NBER WORKING PAPER SERIES

TRANSFORMING NATURALLY OCCURRING TEXT DATA INTO ECONOMIC STATISTICS: THE CASE OF ONLINE JOB VACANCY POSTINGS

Arthur Turrell
Bradley J. Speigner
Jyldyz Djumalieva
David Copple
James Thurgood

Working Paper 25837
http://www.nber.org/papers/w25837

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2019

The authors would like to thank Katherine Abraham, James Barker, David Bholat, David Bradnum, Emmet Cassidy, Matthew Corder, Rodrigo Guimaraes, Frances Hill, Tomas Key, Graham Logan, Michaela Morris, Michael Osbourne, Kate Reinold, Paul Robinson, Ayşegül Şahin, Ben Sole, Vincent Sterk, and the participants of the NBER Conference on Big Data in the 21st Century. We would especially like to thank William Abel for his help throughout the project. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research or the Bank of England. This work was funded by the Bank of England.

Transforming Naturally Occurring Text Data Into Economic Statistics: The Case of Online Job Vacancy Postings
Arthur Turrell, Bradley J. Speigner, Jyldyz Djumalieva, David Copple, and James Thurgood
NBER Working Paper No. 25837
May 2019
JEL No. C55,E24,J63

ABSTRACT

Using a dataset of 15 million UK job adverts from a recruitment website, we construct new economic statistics measuring labour market demand. These data are 'naturally occurring', having originally been posted online by firms. They offer information on two dimensions of vacancies—region and occupation—that firm-based surveys do not usually, and cannot easily, collect. These data do not come with official classification labels so we develop an algorithm which maps the free form text of job descriptions into standard occupational classification codes. The created vacancy statistics give a plausible, granular picture of UK labour demand and permit the analysis of Beveridge curves and mismatch unemployment at the occupational level.

Arthur Turrell
Advanced Analytics
Bank of England
Threadneedle St.
London EC2R 8AH
United Kingdom
a.e.turrell@gmail.com

Bradley J. Speigner
Bank of England
Threadneedle St.
London EC2R 8AH
United Kingdom
bradley.speigner@bankofengland.co.uk

Jyldyz Djumalieva
Nesta
58 Victoria Embankment
London, EC4Y 0DS
United Kingdom
jyldyz.djumalieva@nesta.org.uk

David Copple
Bank of England
Threadneedle St.
London EC2R 8AH
United Kingdom
david.copple@bankofengland.co.uk

James Thurgood
Bank of England
Threadneedle St.
London EC2R 8AH
United Kingdom
james.thurgood@bankofengland.co.uk

Code for assigning occupational labels to job ad text is available at
https://github.com/aeturrell/occupationcoder

1 Introduction

This paper presents an example of converting naturally occurring[1] data into economic statistics for use in research and analysis. The raw data consist of millions of individual job adverts as posted online by firms and recruitment agencies on the website Reed.co.uk in the UK. The objective is to process, clean, re-weight, and use this data as a measure of job vacancies by occupation and by region over time, and according to existing official statistical classifications. The methods developed for this purpose could be applied to other naturally occurring datasets. There have been no UK official statistics on vacancies by region and occupation since the JobCentre Plus data were discontinued and we show how these data can fill an important gap in our understanding of labour market demand.

One of the major benefits of using individual online job postings is that they are a direct measure of the economic activity associated with trying to hire workers. Another is the sheer volume they offer – of the order of 10^5 individual vacancies at any point in time for the UK. These large numbers allow for very granular analysis.

As well as demonstrating the creation of new economic statistics on vacancies, we make a major contribution in the method we use to transform the text of job adverts into time series data labelled by official classifications (here the UK Office for National Statistics', or ONS', Standard Occupational Classification, or SOC, codes). Our algorithm draws on methods from computer science and natural language processing, and makes use of both the job title and job description.[2] It could be adapted and applied to the SOC classifications of other countries or regions, or to other types of text and classifications. It could also be used by employers to check what occupation their job adverts fall under, to better target their advert or adjust compensation.

The newly created vacancy time series, split by occupation, are compared to existing data on UK job vacancies, namely the ONS' Vacancy Survey and JobCentre Plus data. We consider the likely biases of the Reed-derived vacancy time series. To demonstrate the utility of processing the text of these data we use them to estimate Beveridge curves by occupation and to calculate the rate of mismatch unemployment (by occupation) for the UK using the mismatch framework of Şahin et al. (2014).

The structure of the paper is as follows: §2 sets out relevant previous literature relevant to vacancy statistics, §3 describes the online job vacancies data in the context of other data on vacancies, §4 describes the algorithm we develop to assign vacancies to official statistical classifications, §5 describes the processed

[1] As opposed to data collected for the express purpose of constructing statistics, these data are a side-product of other economic activity.

[2] Computer code available at `http://github.com/aeturrell/occupationcoder`.

data, §6 explores some uses of this data in economic analysis, and §7 concludes.

2 Literature

Vacancy data have long been collected via surveys; Abraham (1983) reviews a number of regional surveys which began to do this from the 1960s to the 1980s before national survey data on vacancies began to be widely collected. In the UK and US, there are now designated national statistics measuring job vacancies using surveys: the ONS Vacancy Survey and the JOLTS (Job openings and Labor Turnover Survey) respectively.

The Vacancy Survey was introduced in 2001 and each month surveys around 6,000 firms on the total number of vacancies that they have open (Machin, 2003) – a measure of the stock of vacancies. The firm-level data collection allows for cross-sectional data by both firm size and industry. Data are collected on the Friday between the second and eighth of each month and are thereafter available at monthly frequency with a 40 day lag. No breakdown of vacancies by region or occupation is available. These dimensions are especially difficult for survey data to collect because firms may not be familiar with occupational codes and asking them to submit, instead of a single number, up to 368 numbers reflecting each of the 4-digit UK occupational codes would be a significant change in the administrative burden imposed by the survey. Similarly, regional data are difficult to collect via this method as it is more cost effective and potentially more accurate to contact only a firm's head office for vacancy numbers. Due to the sample being drawn from a business register, new firms are underrepresented, though this bias is only estimated to create errors of $\pm 20,000$ for vacancies levels in the hundreds of thousands. Although the scale and quality of vacancy data collection has changed substantially since the 1960s, the methodology has not. Collecting survey data is expensive, has a relatively long lag, and is ill-suited to providing occupational or regional information.

Administrative data are an alternative source of information on job vacancies that is acknowledged to be "cheap and relatively easy to produce" (Bentley, 2005). These are most often vacancies notified to government employment service offices. In the UK, the main source of these data are JobCentre Plus (JCP) vacancies. They were discontinued in 2012 and underwent significant changes in 2006 so that the longest recent usable continuous time series runs from July 2006 to November 2012. The JCP had aggregate coverage of around a third of UK vacancies prior to 2003 (Machin, 2003) but with large variation between regions, between sectors, and over time depending on the point in the business cycle and the policies of JCP offices. Burgess and Profit (2001) note that these vacancies have a disproportionate share of low-skilled, manual jobs and are more likely to be matched to the long-term unemployed, while Patterson et al. (2016), looking at more recent data than Machin (2003), find that they over-represent some sectors. Problems with JCP data included that a significant percentage of the entire vacancy stock was not always updated when filled or withdrawn by employers, biasing the stock upwards by numbers as high as the multiple tens of thousands out of vacancies in the few hundreds of thousands. These data have been used in several other studies; namely Coles and Smith (1996), Smith (2012), and Manning and Petrongolo (2017). These data were not included in the ONS' labour market statistics releases between 2005 and their discontinuation because of concerns over their appropriateness as a labour market indicator (Bentley, 2005). The number of ways for firms to communicate to JCP offices increased at that time, leading to structural breaks in the series, and the reliance on firms to notify JCP offices when vacancies were filled or withdrawn made the outflow series, and therefore the stock, vulnerable to bias. Indeed, the onus was on JCP offices to follow up with employers and, as this did not happen consistently or for every position, a large amount of what has been described as 'vacancy deadwood' built up.

We use job adverts which have been generated as a result of firms attempting to hire workers, but from a privately run website, Reed.co.uk, rather than from a government run employment office. This will have implications for the nature of the jobs advertised. The ads are run at a cost to the posting party so that concerns about an ever growing stock of vacancies which have, in reality, been filled or withdrawn do not apply. Other job advertisement website data have been used for the analysis of vacancy statistics, including Deming and Kahn (2017) with Burning Glass data, Marinescu (2017) using data from CareerBuilder.com, and Mamertino and Sinclair (2016) using data from Indeed.com. As explained by Cajner, Ratner et al. (2016), there have been significant discrepancies between the stock of vacancies implied by two US series, the JOLTS and the Conference Board Help Wanted Online, which may be caused by changes in the price charged to employers to post online job vacancies.

Previous work has found that online job vacancy postings can give a good indication of the trends in aggregate vacancies (Hershbein and Kahn, 2018). There has been a secular trend increase in the number of vacancies which are posted online, as evidenced by the replacement in the US of the Help Wanted Index of print advertisements with the Help Wanted Online Series. Although they may not offer full coverage, online vacancy statistics can powerfully complement official statistics on vacancies, which tend to be based on surveys of firms.

Our paper adds to a growing literature on the analysis of text in job vacancies. Marinescu and Wolthoff (2016) use job titles to explain more of the wage variance in US job vacancies in 2011 than SOC codes alone do. Deming and Kahn (2017) use job vacancy descriptions that have been processed into keywords to define general skills that have explanatory power for both pay and firm performance beyond the usual labour market classifications. Azar et al. (2018) leverage online job vacancies, with job title text cleaned and standardised, to estimate the labour market concentration according to the Herfindahl-Hirchsman index. And Hershbein and Kahn (2018) ask whether the within-occupation skills demanded in job vacancy text are accelerated during recessions.

We show how online job advert text can be used to generate occupational labels. Until recently, methods which existed to label vacancy text with official classifications were proprietary, limited in the number of searches, or did not make use of the job description field. While writing up our results we became aware of similar approaches being developed for the US (Atalay et al., 2017), Germany (Gweon et al., 2017) and for the International Labour Organisation occupational classification (Boselli et al., 2017*a*,*b*).

For demonstrating the usefulness of the data, we use the search and matching theory of the labour market (Mortensen and Pissarides, 1994) in which job vacancies represent the demand for labour. Labour market tightness, $\theta = \frac{V}{U}$, where V is the stock of job vacancies and U is the unemployment level, is an important parameter in this framework. At the centre of theories of mismatch is the matching function $h(U, V)$ which matches vacancies and unemployed workers to give the number of new jobs per unit time as described in Petrongolo and Pissarides (2001). In the applications part of the paper, we use econometric estimates of the Reed data which are published in full in Turrell et al. (2018).

