Building a Granular Dataset of UK Companies

Alfred Holmes

September 2018

Abstract

Given employment size distributions of enterprises by location and industry individually, and granular company registry data, it is possible to produce a detailed granular dataset of UK companies (containing the age, employment size, number of branches and location through time of each company) by assuming a simple growth model. To study the accuracy of the size assignments, it is possible to develop simulations to test the ability of the assignment methods to reproduce model parameters.

1 Introduction

Contemporary economic modeling requires detailed granular data. Typically highly granular data is not available due to difficulties collecting such data or privacy issues in releasing the information. In the case of UK companies data, an ideal dataset would contain the size (employment and turnover) along with location through time and branches of the companies operating in the UK. The UK Companies House service provides basic company data on active companies, which can, as detailed in this report, be used to provide the backbone of a dataset. The entities registered on companies house are used for legal purposes for the government to track companies operating in the UK, and this legal information is freely available and so does not contain useful data such as the location of branches, number of employees and turnover information[1]. The ONS does produce detailed agglomerated data, giving size distributions by location and company type, but these distributions do not give data about different companies. Using these distributions, the companies house data and simple models of company growth, it is possible to assign sizes to the companies house entities to compile a detailed granular dataset of UK companies and their branches. This report explains the process of producing that dataset and its evaluation using other ONS data.

1.1 Data

1.1.1 Companies House

Companies House release a monthly snapshot of basic data (Company ID, Address, SIC Code) of active companies on their service. This data is only available for the current month - they do offer the purchase of a DVD with historic data - but previous months (approximately one for each year) can be accessed through websites which archive the internet. There is also an API that can be used to get more detailed data (address changes, number of registered people of significant control) which has been used in this study to get company migrations. These migrations can be inferred by changes in location between two snapshots, but in using the API the exact date of the change of location is known. The API is also required to get the correct date of death of companies, since the monthly snapshots only contain active companies.

1.1.2 ONS

The office of national statistics provide a large range of agglomerated statistics. For use assigning values to companies, it's useful to have one statistic broken down in multiple ways. For example in the business activity, size and location tables, ONS release the number of enterprises with 0-4 employees in each local authority as well as the number of enterprises with 0-4 employees with a certain SIC code. This allows properties to be assigned to companies in a more accurate way since the extra data reduces the size of the solution space of the problem.

1.2 Description of Terms

Since there are different types of entities in each of the datasets, it is worth defining terms to improve the clarity and precision of descriptions.

- Company - an entity registered on Companies House

[1] For large companies, they are required to publish turnover information, but this is not in a machine readable format

- Enterprise - a business that is reported by ONS. We assume that an enterprise is a collection of companies.
- 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.

2 Predicting Enterprise and Local Unit size distributions from companies

2.1 Assigning Companies to Enterprises

Each enterprise is made up of one or more companies. A quick way of choosing the groups of companies that make up an enterprise is to match the addresses. The motivation for this is that enterprises may register companies to handle a fraction of their total business, for example on companies house TESCO is a registered company as well as TESCO Holdings, but in making the assumption that TESCO will be reported as one enterprise, it makes sense to omit the TESCO Holdings. This results in a more linear relationship between the number of companies and enterprises in each local authority.

Figure 1: Improvements in grouping enterprises by addresses

After assigning companies to enterprises in this way there is a non trivial number of enterprises missing. This could be because some companies have exempted themselves from reporting to companies house, or an issue with grouping the companies by address. In total for 2017, after the companies have been grouped together there are about 1.6 times as many enterprises as company groups. An effective way to get around this is to assign ratios by SIC code and assume that the enterprises that aren't registered are a representative sample of all the enterprises.

Figure 2: Time series SIC ratios, gradient and average value histograms

2.2 Assigning sizes to Enterprises

Fitting a log normal distribution to the ONS data for employment size by local authority, and then picking numbers from that distribution to assign to enterprises predicts the total employment, and the sizes of the ONS sizebands well (as expected) but doesn't predict the national SIC size distributions well. The same is true for trying to make predictions using the SIC size distributions, indicating that the size of enterprises depends both on the location and the industry. Unfortunately the ONS doesn't release a dataset of the sizes of companies by SIC by local authority[2]. In order to get around this lack of data, it is possible to use the companies house data to get detailed SIC code distributions for each local authority and then force the assigning of companies to respect both the size distributions[3].

