Building a Granular Dataset of UK Companies

Alfred Holmes

August 2018

Abstract

Open data on UK companies is available from two sources: The Office for National Statistics (ONS) and Companies House (CH). It is possible to combine the two to estimate the properties of individual companies and in doing so compile a dataset of reasonably accurate detailed granular data on the age, turnover, location through time and employment size of individual companies through the years 2012 - present. This report describes the techniques used to build the dataset.

Motivation

Contemporary economic modeling requires detailed granular data, but typically the available data is aggregated either due to old methods being used or to avoid breaking privacy laws by disclosing too much personal data. This project began as a way to gain granular data on the number of jobs (employment opportunities) moving from each local authority to each other local authority by combining multiple data sources for use in an agent based model. The techniques presented here could be used to infer many properties of individual companies operating in the UK as well as producing all manner of summary statistics and visualizations.

Data and Code

The code used for this analysis, along with data, is available in a GitHub repository.

1 Raw Data

1.1 Companies House Snapshots

Companies House (CH) is the database of registered companies[1] in the UK. The database contains the registered address, standard industrial classification (SIC) code, age, number of mortgages, and sometimes turnover information depending on the size of the company. Every month CH releases a snapshot of all the active[2] registered companies. This is the easiest way to get company data of companies that are currently operating.

[1] Need to check definition of a company

[2] According to Companies house records - the company hasn't told companies house that it is inactive.

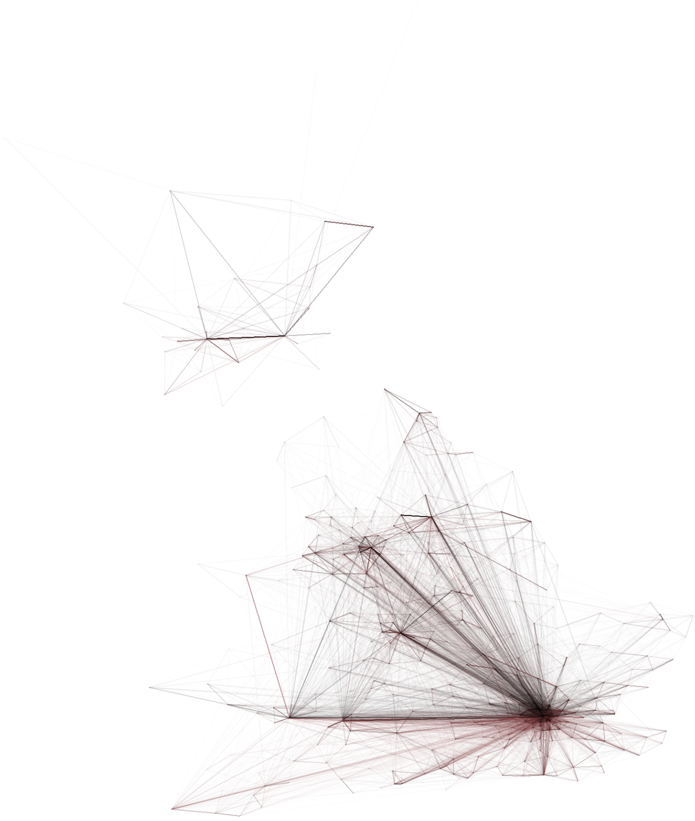

Figure 1: Visualization of all company migrations of active companies. Higher opacity represents more companies making the migration, red lines represent migration from south to north, black north to south.

1.2 Companies House API

CH does have an open API which can be used to get more detailed data by processing filing histories of individual companies. This can be used to see the evolution of a set of companies through time. The API has to be used on an individual company basis so to get data for a particular company it's Company ID is required. As the API doesn't have a call to dump a list of company numbers, to use the API the company numbers need to be acquired. Company numbers can be taken from the CH snapshots. If looking at aspects of companies through time it may be important to have the company numbers of companies that are no longer active but were active during the time of the sample. Archived versions of the CH snapshot can be used to get company numbers of inactive companies from 2012 and onwards. Scripts to get company migration data can be written, although inconsitent structures of the filing elements makes the processing of this data difficult. The API also has to be used to get accurate timings for the death of companies since in the snapshots only active companies are listed.

1.3 Office of National Statistics

The Office of National Statistics (ONS) releases fairly detailed business data on a yearly basis. The data broken down by SIC code and location - up to local authority level. The data contains information about enterprises, defined by ONS as ‘units with a certain level of autonomy’, and local units which are parts of an enterprise. For this analysis of the ONS and CH data, it is assumed that an enterprise is a collection of CH companies. Data before 2014 needs to be accessed through the archive of the old ONS website. ONS also releases data (position, area, name - id look up) on each of the UK’s local authorities as well as postcode look up tables to match postcodes to local authorities.

2 Combining CH Granular Data with ONS reports

Since the Companies House open data doesn’t contain some of the most important properties of a company (number of employees and turnover in particular) we assign companies properties in a way that both matches conditionally probabilistic growth models. The conditions being that the granular dataset reproduces the ONS aggregated statistics and that the companies die when the real CH companies die[3].

2.1 Assumptions

For the years 2012 - Present, there are about 10^6 more companies in the CH snapshot than are reported on ONS. This means that in order to combine the two datasets there needs to be a way of deciding which companies to include and which to discard.

2.2 Power Law with number reported by ONS and number of companies registered on CH

If one assumes that each enterprise reported by ONS is represented by a collection of CH companies then one way of interpreting Figure 2 is that the probability of a company being reported by ONS depends on the number of companies in the area. One possible explanation for this is that companies have a certain probability of forming partnership groups in such a way as for the partnership to be reported as one enterprise by ONS. In this way the number of partnerships could follow a similar power law relationship to the total number or registered companies depending on the mechanics.[4]

[3] Due to the lack of data, we have been unable to test the accuracy of this method of prediction

[4] It is quite straightforward to simulate non linear relationships in this way, but difficult to get a power law so there could be better ways to remove this

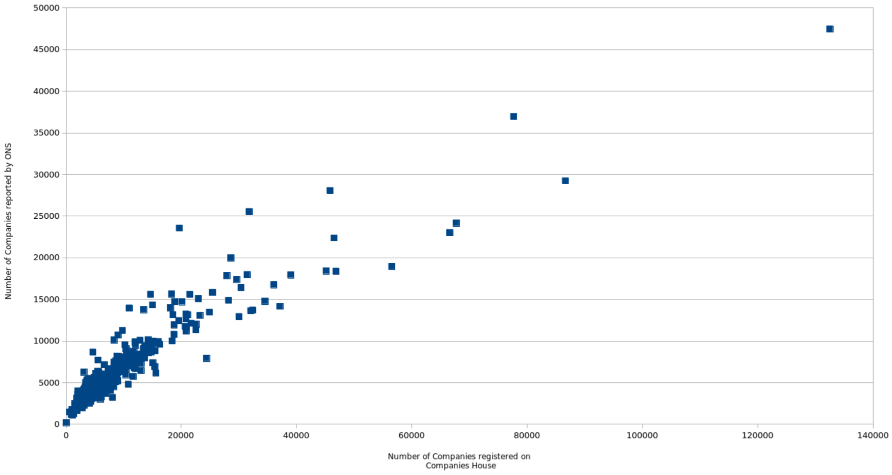

Figure 2: Total companies operating (ONS) in local authorities against active number registered on CH

3 Results

3.1 Employment Migration Data and Visualization

3.2 Generating other ONS Data from Dataset

3.2.1 Employment totals

4 Structure of Dataset

When dealing with highly granular data of this kind, a difficult question is how to store the data.