3 Data

Our raw data are approximately $15,242,000$ individual jobs posted at daily frequency from January 2008 to December 2016 on Reed.co.uk, a job advertisement website. The site facilitates matching between firms and jobseekers. Firms who wish to hire workers, or recruitment agencies acting on their behalf, pay Reed to take out adverts on the site. As of February 2019, the cost of a single job advert to be posted anytime in the next 12 months and to remain live for 6 weeks is £150 + tax.[3] Reed has a direct business relationship with the firm or recruitment agency who post the advert.

The fields in the raw data which are typically available include a job posted date, an offered nominal wage, a sectoral classification (similar to the ONS sectoral section classification), the latitude and longitude of the job, a job title, and a job description. Our data are unusual compared to the recent literature in that they come from a job advertisement and employee recruitment firm (a recruiter) rather than from an aggregator or from a survey. There are two different kinds of website which post job advertisements.

[3] Unfortunately we do not have a time series of advert posting costs.

Aggregators use so-called 'spiders' to crawl the internet looking at webpages, such as firm recruitment sites, which host job vacancies and then record those job vacancies.[4] In contrast, firms post vacancies directly with recruiters. Recruiters may have access to private information about the job vacancy which an aggregator would not. In our case, an example of such information is the offered salary field. Additionally, the likelihood of duplicates is lower in a recruitment firm dataset because jobs are only added to the site as the result of direct contact with a firm. Aggregators are more likely to pick up the same job multiple times from different ad sites though they expend considerable effort in removing duplicate listings.

A feature of all data collected online is that it tends to contain superfluous information, at least relative to survey data, and, similarly to survey data, may have entries that are incomplete or erroneous. However, perhaps because of the cost of posting, there are very few incomplete entries in the Reed data. The most frequently encountered erroneous information is in the form of offered wages (not always shown to jobseekers) which appear too low (as they are not compliant with the minimum wage law) or unrealistically high. We do not use the wage data for the creation of occupational labels.

The sectoral field of each vacancy has strong similarities to ONS Standard Industrial Classification (SIC) sections and we constructed a manual mapping from the Reed sectors to the SIC sections. The data contain fields for latitude and longitude which are used to map each vacancy into regions given by Nomenclature of Territorial Units for Statistics (NUTS) codes. As the data are for the UK, the NUTS characters are counted only after the 'UK' designation. An example 3-character NUTS code would be 'UKF13', where the 'F1' designates Derbyshire and Nottinghamshire (UK counties), and 'F13' South and West Derbyshire.

We also use a number of other datasets from the ONS, including the *Labour Force Survey* (LFS) (Office for National Statistics, 2017), the aforementioned *Vacancy Survey*, and sectoral productivity measures.

3.1 The stock of vacancies and its potential bias

We consider how to estimate a stock of vacancies from the Reed job adverts and what biases might affect this estimate. We want to turn the Reed job adverts into a measure of job vacancies which are as close to the US JOLTS (Job Openings and Labor Turnover Survey) definition of vacancies as possible. JOLTS defines job vacancies as all positions that are open (not filled) on the last business day of the month. A job is vacant only if it meets all of the following conditions:

1. A specific position exists and there is work available for that position. The position can be full-time or part-time, and it can be permanent, short-term, or seasonal, and;

[4] Examples of research using datasets from aggregators include Deming and Kahn (2017) (Burning Glass), Marinescu (2017) (CareerBuilder.com), and Mamertino and Sinclair (2016) (Indeed.com).

2. The job could start within 30 days, whether or not the establishment finds a suitable candidate during that time, and;

3. There is active recruiting for workers from outside the establishment location that has the opening.

The ONS Vacancy Survey uses a similar definition but without the stipulation that the job could start within 30 days (Machin, 2003). Both definitions are of job vacancies as a stock – that is all jobs which are open at a particular time, rather than newly opened within a particular time window.

The Reed job adverts constitute a flow of new vacancies, arriving in daily time periods. In order to satisfy the JOLTS definition, we need to transform this flow of vacancies into a stock and ensure that all three conditions are met. We can be fairly certain that the first JOLTS condition is satisfied. As posting a vacancy incurs a cost, it seems unlikely that firms or recruitment agencies would post vacancies for which there is not an available position, at least on any large scale.

We cannot be sure about Reed adverts satisfying the second JOLTS condition, but it seems reasonable to assume that, once filled, most positions could start within 30 days because the adverts do not have a start date field. This suggests an implicit start date of as soon as the position is filled. Typically, for job-to-job flows, the limiting factor in a new firm-worker match is the workers' notice period.

The third JOLTS condition is satisfied by the posting of the vacancy on a third party website. It seems very likely that most job adverts posted on Reed will satisfy these three conditions.

Now we must consider how to transform the job adverts, which are a flow in units of adverts per day, into a stock of vacancies. As entries are removed from the site after being live for 6 weeks, the stock is simply the number of vacancies which were posted in the last 6 weeks or fewer. More explicitly, in discrete time, let the flow of adverts be $\dot{V}_d$ with d referring to a day. To retrieve stocks, the data are transformed as follows (where the time index refers to monthly frequency):

$$V_m = V_{m-1} + \sum_{d \in m} \left(\dot{V}_d - \dot{V}_{d-6\times 7} \right) \tag{1}$$

Note that this implicitly assumes that job adverts are filled or withdrawn by the employer after 6 weeks. There is no information on whether or not positions are filled within the Reed job advert data. This is typical of online vacancy data which are not matched with data on recruitment and most survey data: we cannot properly distinguish between advert outflows (that is, job adverts which are removed from the site) that are due to employers who have decided to stop trying to recruit and those that are due to a position being filled. In the Reed case, when an advert is not reposted after 6 weeks, it could be for either of these two reasons. This is an outflow-type identification problem. However, because we will later work with data at the occupational level which is matched to survey data on hires also at the occupational level, we will be able to distinguish between the two cases.

Similarly, if an advert is reposted it could be because either the position was not filled, or the firm has decided to hire further employees. However, in this case and with all else equal, we would see whether the number of vacancies had increased or not. As with the outflow identification, it will not matter at the occupational level for which we have data on hires from surveys.

At the occupational level, then, we need not be concerned that econometric estimates of the effect of vacancies on hires estimated on Reed data will be biased by the inability to distinguish between types of outflow or inflow. However, the JOLTS definition requires jobs to be unfilled to be a vacancy, as does the definition used in many other analyses of vacancies (Abraham, 1983) which describe them as being current, unfilled job openings which are immediately available for occupancy by workers outside a firm and for which a firm is actively seeking such workers (for full-time, part-time, permanent, temporary, seasonal and short-term work). Therefore our assumption, enforced by the data, that the stock of vacancies is built up from equation (1) could lead to some biases in this measure of the stock.

Let us consider these stock-flow biases. The first is that posted job adverts are filled before the 6 weeks are up, which would bias the vacancy stock derived from the Reed data upwards. This is an aggregate outflow bias. The extent of this bias overall depends on the average duration of a vacancy, which is known to vary across the business cycle (Abraham, 1983; Abraham and Wachter, 1987). The discontinued DHI-DFH Mean Vacancy Duration Measure for the USA falls markedly during recessions, to 2-3 weeks, and increased to over 4 weeks in mid-2018 (FRED, 2019). No official statistics are currently available on the mean duration of vacancies in the UK. If we were to assume that vacancies were to endure for the 2-4 weeks implied by the US data, it would mean that our aggregate vacancy stock is biased upwards. However, as we will shortly adjust the mean level of vacancies in the Reed data to match the ONS' measure of overall vacancies, this aggregate upward bias will be corrected.

Vacancy durations also vary by occupation (Abraham, 1983; Abraham and Wachter, 1987), and this poses more of a problem because it means that the stock of vacancies will be differentially biased by occupation. This is a differential outflow bias. Those occupations with short vacancy durations will have vacancy stocks which are biased upwards. We will shortly re-weight the Reed data using the fact that sectoral counts are available in both the ONS' measures of vacancies and the Reed data. By doing so, we will eliminate bias which exists across sectors. This will reduce some of the biases by occupation but, unfortunately, these biases cannot be eliminated entirely because there is no one-for-one relationship between occupation and sector. This is likely to be a problem for aggregator job advertisement sites too; if their data ultimately come from sites like Reed, who have a fixed period when a job is live, they similarly do not know if and when the vacancy was filled within that period. We also cannot exclude the possibility that some firms' hiring strategies are adapted to the method by which they post the vacancy. If they have paid for an advert with a duration of 6 weeks, they may decide to only review applications to select a preferred match once that time has expired. This strategy is typical of graduate schemes, for example.

Unfortunately, the differential outflow bias by occupation could also create bias in estimates of matching efficiency. Upward biases in the stocks of some occupations will bias the matching efficiency of those occupations downwards. We consider which occupations may be affected by this: the DHI-DFH Mean Vacancy Duration Measure (FRED, 2019) for the US offers a sectoral split which shows that more highly skilled vacancies, for example in financial services and business and professional services have longer vacancy durations on average than leisure and hospitality and construction. This makes intuitive sense in the context of specialisation. So an important caveat of our results is that heterogeneous vacancy durations are likely to bias the matching efficiency of low skill occupations downwards. The reweighting we apply in the next section will reduce, but not eliminate, this bias.

3.2 Coverage and representativeness biases

We now examine bias with respect to coverage and representativeness for the Reed vacancies, as well as describing the steps we take to reduce these biases.

These two types of bias exist at the aggregate level. Vacancies posted online are unlikely to have coverage of 100% of vacancies advertised in the economy, and the Reed stock of vacancies, obtained from equation (1), has aggregate coverage of around 40% relative to the Vacancy Survey. In addition, the composition of the vacancies that are posted online is likely to be quite different from reality. These problems of bias and coverage exist for all job vacancy data based on job adverts, including the widely used JobCentre Plus data, and have long existed in the empirical literature on job vacancies. Prior to the advent of national vacancy statistics, most previous empirical work is based on the use of vacancies advertised at job centres, which have the same problems though for different reasons.

Additionally, vacancies as posted online do not have some of the problems that data collected by surveys have. Surveys are likely to have non- or incomplete-response bias, overestimation of the vacancies posted by large firms, underestimation of vacancies from recently created firms, and, when comparing vacancies and unemployment, could be biased by frequency mismatch between surveys Abraham (1983). Non-response bias is not relevant for job adverts posted online, differentials due to firm size may exist but are more likely to be caused by the ability to advertise positions (rather than size itself), and as postings are typically at daily frequency there can be no large role for frequency mismatch. The cost of posting adverts online with a recruiter means the problem of 'discouraged vacancies' is likely to be small.