2.3 Model for company growth

In order to make the solution space of size assignments smaller, we assume a growth model for enterprises. This has two benefits: allowing the knowledge of enterprise ages be used to assign company sizes and giving the enterprises sizes a realistic time evolution. As first proposed by Gibrat in 1930 suppose that that the size of a company with age i is given by

$$X_i = \prod_{j=1}^{i} (1 + \epsilon_j) \tag{1}$$

where $\epsilon_i \sim \mathcal{N}(\alpha, \beta^2)$, with α and β depending on the local authority and SIC code. This means that if X represents the size of an enterprise picked at random, with a given SIC and location, then

$$X = \sum_i \mathbf{1}_{Zi} X_i \tag{2}$$

where $\mathbf{1}$ is an indicator function and Z is a random variable such that $\mathbb{P}(Z = i) = \frac{n_i}{N}$ where N is the number of enterprises with the given SIC and local authority and n_i is the number of companies with age i. If the distribution of X is known, the parameters α and β can be calculated by matching $\mathbb{E}(X)$ and $\mathrm{Var}(X)$[4].

[2] Regional size distributions are available but hopefully don't predict things well
[3] details in the Appendix
[4] See section 4.3 for details

2.4 Local Units From Enterprises

In order to measure useful quantities, like employment, it is important to have data on the local units. In the case of employment per local authority, there is a strong linear relationship between the number of local units and the number of enterprises in each local authority. The interpretation of this is that in areas with more registered companies (and therefore more enterprises) there is more economic activity which results in more local units being present, not that on average a company has α local units in the local authority that it is registered. Using this approach the number of local units can be calculated from companies data.

3 Results

3.1 Dataset

3.2 Total Employment

3.3 Employment Migration

4 Appendix

4.1 Size Bands and Log Normal parameter fitting

The ONS typically release enterprise and local unit size data in bands, so they give the number of companies, n_i with size $x \in [a_i, a_{i+1})$ by some other parameter (SIC code or location). 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

$$l(\mu, \sigma) = \sum_i n_i \log\left(\Phi\left(\frac{\log a_{i+1} - \mu}{\sigma}\right) - \Phi\left(\frac{\log a_i - \mu}{\sigma}\right)\right) \tag{3}$$

since $\mathbb{P}(a_i < X < a_{i+1}) = \mathbb{P}(\log a_i < \log X < \log a_{i+1}) = \Phi\left(\frac{\log a_{i+1} - \mu}{\sigma}\right) - \Phi\left(\frac{\log a_i - \mu}{\sigma}\right)$.
For the employment size distributions of enterprises and local units the the fitted parameters predict the ratios $\frac{n_i}{N}$ well and in the case of local authorities the distributions predict the total employment well so it is perhaps reasonable to assume that the actual distribution is log normal. Whether it is possible for other distributions (Yule or other fat tailed distributions) to reproduce these results is something worth testing, but for the purposes of predicting enterprise and local unit sizes, the log normal distribution seems to work well.

4.2 Log Normal Sum Approximation

Given an indexed set of random variables $\{X_i\}$ such that $\log X_i \sim \mathcal{N}(\mu_i, {\sigma_i}^2)$ an important question for this study is whether the random variable $X = \sum_i \mathbf{1}_{Zi} X_i$ has a distribution that is approximately log normal, and whether there exist non trivial requirements of the distribution (μ_i, σ_i) given the distribution of Z for X to be log normally distributed[5]. If this were the case then the fact that there is a conservation of the log normal distributions and that the distribution of Z is known, then a more accurate prediction of (μ_i, σ_i) could be calculated to get a better prediction of company size. Running simulations with the given data, to make sure that the parameter choice is relevant, results in X being approximately log normally distributed for both the local authority distributions and SIC distributions.

