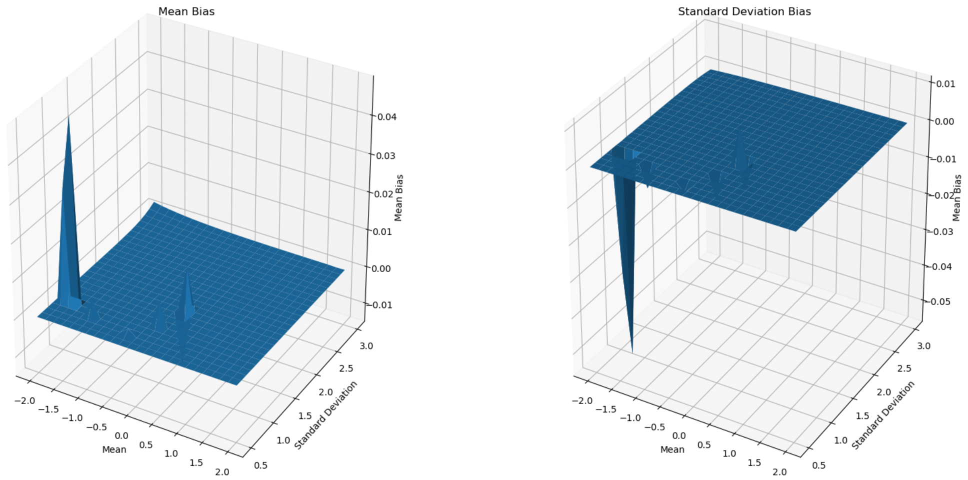
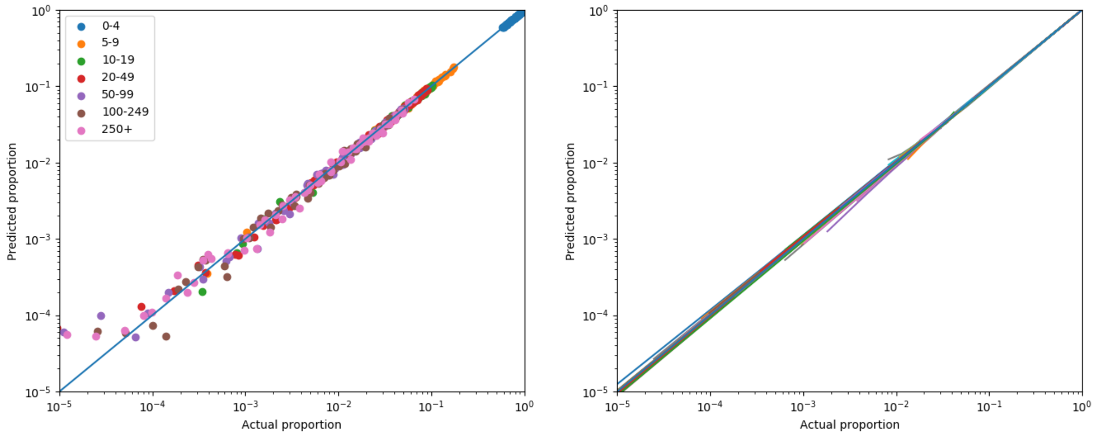


Figure 2 shows that the parameter estimation generally works quite well for the typical parameters in the observed distributions. In order to test the fit of the fitted parameters, we estimate the parameters from the ONS binned data and then use the distribution parameters to calculate the expected proportion of companies with in each bin and then plot those proportions. Although not the most mathematically rigorous method, this is a quick way to get a clear picture about the parameter estimation.

Figure 3: Plot to evaluate parameter estimation. 100 points plotted with $(\mu, \sigma) \in [-1, 1] \times [0.5, 3]$. The left plot contains 7 points for each parameter pair, and the right plot is the same data but connecting each set of points up.