There are many factors which affect the coverage of online job vacancies. Technological diffusion is one; given that no vacancies were posted on the World Wide Web before 1990, and that newspaper circulations have fallen substantially since the 1980s, there has been a drift of job vacancies from adverts in newspapers to adverts placed online. Over time, the coverage of online vacancies has improved. Barnichon (2010) shows that this drift in coverage closely follows the S-shape typical of technological diffusion for the US, and that it also closely follows the similarly S-shaped fraction of internet users in that country. At the start of the period we study, 78% of the UK population were internet users, suggesting that the equivalent transition in the UK was already well under way by 2008.[5] Another reason why there are coverage differences for online adverts posted with a recruiter versus surveys is the cost of posting vacancies online. (Cajner, Ratner et al., 2016) find that changes in the cost of posting vacancies online had a significant influence on the aggregate stock of vacancies as represented by online sources versus other sources. The (time-dependent) reweighting we will use will correct for both of these biases.[6]

The extent to which the composition of job adverts posted online is biased relative to the composition of all vacancies in the economy is a more difficult issue to resolve. As there is a non-trivial cost to posting a job advert online, at least with a recruiter, those which are will need to have an expected return for the firm greater than that cost. Additionally, some job vacancies may get a better response if posted via other media, e.g. newspaper or shop window. There may be other pressures which determine whether vacancies appear online or not, for instance the quality of alternative channels for matching between jobseekers and firms.

Because of being online, having a posting cost, and other factors, it is likely that Reed job adverts are biased to over-represent middle and higher skilled vacancies. This is a differential representativeness bias. The bias may not matter much for the uses demonstrated here as long as it is reasonably fixed over time. Bias which is changing over time is the most detrimental to any analysis because (cross-section) fixed effects cannot absorb the bias effect. A fixed bias would imply that the stock of vacancies expressed as a

[5] World Bank series: Individuals using the Internet (% of population) International Telecommunication Union, World Telecommunication/ICT Development Report and database.

[6] The cost of posting vacancies with Reed is not differentiated by sector or occupation.

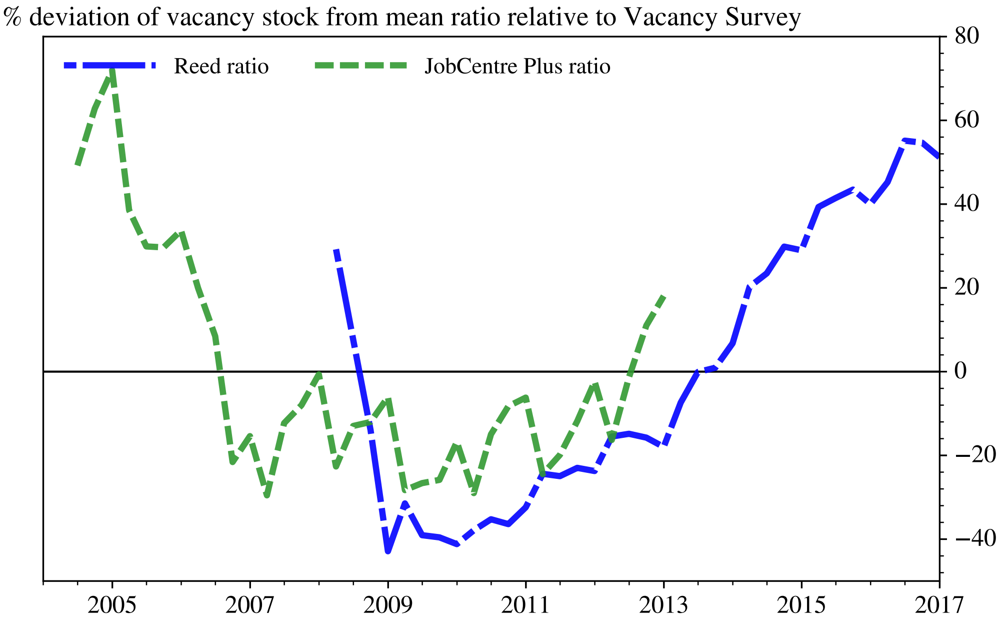

Figure 1: The percentage deviations of both the JobCentre Plus and Reed stocks of vacancies from their mean ratio relative to the Vacancy Survey stock of vacancies. Source: Reed, ONS, National Online Manpower Information System (NOMIS).

ratio relative to the Vacancy Survey stock was also fixed over time. In Figure 1 we show the percentage deviations of both the JobCentre Plus and Reed stocks of vacancies from their mean ratio relative to the Vacancy Survey stock of vacancies. The figure shows that neither are fixed over time and both likely suffer from a changing level of bias. On the basis of the simple measure shown in Figure 1, bias does not seem to be more of a problem for the Reed data than for the widely used JobCentre Plus vacancy data but it nonetheless does exist.

We can examine how much this bias is a problem at a more disaggregated level by taking advantage of the appearance of sectoral fields in both the Vacancy Survey and the Reed data. The mean annual ratios of the Reed to the Vacancy Survey stock of vacancies by sector are shown in Figure 2. The annual coverage ratios of the sectoral vacancy counts of the Reed data relative to the Vacancy Survey data are closer to unity for some sectors than for others, e.g. professional, scientific, and technical activities have a higher average coverage ratio than human health and social work activities. Such biases inevitably affect the stock of vacancies in the (unweighted) Reed data. For professional and scientific activities, information and communication, and administration, the Reed data are of comparable magnitude to the ONS estimates of vacancies. This could be because those sectors are well represented by the Reed data, but there could also be measurement differences which mean that the composition is different. Around 64% of vacancies

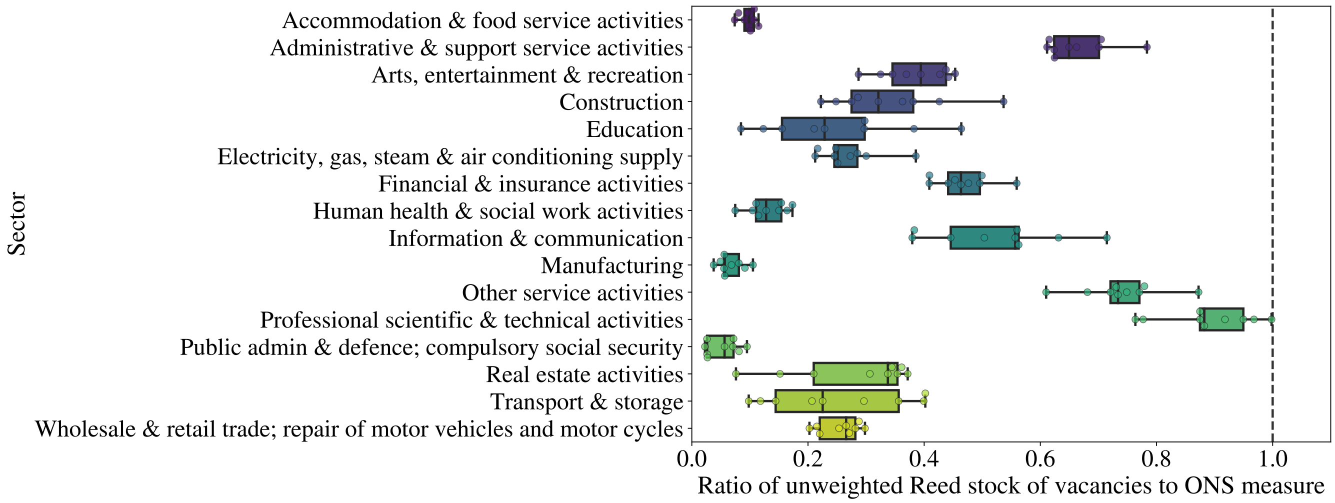

Figure 2: Mean annual ratios of the Reed to Vacancy Survey stock of vacancies by sector give an indication of where the Reed data have higher coverage (first moment close to unity) and where the bias remains relatively static over time (small second moment). Source: Reed, ONS.

have an annual ratio relative to the ONS survey with a median of greater than 20%. All are below unity, as would be expected if they were representative of the ONS equivalent sectoral counts.[7] The largest differences in magnitude between vacancies by sector in the Reed data and the ONS data are for public administration and manufacturing. Together, these account for around 9% of vacancies in the last quarter of 2016 according to the Vacancy Survey.

As noted, fixed biases can be absorbed by cross-section fixed effects. This does mean that there is potential for matching efficiencies calculated from these data to be biased. The Reed stock of vacancies is likely to be biased downwards for lower skill occupations, making the matching efficiencies of these occupations biased upwards. This contrasts with the differential outflow bias noted earlier, which biases the same occupations' matching efficiencies downwards. We do not know which dominates.

Some of these representativeness biases may be overcome or mitigated by reweighting the Reed stocks of vacancies by occupation. We do this by using the monthly sectoral (Standard Industrial Classification) disaggregation of the Vacancy Survey and the fact that the Reed monthly stock of vacancies also has a sectoral breakdown. Their ratios are used as weights. Reweighting can almost completely eliminate any aggregate vacancy stock bias. It will reduce the online representativeness bias and the differential occupational representativeness bias only to the extent that sectoral differences are correlated with these other compositional differences. Both online and occupational representativeness are likely to be strongly corre-

[7] If the Vacancy Survey is taken to be a true benchmark, values above unity would mean that there was duplication or misclassification in the Reed data.

	JobCentre Plus	Vacancy Survey	Reed	Reed (weighted)
JobCentre Plus	1	0.71	0.68	0.69
Vacancy Survey	-	1	0.93	0.98
Reed	-	-	1	0.90
Reed (weighted)	-	-	-	1

Table 1: Correlation matrix of aggregate vacancy data. Source: Reed, ONS, National Online Manpower Information System (NOMIS).

lated with skill level, and skill level and sector are also strongly correlated. So we expect that reweighting by sector has a substantial effect on these two biases and the differential outflow bias of §3.1 but cannot be sure of the quantitative extent of it. These biases, and others discussed in §2, exist in the widely used JobCentre Plus data too.

In the reweighting, the stock weight of an individual vacancy v in sector i and month m is given by

$$\omega_{i,m} = V_{i,m}^{\text{vs}}/V_{i,m}$$

with $V_{i,m}^{\text{vs}}$ the monthly stock of vacancies by sector according to the Vacancy Survey, and $V_{i,m}$ the stock of vacancies from the Reed data. Note that the correlation of the re-weighted Reed data with the aggregate Vacancy Survey is just smaller than unity. This is because of small differences between the ONS' sectoral vacancy stocks and the ONS' aggregate measure of vacancies due to rounding and seasonal adjustments. In subsequent sections, we use the weighted Reed data.

The aggregate time series of the Vacancy Survey, raw Reed stock of vacancies, and JobCentre Plus vacancies are shown in Figure 3. Neither of the latter have the same overall level of vacancies as the official statistics. The weighted Reed data, with lower bias, has increased variance relative to the unweighted series but provides a good fit to the Vacancy Survey data. The correlations between the series, shown in Table 1, show that the aggregate, unweighted Reed vacancy time series is better correlated with the Vacancy Survey measure than the JobCentre Plus data.