4.3 Gibrat process parameter fitting

Assuming equations (1) and (2)

$$\mathbb{E}(X) = \sum_i \mathbb{E}(\mathbf{1}_{Zi} X_i) = \sum_i \frac{n_i}{N} \mathbb{E}(X_i) \tag{4}$$

$$\mathbb{E}(X_i) = \prod_{j=1}^{i} (1 + \epsilon_i) = (1 + \alpha)^i \tag{5}$$

Assuming $\mathbb{E}(X)$ is known,

$$\mathbb{E}(X) = \sum_i \frac{n_i}{N} (1 + \alpha)^i \tag{6}$$

[5] A trivial example is that $\forall i,\ (\mu_i, \sigma_i) = (\mu, \sigma)$, so $X = kX_i$ and is therefore log normally distributed.

is a polynomial in $(1+\alpha)$ of quite a high order (depending on the time step) which can be solved numerically to find α. The variance is much the same, given that

$$\mathrm{Var}(X) = \sum_{i,j} \mathrm{Cov}(\mathbf{1}_{Zi} X_i, \mathbf{1}_{Zj} X_j) \tag{7}$$

$$= \sum_{i,j} \delta_{ij} \frac{n_i}{N} (\beta^2 + (1+\alpha)^2)^i - \frac{n_i n_j}{N^2} (1+\alpha)^{i+j} \tag{8}$$

which is also a polynomial, this time in $(\beta^2 + (1+\alpha)^2)$, that can be solved numerically knowing α. β can then be calculated provided that the underlying known ages do not themselves create too much variance in the final distribution, such that solving the above polynomial implies that $\beta^2 < 0$ which is generally the case if the assumption is made that α and β depend only on location or only on SIC code.

4.4 Model of Company Growth

Since the ONS data shows the Suppose that the number of employees X_i^s belonging to an enterprise registered in local authority l, with SIC code s and that is i months old has size:

$$X_i = \prod_{j=1}^{i} (1 + \alpha_j + \beta_j) \tag{9}$$

Where $\alpha_j \sim \mathcal{N}(\mu_l, \sigma_l^2)$ and $\beta_j \sim \mathcal{N}(\nu_s, \tau_s^2)$. So α is a property of the local authority and β is a property of the SIC code. Dropping the subscripts, we assume that μ, ν, σ, and τ are suitably small such that $\log(1+\alpha+\beta)$ is approximately normally distributed. This means that, X_i has a log normal distribution, and by the log normal sum approximation[6] this implies that the national size distribution of companies is log normally distributed, as well as the cross sections by SIC code and location, as seen in the data.

4.4.1 Parameter Fitting

Assuming this model, it is possible to fit the parameters by matching the mean and variance of the observed size distributions. Here we will focus on using the local authority distributions, but the process is the same for the SIC codes.
Let X be a random variable representing the size of a firm picked at random from a local authority l. After fitting a log normal distribution to the ONS data[7], the mean and variance of the observed size distributions are known. From (9):

$$X = \sum_{s \in S} \sum_{i \in A} \mathbf{1}_{Z(s,i)} X_i^s \tag{10}$$

where $\mathbf{1}$ is an indicator function, and Z is a random variable which represents drawing a random SIC code and age. This means that, if N is the total number of companies, and n_{si} is the number of companies, age i with SIC code s in the local authority, then:

$$\mathbb{E}(X) = \frac{1}{N} \sum_{s \in S} \sum_{i \in A} n_{si} \mathbb{E}(X_i^s) \tag{11}$$

$$= \frac{1}{N} \sum_{s \in S} \sum_{i \in A} n_{si} (1 + \mu_l + \nu_s)^i \tag{12}$$

and

$$\mathrm{Var}(X) = \sum_{s,t \in S} \sum_{i,j \in A} \mathrm{Cov}(\mathbf{1}_{Z(s,i)} X_i^s, \mathbf{1}_{Z(t,j)} X_j^t) \tag{13}$$

$$= \sum_{s,t \in S} \sum_{i,j \in A} \delta_{st} \delta_{ij} \frac{n_{si}}{N} \left(\sigma_l^2 + \tau_s^2 + (1 + \mu_l + \nu_s)^2\right)^i - \frac{n_{si} n_{tj}}{N^2} (1 + \mu_l + \nu_s)^i (1 + \mu_l + \nu_t)^j \tag{14}$$

. From (11) the model provides an equation for each local authority, and similarly for each SIC code, so it is possible to calculate μ_l and ν_s for all l and s and in the same way (13) provides equations for σ_l and τ_s. Luckily the calculation of the change in the total variance with changing the parameters σ and τ scales like N^2 not N^4 since the second term in the summand of 14 can be calculated once beforehand, so numerical methods can be used with the whole dataset.

[6] See appendix, section 4.2
[7] See appendix, section 4.1