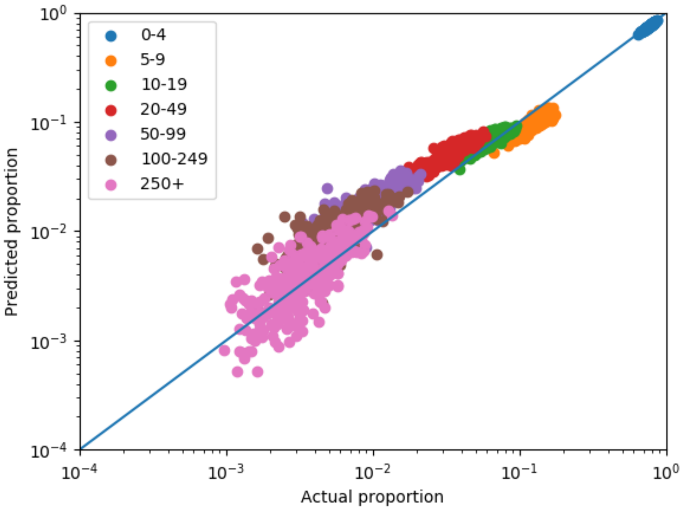


The results in figure 3 show that if the size distribution of companies is approximately log normal then using the available ONS data, it is possible to recover the parameters of the underlying distributions.

4.1.2 Matching Enterprise Employment

The same procedure can then be applied to the ONS data. Figure 4 indicates that although the underlying distribution of enterprise sizes is not lognormal, the fitted distribution does an adequate job of predicting the enterprise sizes.

Figure 4: 2014 ONS Employment size. Each point represents a size bin of a local authority.



5 Assigning Sizes to Enterprises

For this section we will focus on the year 2014. The same approach can be used for all the other years in the sample. To assign the size to companies we assume the following:

- Companies who do not report their assets have assets of 0 value.[2]
- If company a has more assets than company b and a and b have the same SIC code, then company a employs more people and company b.

We then pull n numbers from the fitted lognormal distribution of company sizes for each SIC code, where n is the total number of companies with each SIC Code, and sort the list. We then sort the companies by asset size and assign company i size i. Using the same technique as in 4.1.2 we then plot the predictions (figure 5).

5.1 Evaluating the the assigned sizes

One simple test and use for the data set is to try and reproduce the ONS reported private employment rates. To do this we use the sizes generated using the SIC code distributions and add up the enterprise sizes. Due to the issues raised in section 3.1, we test the employment predictions by looking at the ratios of total employment, rather than plotting absolute employment statistics.

We see in figure 6 that the employment predictions are reasonable - with a gradient of approximately 1. There is a fair amount of error in the data but this is to be expected, as the prediction assumes that all the workers for a company work in that companies registered office location, which is not true for many jobs. This is seen in the outlier, Westminster, since central London is a common place for large company headquarters with companies such as *Marks and Spencer*[1].

6 Improvements

To improve this dataset the most important issue to sort out is the issue discussed in section 3.1 since this would allow a much finer understanding of how the ONS data relates to the companies house entities. We found that if the assumption that all enterprises with a given SIC code, s have a probability p_s of having an enterprise (set of companies with the same set of addresses) is made then by finding p_s for each s we could improve the predictions. The problem with this approach is that it creates two granular sets of enterprises, the ones from companies house and the set with the correct number of companies, and it is difficult to use the companies house data to then infer the properties of the second set.

[2] This assumption is made since when processing the data it is difficult to distinguish whether a company did not report assets or is exempt from reporting the data.