4 Matching job vacancy text to occupational classifications

We wish to apply occupational labels to the job vacancies by making use of the text of the job title, job description, and job sector. Using the text of the vacancies can be a powerful method to capture the stocks of different kinds of vacancies, as can be demonstrated with a simple count, by year, of the roots of the words 'data scientist', 'analyst', 'chef', 'nurse', and 'teacher'. Figure 4 shows the results of this

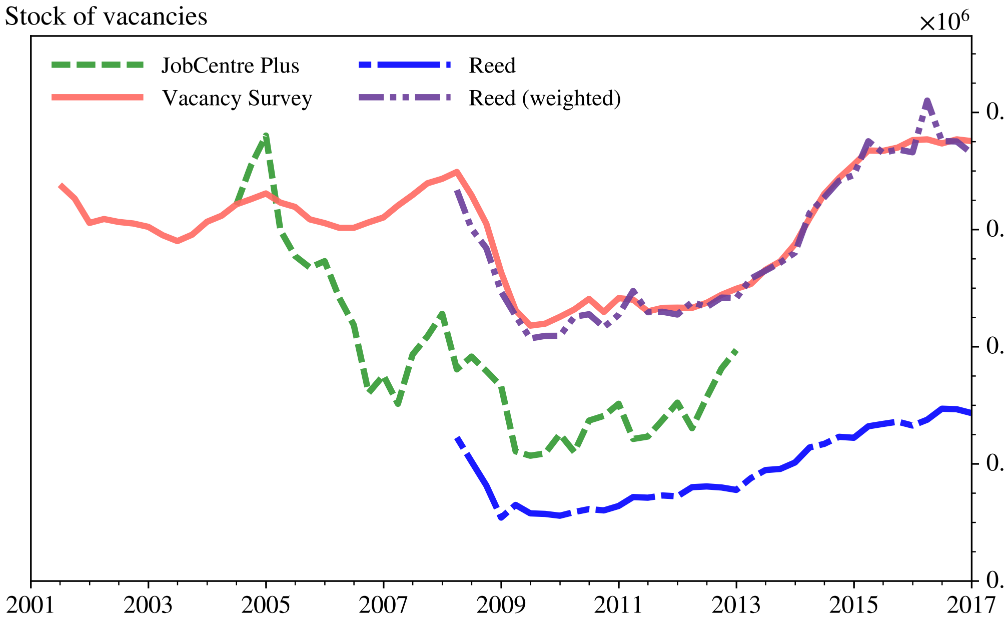

Figure 3: The aggregate stock of vacancies from three data sources. Source: Reed, ONS, National Online Manpower Information System (NOMIS).

count and documents the rise of data science as a distinct job from 2011-2016. From this figure, we cannot know whether data scientist is rising due to new demand, extra terms in existing occupations, or because of substitution away from other roles, e.g. statistician. However, it would be prohibitively labourious to create a list of all possible job titles and search them individually. Ideally we would want to count according to a well-defined and comprehensive classification which would put jobs into buckets according to a taxonomy. As long as the level of granularity is not too fine, this would put jobs like data scientist into buckets with jobs that require very similar skills and produce meaningful counts at the level of occupations. We develop and use an automated method for applying standardised occupational labels to job text. In order to use the Reed data most effectively for economic statistics, we label it with these standard classifications because they also exist in other official data, for example on unemployment.

4.1 Matching algorithm

In this section we describe the steps required to match job adverts in the Reed data to official Standard Occupational Classification (SOC) codes. We use the job title, job sector and job description text as the inputs into an algorithm which outputs a 3-digit SOC code. We choose the 3-digit level rather than more granular levels as there is a trade-off between more granularity and more accuracy in classifying jobs according to the correct SOC codes. As the SOC system is hierarchical and nested, with four levels as

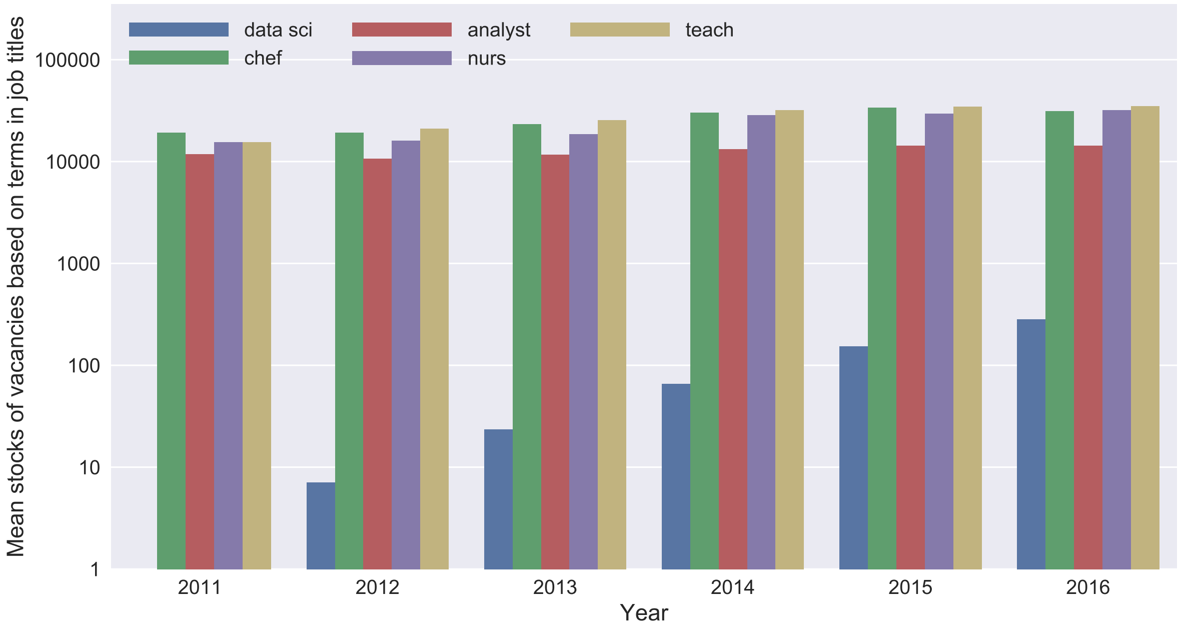

Figure 4: Counts of terms in job vacancy text designed to capture the job titles of ‘data scientist’, ’analyst’, ‘chef’, ‘nurse’, and ‘teacher’. Note the logarithmic scale. Source: Reed, LFS.

shown in Figure 5, generating SOC codes at the 3-digit level also delivers vacancies labelled by 1- and 2-digit codes.

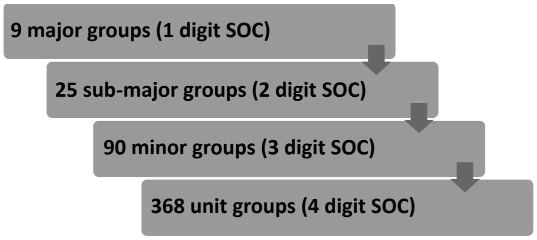
9 major groups (1 digit SOC)
25 sub-major groups (2 digit SOC)
90 minor groups (3 digit SOC)
368 unit groups (4 digit SOC)

Figure 5: Schematic of SOC code hierarchy.

In order to perform matches, we need reference information about all 3-digit SOC codes. We compile all publicly available text data, consisting of all known possible job titles and a short official description, for each SOC code and create a single text string from it. We use term frequency-inverse document frequency (tf-idf) vectors to represent these SOC code strings with a matrix with dimension $T \times D$ where t is a term from the text associated with a SOC code.[8] Our terms are comprised of all 1–3-grams[9] of salient words, that is words which are likely to have a useful meaning in a job vacancy context (we will define this formally below). For example, the phrase “must know how to cook Italian recipes” might reduce to a salient-words only phrase “cook Italian recipes”. This has 2-grams “cook Italian” and “Italian recipes” as well as 1-grams

[8] We use the SCIKIT-LEARN Python package to do this.

[9] An n-gram is all combinations of words with n words, so all 1–3-grams consist of all combinations of words with a length less than or equal to three words.

"cook", "Italian", and "recipes". The term frequency vector of this phrase would have entries equal to zero apart from for the columns representing these five terms.

Rather than term frequency, which is defined as the pure count of the number of times a term appears, we use tf-idf to represent SOC codes. The 'idf' part of tf-idf down weights words which are common across the corpus and so less useful in distinguishing one vector from another. As an example, the word 'work' may be salient as part of some n-grams but, as a single word, could also be very common in job adverts. Let d be a document (in this case a text string corresponding to a 3-digit SOC code) with D documents (the number of unique 3-digit SOC codes) in total. The specific form of tf-idf we use is then given by

$$\text{tf-idf}(t, d) = \text{tf}(t) \times \text{idf}(t, d) = \text{tf}(t) \times \left[\ln\left(\frac{1+D}{1+\text{df}(t,d)}\right) + 1\right]$$

where the document frequency, $\text{df}(t, d)$, is the number of documents in the corpus that contain term t and term-frequency, $\text{tf}(t)$, is the frequency of t. Each document can be represented as a vector of tf-idf scores, $\vec{v}_d$. These are normalised via the Euclidean norm so that $\hat{\vec{v}}_d = \frac{\vec{v_d}}{||v_d||}$.

The algorithm has four main stages; cleaning of vacancy text, exact matching on titles, identification of similar SOC codes, and fuzzy matching. The full flow of the algorithm is shown in Figure 13 of Appendix A. In more detail, the steps to match each vacancy in the dataset are:

- clean and combine the text of the job vacancy title, sector, and description, expanding any recognised acronyms in the process
- check whether the job title matches any known 3 digit SOC code titles (if so, use this as an exact match)
- express the given vacancy as a vector using tf-idf
- identify the five 3-digit SOC code vectors 'closest' to the vacancy vector by cosine similarity
- choose amongst these five from the best fuzzy match between the vacancy job title and all known 3-digit SOC code job titles

The cleaning process for text converts plural forms to singular forms (with some exceptions), expands abbreviations, removes stopwords[10] and non-salient words, and removes digits, punctuation, and extra spaces.

[10] Words which are not informative, typically conjunctions such as 'and'.

Real job vacancies are represented in the vector space by calculating their tf-idf score in the space of terms from the original corpus of SOC code descriptions and titles. A job vacancy is expressed as $\hat{\vec{v}}'$. In our algorithm, an arbitrary 3-digit SOC code is represented by $\hat{\vec{v}}_d$. To calculate which SOC codes are closest to $\hat{\vec{v}}'$, we solve

$$\arg\max_d \left\{ \hat{\vec{v}}' \cdot \hat{\vec{v}}_d \right\}$$

for the top five documents. This process allows us to estimate how 'close' a given posted job vacancy is to the 'golden image' jobs defined by each 3-digit SOC code string. Of the top five matches found in this way, the known title with the closest fuzzy match is chosen. This is implemented via the Python package fuzzywuzzy, which is based on Levenshtein distance (Levenshtein, 1966). We experimented with just taking the closest SOC code match by cosine similarity but using the Levenshtein distance to select among the five closest SOC code vectors provided better performance. We did not experiment with alternatives to Levenshtein distance.

In order to implement the algorithm, it was necessary to create three look-up dictionaries. The *known titles dictionary* represents known job titles and their associated SOC codes as tf-idf vectors and is also used to identify any exact job title matches, and for fuzzy matching. The text which is used to create the vectors for each 3-digit SOC code combines all of the known possible job titles for that SOC code in addition to a short official job description of it. The job titles are drawn from a set of around 10^4 possible titles covering all SOC codes. Publicly available ONS resources were used to generate this dictionary; the ONS Standard Occupational Classification, an extract from which may be seen in Table 2, and the Standard Occupational Classification 2010 Index, an extract from which is shown in Table 3. As shown in Figure 5, the ONS standard occupational classification system is a hierarchical structure with four levels. The ONS Standard Occupational Classification includes descriptions of each job. The Standard Occupational Classification Index 2010 extends the ONS occupational classification to capture around 30,000 alternative job titles across all unit groups. The *known titles dictionary* combines descriptions and all known titles from both sources to act as a reference list to match 'raw' job vacancy titles against. Example entries are given in Table 4.

We compiled an *acronym expansion dictionary* for processing the raw job title and job sector. It takes common within-occupation acronyms and expands them for clarity and to improve the quality of matches to the *known titles dictionary*. An example is the expansion of 'rgn' to 'registered general nurse'. The abbreviations were drawn from those commonly found in the job vacancies. The dictionary consists of a list of 219 abbreviations. Replacements of acronyms with their expansions increase the likelihood of an

Major Group	Sub-Major Group	Minor Group	Unit Group	Group Title
3				Associate professional and technical occupations
	31			Science, engineering and technology associate professionals
		311		Science, Engineering and Production Technicians
			3111	Laboratory technicians
			3112	Electrical and electronics technicians
			3113	Engineering technicians
			3114	Building and civil engineering technicians
			3115	Quality assurance technicians
			3116	Planning, process and production technicians
			3119	Science, engineering and production technicians n.e.c.
		312		Draughtspersons and Related Architectural Technicians
			3121	Architectural and town planning technicians
			3122	Draughtspersons

Table 2: An extract from the ONS occupational classification structure which forms the basis of our *known titles dictionary.*

exact match or a strong fuzzy match. The abbreviations were initially collected from a sample of 100,000 postings, where the set of words used in that sample was compared to the set of words in the official classification reference corpus. The abbreviations were detected by checking for words which existed in the raw job postings but were not present in the set of the official classification words. Those that occurred at least 5 times were investigated by searching for likely elaborations based upon the raw job titles and descriptions. Table 5 shows an extract from the *acronym expansion dictionary.*

We also created a *known words dictionary* that contains all words present in the ONS reference corpus (official and alternative ONS job titles and job descriptions). It is used to remove extraneous information from the titles of job vacancies; any term that is not in the dictionary is treated as a custom stopword and removed from the job vacancy titles before matching. This defines what we mean by salient terms. If a term does not exist in our ONS reference corpus, then we cannot use it for exact or fuzzy job title matching. This means that the term does not help in matching and may hinder it by preventing the detection of an exact title match or strong fuzzy title match. This dictionary is generated from the known titles dictionary but excludes official minor and unit group descriptions. These descriptions were excluded since they tend to contain more general words that might be irrelevant to a job. While descriptions are used when calculating cosine similarities, for exact and fuzzy job title matching, it was decided to use a stricter list of stopwords in order to increase the quality of the matches. Several additional words are deleted from the dictionary (and therefore from the job vacancy titles during matching). These words are ‘mini’, ‘x’, ‘london’, ‘nh’, ‘for’, ‘in’, ‘at’, ‘apprentice’, ‘graduate’, ‘senior’, ‘junior’, and ‘trainee’. There were two reasons for this. First, words which only qualify the level of seniority, but do not change the occupation, may inhibit matching; so we wished to have ‘senior financial analyst’ classified in the same way as ‘financial analyst’. Second, there are

SOC 2010	INDEXOCC	IND	ADD
1221	Manager, centre, holiday		
1225	Manager, centre, leisure		
1139	Manager, centre, mail	(postal distribution services)	
1181	Manager, centre, postgraduate	(health authority: hospital service)	
1251	Manager, centre, shopping		
1259	Manager, centre, skills		
1225	Manager, centre, sports		
1251	Manager, centre, town		
1259	Manager, centre, training		
1133	Manager, chain, supply		
2424	Manager, change, business		
2134	Manager, change, IT		
2134	Manager, change		(computing)
2134	Manager, change	(telecommunications)	
2424	Manager, change		
3545	Manager, channel		
1139	Manager, charity		
7130	Manager, check-out		
1225	Manager, cinema		
1225	Manager, circuit		(entertainment)
1190	Manager, circulation		
1225	Manager, circus		
3538	Manager, claims		
6240	Manager, cleaning		
1255	Manager, cleansing		
3545	Manager, client		(marketing)
3538	Manager, client	(bank)	
2462	Manager, client	(British Standards Institute)	
3538	Manager, client	(financial services)	

Table 3: An extract from Standard Occupational Classification Index 2010 which forms part of our *known titles dictionary*.

SOC code	Titles
214	conservation and environment professionals conservation professionals environment professionals conservation adviser countryside adviser environmental chemist marine conservationist coastal nature conservationist conservationist ecological consultant environmental consultant ecologist environmental engineer geoenvironmental engineer contaminated land engineer landfill engineer . . .
215	research and development managers research and development managers head research and development analytics manager creative manager research and development design manager process development manager manufacturing development manager research and development information manager research and development consumer insights manager insights manager laboratory manager passenger link manager government product manager . . .

Table 4: An extract from the *known titles dictionary*.

Term	Replace with
‘rgn’	registered general nurse
‘ifa’	independent financial adviser
‘nqt’	newly qualified teacher
‘flt’	fork lift truck
‘ce’	community employment
‘rmn’	registered mental nurse
‘eyfs’	early years foundation stage teacher

Table 5: An extract from the *acronym expansion dictionary*.

words which are not common stop words and also exist in the official ONS titles but which do occur very frequently in job titles and so are not particularly informative. These were identified via our exploratory analysis.

4.2 Evaluating the performance of the occupation coding algorithm

There is no perfect metric against which to score the quality of SOC code assignment by our algorithm. Official classifications can be applied inconsistently. Schierholz et al. (2016) surveys disagreements amongst those who code job titles into occupational classes, finding that the agreement overlap between coders is around 90% at the first-digit of the code (the highest level, for instance “Managers, Directors and Senior Officials”) but reduces to 70–80% at the 3-digit level that we work with for SOC codes (for instance, “Managers and proprietors in agriculture related services”). Automated approaches which use job title alone have even lower levels of agreement; Belloni et al. (2014) showed that algorithms which use job title alone agree on only 60% of records even at the top, 1-digit level of the International Standard Classification of Occupations. Other evidence of poor consistency of coding comes from Mellow and Sider (1983), who find an agreement level of only 57.6% percent for 3-digit occupational classifications, and Mathiowetz (1992). Additionally, not all job titles can be unambiguously assigned to an occupation. The algorithm which we contribute to match job vacancies (using both title and description) to SOC codes appears to reach at least the same level of agreement as do human coders.

To evaluate the quality of the labelling algorithm we developed, we asked the ONS to code a randomly chosen subset of our data using their proprietary algorithm. This algorithm is designed to process the responses to survey questions. The naturally occurring vacancy data contain job titles which often have superfluous information (for instance, "must be willing to travel") which can confuse a naive algorithm. Proprietary algorithms and algorithms used by national statistical offices are typically designed for survey data, in which job title entries tend to be more easy to parse and there is less extraneous information. Our algorithm must cope with a more challenging environment. We submitted 2×10^5 example vacancies to the ONS to run through their automated SOC code labelling process. Due to superfluous or missing information in the job title of the kind that would be unlikely to occur in survey data, their algorithm could only confidently produce a label for around 34% of these. Note that our algorithm similarly uses the job title to determine the SOC code to apply, but that it additionally uses the job description. Of the 34% which the ONS' approach could confidently give labels to, our method of coding based on job title and job description found the same label for 91% of the vacancies.

We also performed a smaller evaluation with manually assigned SOC codes. Volunteers, some associated with the project, were given parts of a list of 330 randomly chosen job titles from vacancies posted in 2016. Job titles were manually entered into the ONS online occupation coding tool, which returns a short list of the most relevant SOC codes, and volunteers then make a subjective selection of the most relevant SOC code. This is then compared with the output of our algorithm, with only a match at the 3-digit level being counted as accurate. The results from both are shown in Table 6. The results are similar to the levels of agreement seen between human coders. This algorithm is used in all applications of SOC codes to the Reed data.

In creating the algorithm, several areas of possible future improvement became clear. It always assigns a SOC code, but it could instead assign a probability or confidence level to each SOC code and so allow for a manual coder to judge marginal cases. Historical data on vacancies and on employment might also be used in marginal cases. We also found that occupations often come with both a level, e.g. manager, and a role, e.g. physicist. Better SOC assignment might result from explicitly extracting these two different types of information, and perhaps distinguishing between the higher and lower levels using offered salaries.

In interpreting the results based upon our SOC coding algorithm, it is useful to note that the less granular levels of classification are likely to have fewer incorrect classifications. There is a trade-off, as

	Manually assigned	Proprietary algorithm
Sample size	330	67,900
Accuracy	76%	91%

Table 6: Summary of evaluation of SOC coding algorithm against ONS coding (3-digit level). Source: ONS.

going to more aggregate classifications loses some of the rich heterogeneity which we find in the data.