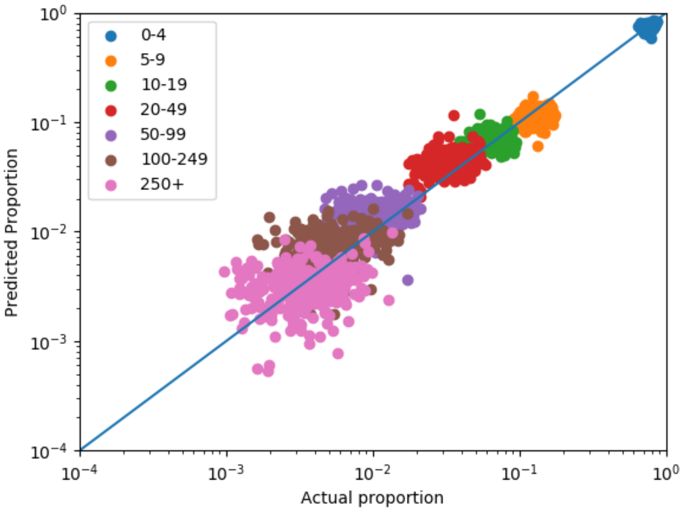


Figure 5: Local Authority Employment Reconstruction using SIC code distributions

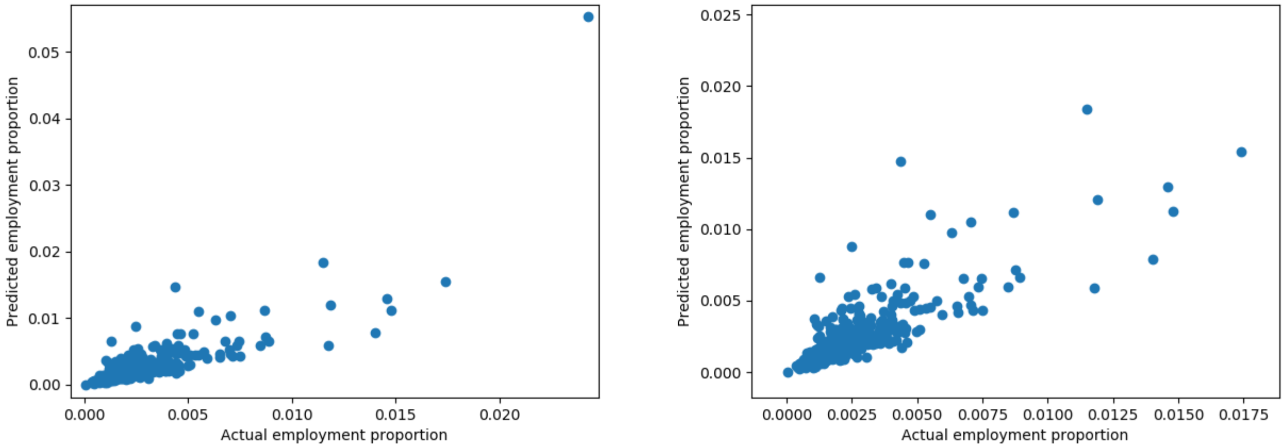


Figure 6: Local Authority Employment Reconstruction using SIC code distributions. The right is a zoomed in version of the left

Another important improvement would be to map and assign branches to companies. ONS release data on the number and size of local units in each local authority, so this could be used to assign branches and associated employees to enterprises. Since the technique for assigning employees to enterprises in section **??** doesn't use the location of the company, there are no ties for the employees of an enterprise registered in a particular local authority to work there. From the ONS reports, it is not clear how the location of an enterprise is assigned - whether it is the registered office address of the enterprise, or where the company is acting. A clever assignment of branches to the enterprises would allow the size assignations to predict employment accurately and also track the flow of employment around the UK using the companies house migration data.

References

[1] Companies House Snapshot: list of basic company information for the current month's UK registered active companies
`http://download.companieshouse.gov.uk/en_output.html`

[2] Companies House Snapshot Archive
`https://web.archive.org/web/20120901000000*/http://download.companieshouse.gov.uk/en_output.html`

[3] Accounts Data Product
`http://download.companieshouse.gov.uk/historicmonthlyaccountsdata.html`

[4] UK Company Data GitHub repository
`http://github.com/alfredholmes/uk-company-data`

[5] Companies House API
`https://developer.companieshouse.gov.uk/api/docs/`

[6] Business Activity Size and Location
2012:
`http://webarchive.nationalarchives.gov.uk/20140511113648/http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-254601&format=hi-vis`
2013:
`http://webarchive.nationalarchives.gov.uk/20140510023623/http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-313744&format=hi-vis`
2014 - 2017:
`https://www.ons.gov.uk/businessindustryandtrade/business/activitysizeandlocation/datasets/ukbusinessactivitysizeandlocation`

[7] ONS Employment Statistics
`https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/datasets/summaryoflabourmarketstatistics/current`

[8] ONS Postcode Data
`http://geoportal.statistics.gov.uk/datasets/postcode-to-output-area-to-lower-layer-super-output-area-to-middle-layer-super-output-area-to-local-authority-district-august-2018-lookup-in-the-uk`