Since we developed our approach, we became aware of several recent similar approaches. Atalay et al. (2017) labels job vacancy adverts appearing in US newspapers with SOC codes. Their approach shares some similarities with ours, including the use of cosine similarity, but is also different in several respects: our model is created from the official job category descriptions, while theirs is created from the vacancy text; while we use tf-idf to create a vector space, they use continuous-bag-of-words; and finally they match to US SOC codes, while we match to UK SOC codes. We think that one advantage of creating the vector space from the official descriptions of SOC codes is that it only retains words or terms which are relevant to solving the problem of finding the right SOC code and discards all other words. This is not true when the vector space is created from the vacancy text. The vector space created the former way is inherently limited by the cardinality of SOC codes, which is a benefit, rather than potentially growing indefinitely as more job adverts are added in the latter approach. Working with self-reported job title data from the German General Social Survey, Gweon et al. (2017) develop three different statistical approaches to apply occupational classifications. Boselli et al. (2017*a*,*b*) take a different approach and manually label around 30,000 vacancies to then use a supervised machine learning algorithm to classify a further 6 million vacancies using ISCO (International Standard Classification of Occupations) codes. We believe that the use of supervised machine learning to train a model could potentially produce more accuracy in matching (where accuracy is measured relative to the labels that a human coder would select). However, the maintenance cost of the supervised approach is higher; if the SOC code standard changes, our approach would be trivial to update with the new master descriptions of each SOC code, but a supervised machine learning approach would need to be re-trained with, presumably, 30,000 more vacancies labelled by humans. Similarly, applying the same approach in different countries would require model retraining. Future work could usefully compare or combine all of these methods on the same SOC matching problem.

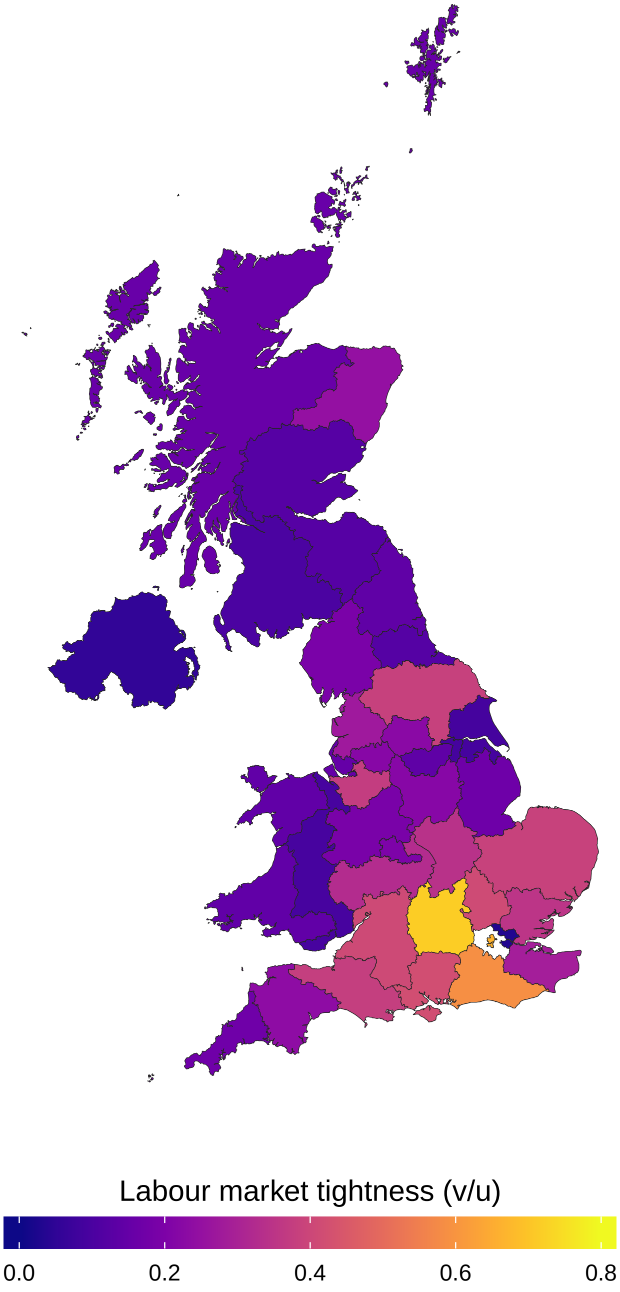

Figure 6: Map of mean UK labour market tightness, $\theta = \left(\frac{v}{u}\right)$, by 2-character NUTS code over the period 2008Q1–2016Q4. Some of the NUTS classifications are different in the ONS data relative to the NUTS2010 standard (EUROSTAT., 2011). This causes problems for London (UKI). We map UKI1 to UKI3 and UKI2 to UKI5. This neglects the UKI6 and UKI7 categories in NUTS2010. These are shown as white in the plot.

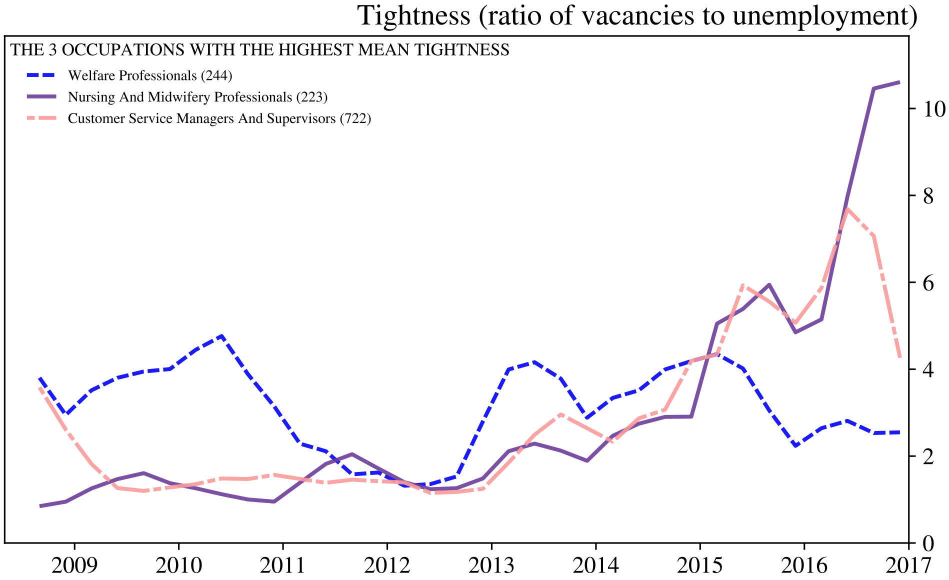

Figure 7: The 3-digit SOC codes with the three highest mean tightnesses over the full time period. Source: Reed, LFS.

5 Analysis of processed data

Once labelled with both regional NUTS codes and occupational SOC codes, the data allow for an entirely new perspective on the heterogeneity of labour demand within the UK. In this section, we report assorted summary statistics which illustrate this. Figure 6 shows the labour market tightness, $\theta = \frac{v}{u}$, by 2-character NUTS code. Unemployment data come from the *Labour Force Survey*. The picture reflects regional incomes, with the South East having higher tightness than Northern Ireland. However, there are isolated regions of tightness outside of the South East.

We can also look at changes in tightness which occur at an extremely disaggregated level, although some caution should be taken in inferring too much from changes at the lowest possible levels of disaggregation given that the data have been reweighted from a biased source. In Figure 7 we plot the rolling 2 quarter means of the three highest mean tightnesses for 3-digit SOC codes. The appearance of nurses and welfare professionals in the three most tight occupations is consistent with the UK Government's 'Shortage Occupation List'. Not shown are the bottom three occupations, which were elementary sales occupations, process operatives, and elementary agricultural occupations. While these are likely to have low tightnesses in part due to genuinely low demand, it is also likely that these jobs are not commonly posted online by firms.

Another useful check on the newly compiled vacancy data are that they satisfy similar stylised facts to

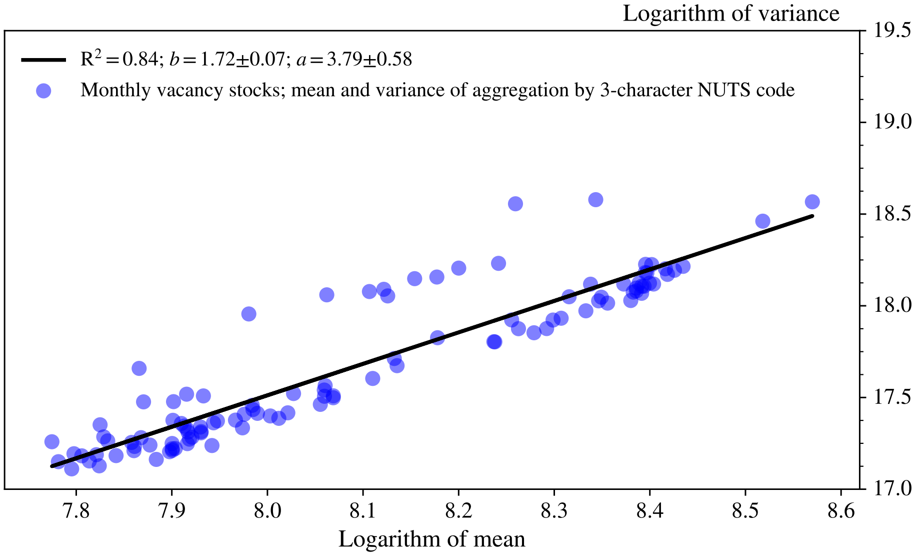

Figure 8: Monthly vacancy stocks, when aggregated by region into mean and variance, show a clear Taylor power law relationship. Source: Reed.

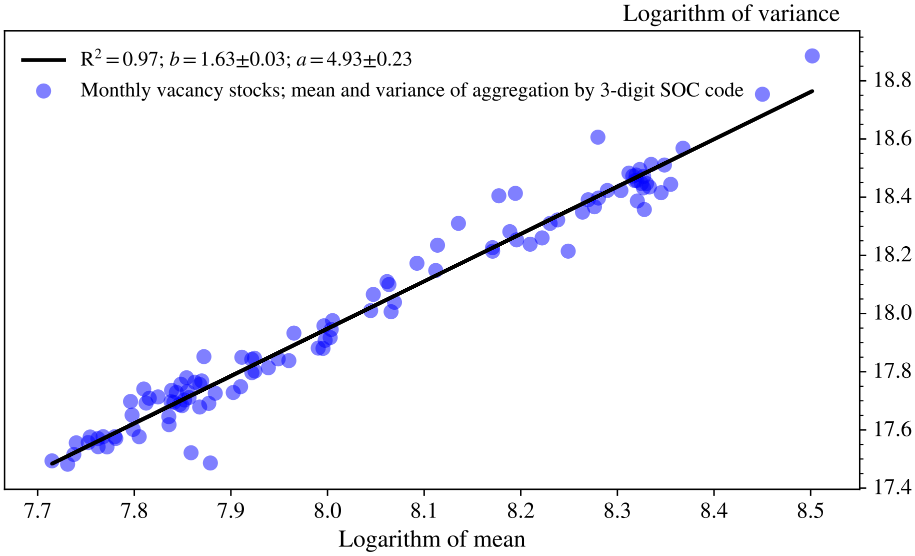

Figure 9: Monthly vacancy stocks, when aggregated by occupation into mean and variance, show a clear Taylor power law relationship. Source: Reed.

the official data produced by the ONS. One such stylised fact is that the monthly sectoral vacancy stocks follow a Taylor power law (Taylor, 1961). Firm sizes have also been shown to satisfy this law (Gaffeo et al., 2012). Let i represent a region, occupation, or sector with $\overline{V}_t = \frac{1}{I}\sum_i V_{t,i}$. Then the monthly mean and monthly variance, σ_t^2, are related to each other as

$$\sigma_t^2 = a\overline{V}_t^b$$

where the power, b, is sometimes called the index of aggregation[11]. The ONS vacancies by sector strongly follow a Taylor power law with $R^2 = 0.857$ and $b = 2.037 \pm 0.061$. We show, in Figures 8 and 9, that the breakdowns by NUTS and SOC respectively do both strongly follow Taylor power laws, giving confidence in the methods used to produce these statistics. We also highlight the existence of these Taylor power laws in vacancy data as they could be useful for the calibration of heterogeneous models of the labour market.

The descriptive statistics of the Reed data at the disaggregated level seem to provide a plausible representation of vacancies by both occupation and region.

6 Use of Reed vacancy data

We demonstrate potential uses of these new economic statistics.

By combining Reed vacancy data labelled by occupation with data on unemployment and hires from the *Labour Force Survey*, we are able to estimate Beveridge curves[12]. These track the relationship between unemployment and vacancies over time. By utilising vacancy data labelled by the text analysis technique developed, we are able to create Beveridge curves at the occupational level.

At the aggregate level, we assume a matching function M which takes the level of vacancies and unemployment in discrete time as inputs and outputs the number of hires (per unit time) as in the comprehensive survey by Petrongolo and Pissarides (2001). Define the aggregate number of hires, h, and matching function, M, with constant returns to scale (homogeneous of degree 1) as

$$h(U,V) = \phi M(U,V) = \phi U^{1-\alpha}V^{\alpha}$$

where ϕ is the matching efficiency and α is the vacancy elasticity of matching. These are structural parameters. Matches and new hires from unemployment are equivalent. At the disaggregated level, hires

[11] $1 < b < 2$ indicates that the variation falls with increasing size of region, occupation, or sector relative to what would be expected from a relationship with constant per vacancy variability, which is $b = 2$ (Kilpatrick and Ives, 2003).

[12] See Elsby, Michaels and Ratner (2015) for a comprehensive review.

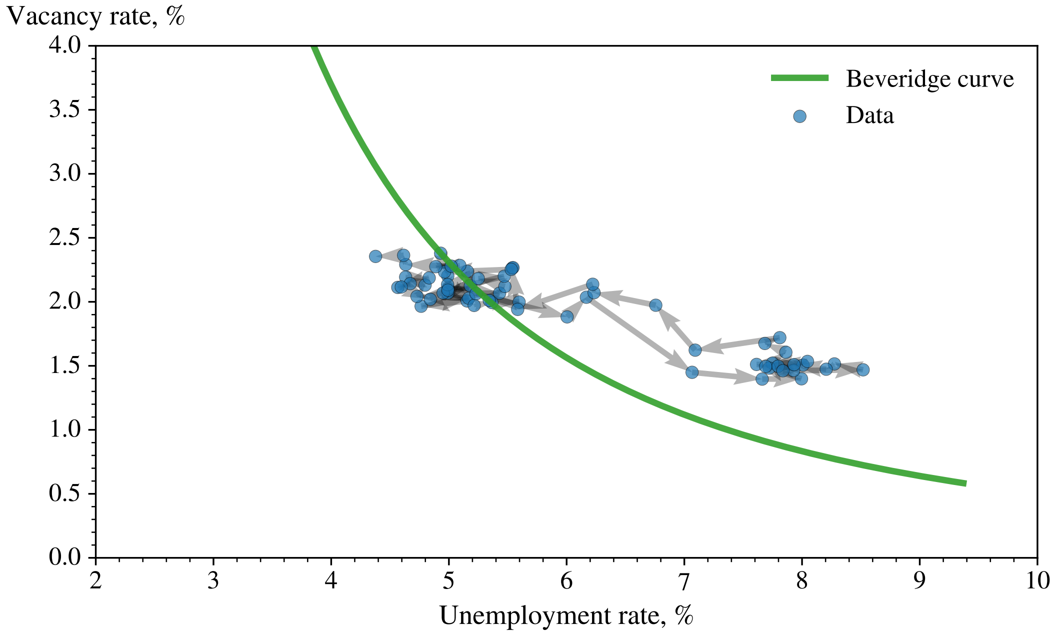

Figure 10: Beveridge curve (line) versus aggregate u-v data at quarterly frequency. Source: Reed, ONS.

are given by h_i. Hires based upon the theoretical matching function and a segment-specific matching efficiency are given by

$$h_i = \phi_i M(U_i, V_i) = \phi_i U_i^{1-\alpha} V_i^{\alpha} \tag{2}$$

The key structural parameters are the scale parameter of the matching function, ϕ, and the vacancy elasticity parameter, $\alpha = \frac{V}{M}\frac{\partial M}{\partial V}$. The scale parameter is often interpreted as an indicator of the level of efficiency of the matching process, hence we refer to it as the 'matching efficiency'. The elasticity parameter contains information about the severity of the congestion externalities that searchers on either side of the labour market impose on each other. Econometric estimates are reported in full in Turrell et al. (2018). When the number of hires is equal to the job destruction rate and $\frac{\mathrm{d}U}{\mathrm{d}t} = 0$, the combinations of possible U and V values for a given set of matching parameters trace out a locus of points in $U - V$ space. This is the Beveridge curve, and its empirical counter-part may be seen by plotting observed U and V values against one another.

Figure 10 shows an aggregate fitted Beveridge curve against aggregate vacancy-unemployment points at quarterly frequency for 2008 to 2017.[13] The aggregate matching efficiency is $\phi = \exp\{0.554 \pm .037\}$ (significant to 1%). Arrows indicate movements over time, and a shift toward higher unemployment during the Great Recession is evident, as is the high tightness in the last quarter of 2016.

[13] In the LFS data, there are discrepancies between the stocks implied by the flows in the longitudinal data and the stocks in the cross-sectional data. Due to this, we calibrate the job destruction rate in the Beveridge curves to give the best fit to the data.

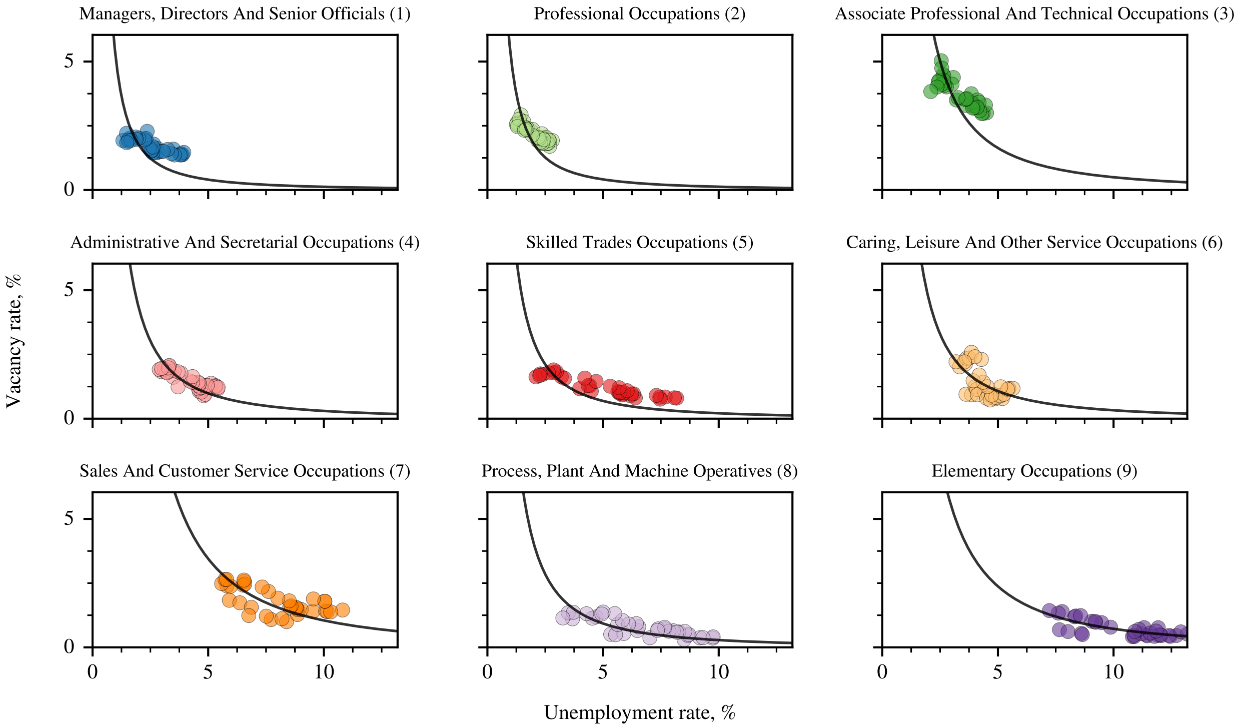

Figure 11: Beveridge curves (lines), estimated with Reed data, and Reed data (points) in u-v space for each 1-digit SOC code at quarterly frequency. Source: Reed, ONS.

Figure 11 shows the disaggregated equivalent of Figure 10, with Beveridge curves and quarterly u-v points for each 1-digit SOC code. The sub-market level Beveridge curves show that a single, aggregate Beveridge curve hides a great deal of important variation in u-v space across SOC codes. There are significant differences between the apparent curves as separated by skill, with the curve for associate professional and technical occupations shifted up relative to other occupations. There are also differences in spread. The driver of the spread varies by occupation; for the Caring, Leisure and Other Service occupation (1-digit SOC code 6), it is largely driven by vacancies, while what variation there is for Managers, Directors and Senior Officials (1-digit SOC code 1) is driven by unemployment. We do not allow matching efficiency or job destruction rates to vary over the time period here so that the Beveridge curve is fixed. In practice there are shifts in Beveridge curves, certainly at the aggregate level, and these are documented for the US in Barnichon et al. (2012). They find that a break in the hires per vacancy shifted the curve so that the implied unemployment rate was 2.8 percentage points lower than the actual unemployment rate. Our short time series makes a similar analysis difficult here but the estimated Beveridge curves at the occupational level provide a good fit for the entire period.

The patterns shown here could be affected by the biases discussed in §3.1 and §3.2. The vacancy stocks of higher numbered occupations are subject to both an upward bias, due to the likelihood of having vacancy durations shorter than the average across occupations and the 6 weeks assumed for the Reed data, and a downward bias, due to their being underrepresented amongst online vacancies posted at cost.

We now turn to the mismatch framework of Şahin et al. (2014), which, for heterogeneous labour markets, can determine the extent of unemployment which arises due to mismatch between jobseekers and job vacancies. Mismatch arises when there are barriers to mobility across distinct parts of the labour market, which we refer to as sub-markets or market segments. Mismatch lowers the overall efficiency of the labour market; given the aggregate level of unemployment and vacancies, it lowers the rate of job finding. The mismatch framework is also used by Smith (2012), Patterson et al. (2016), and Turrell et al. (2018) – from which the econometric estimates used here are drawn.

The Şahin et al. (2014) model provides counter-factuals due to a social planner who allocates the unemployed to search in sub-markets so as to optimise output. The social planner takes into account the matching efficiency and tightness of each sub-market. Mismatch unemployment is defined as the gap between actual unemployment, u, and counter-factual unemployment, u^*. We compute this mismatch unemployment rate in Figure 12 using 1-digit SOC codes. The biases which affect the stock of vacancies also affect estimates of the matching efficiency, producing a bias both upwards and downwards for occupations

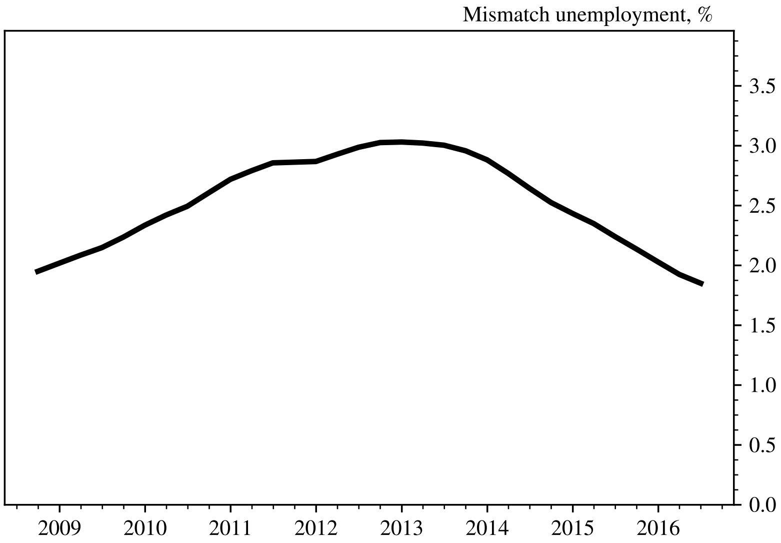

Figure 12: Mismatch unemployment, $u - u^*$ (seasonally adjusted). Source: ONS, Reed.

with short vacancy durations and low online representation respectively. Upwards and downwards bias in matching efficiency make mismatch unemployment seem lower or higher respectively. For following the trend in mismatch unemployment, these biases are unimportant, as they are likely to be relatively fixed over the period under consideration. Following the recession caused by the Great Financial Crisis, mismatch unemployment gradually rose. The maximum inflection point at the end of 2012 coincides with the UK's last negative quarter-on-quarter GDP growth within the time period under consideration; mismatch unemployment subsequently falls more steeply during the recovery. Mismatch between jobseekers and firms has been implicated as one driver of the UK's productivity puzzle (Patterson et al., 2016) but the trend in mismatch unemployment seen here suggests that, while that could have been a factor up until 2013, the role it has played fell substantially between 2013 and 2017.

7 Conclusion

We mapped naturally occurring vacancy data into official occupational classifications in order to construct new economic statistics. The algorithm we have developed is especially useful for firms, recruitment agencies, and other researchers seeking to apply consistent occupational labels to free form job descriptions. The tools and processes developed can be deployed on other vacancy data, but could also be adapted to other types of naturally occurring text data.

We have considered the limitations due to bias and coverage in the Reed vacancy data presented. While there is undoubtedly bias in the data, we have provided a qualitative description of it and how it might affect the estimates of the stock of vacancies. We also quantified the biases by sector and reweighted the data in order to reduce the overall bias, and increase the effective coverage of the data. The bias we find is no worse than in other widely used UK vacancy microdata. Example applications demonstrate the utility of these data for analysis.

These datasets are a complement, not a replacement, for existing survey based approaches to constructing economic statistics because those existing statistics are required to assess the extent of bias and coverage in new datasets, and to create weighting schemes. We have shown that the Reed data, transformed by text analysis, can augment existing official statistics because they can give estimates of vacancies by occupation and region which survey data do not, and because of their vast scale. That scale permits very disaggregated analysis which can substantially benefit labour market research.

Acknowledgements

The authors would like to thank Katherine Abraham, James Barker, David Bholat, David Bradnum, Emmet Cassidy, Matthew Corder, Rodrigo Guimaraes, Frances Hill, Tomas Key, Graham Logan, Michaela Morris, Michael Osbourne, Kate Reinold, Paul Robinson, Ayşegül Şahin, Ben Sole, Vincent Sterk, and the participants of the NBER Conference on Big Data in the 21st Century. We would especially like to thank William Abel for his help throughout the project.

References

Abraham, Katharine G. 1983. “Structural/frictional vs. deficient demand unemployment: some new evidence.” *The American Economic Review*, 73(4): 708–724.

Abraham, Katharine G, and Michael Wachter. 1987. “Help-wanted advertising, job vacancies, and unemployment.” *Brookings papers on economic activity*, 1987(1): 207–248.

Atalay, Enghin, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. 2017. “The Evolving US Occupational Structure.” Discussion paper.

Azar, José A, Ioana Marinescu, Marshall I Steinbaum, and Bledi Taska. 2018. “Concentration in US labor markets: Evidence from online vacancy data.” National Bureau of Economic Research.

Barnichon, Regis. 2010. “Building a composite help-wanted index.” *Economics Letters*, 109(3): 175–178.

Barnichon, Regis, Michael Elsby, Bart Hobijn, and Aysegul Sahin. 2012. "Which industries are shifting the Beveridge curve." *Monthly Lab. Rev.*, 135: 25.

Belloni, Michele, Agar Brugiavini, Elena Maschi, and Kea Tijdens. 2014. "Measurement error in occupational coding: an analysis on SHARE data." Department of Economics, University of Venice "Ca' Foscari" Working Papers 2014: 24.

Bentley, R. 2005. "Publication of JobCentre Plus vacancy statistics." *ONS Reports*, Labour Market Trends.

Boselli, Roberto, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. 2017*a*. "Using machine learning for labour market intelligence." 330–342, Springer.

Boselli, Roberto, Mirko Cesarini, Stefania Marrara, Fabio Mercorio, Mario Mezzanzanica, Gabriella Pasi, and Marco Viviani. 2017*b*. "WoLMIS: a labor market intelligence system for classifying web job vacancies." *Journal of Intelligent Information Systems*, 1–26.

Burgess, Simon, and Stefan Profit. 2001. "Externalities in the Matching of Workers and Firms in Britain." *Labour Economics*, 8(3): 313–333.

Cajner, Tomaz, David Ratner, et al. 2016. "A Cautionary Note on the Help Wanted Online Data." *FEDS Notes, Board of Governors of the Federal Reserve System https://www. federalreserve. gov/econresdata/notes/feds-notes/2016/acautionary-note-on-the-help-wanted-online-data-20160623. html.*

Coles, Melvyn G, and Eric Smith. 1996. "Cross-section estimation of the matching function: evidence from England and Wales." *Economica*, 589–597.

Deming, David, and Lisa B Kahn. 2017. "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals." National Bureau of Economic Research.

Elsby, Michael WL, Ryan Michaels, and David Ratner. 2015. "The Beveridge curve: A survey." *Journal of Economic Literature*, 53(3): 571–630.

EUROSTAT. 2011. "Regions in the European Union. Nomenclature of territorial units for statistics. NUTS 2010/EU-27. 2011 edition."

FRED, Federal Reserve Bank of St. Louis. 2019. "DHI-DFH Mean Vacancy Duration Measure." *https://fred.stlouisfed.org/series/DHIDFHMVDM.*

Gaffeo, Edoardo, Corrado Di Guilmi, Mauro Gallegati, and Alberto Russo. 2012. "On the mean/variance relationship of the firm size distribution: Evidence and some theory." *Ecological complexity*, 11: 109–117.

Gweon, Hyukjun, Matthias Schonlau, Lars Kaczmirek, Michael Blohm, and Stefan Steiner. 2017. "Three methods for occupation coding based on statistical learning." *Journal of Official Statistics*, 33(1): 101–122.

Hershbein, Brad, and Lisa B Kahn. 2018. "Do recessions accelerate routine-biased technological change? Evidence from vacancy postings." *American Economic Review*, 108(7): 1737–72.

Kilpatrick, AM, and AR Ives. 2003. "Species interactions can explain Taylor's power law for ecological time series." *Nature*, 422(6927): 65.

Levenshtein, Vladimir I. 1966. "Binary codes capable of correcting deletions, insertions, and reversals." Vol. 10, 707–710.

Machin, Andrew. 2003. "The Vacancy Survey: a new series of National Statistics." *ONS Reports*, National Statistics feature.

Mamertino, Mariano, and Tara M Sinclair. 2016. "Online Job Search and Migration Intentions Across EU Member States." Institute for International Economic Policy Working Paper Series.

Manning, Alan, and Barbara Petrongolo. 2017. "How local are labor markets? Evidence from a spatial job search model." *American Economic Review*, 107(10): 2877–2907.

Marinescu, Ioana. 2017. "The general equilibrium impacts of unemployment insurance: Evidence from a large online job board." *Journal of Public Economics*, 150: 14–29.

Marinescu, Ioana, and Ronald Wolthoff. 2016. "Opening the black box of the matching function: The power of words." National Bureau of Economic Research.

Mathiowetz, Nancy A. 1992. "Errors in reports of occupation." *The Public Opinion Quarterly*, 56(3): 352–355.

Mellow, Wesley, and Hal Sider. 1983. "Accuracy of response in labor market surveys: Evidence and implications." *Journal of Labor Economics*, 1(4): 331–344.

Mortensen, Dale T, and Christopher A Pissarides. 1994. "Job creation and job destruction in the theory of unemployment." *The review of economic studies*, 61(3): 397–415.

Office for National Statistics. 2017. "Quarterly Labour Force Survey, 1992-2017: Secure Access. [data collection]. 10th Edition." *http://dx.doi.org/10.5255/UKDA-SN-6727-11*, Social Survey Division, Northern Ireland Statistics and Research Agency. Central Survey Unit.

Patterson, Christina, Ayşegül Şahin, Giorgio Topa, and Giovanni L Violante. 2016. "Working hard in the wrong place: A mismatch-based explanation to the UK productivity puzzle." *European Economic Review*, 84: 42–56.

Petrongolo, Barbara, and Christopher A Pissarides. 2001. "Looking into the black box: A survey of the matching function." *Journal of Economic literature*, 39(2): 390–431.

Şahin, Ayşegül, Joseph Song, Giorgio Topa, and Giovanni L Violante. 2014. "Mismatch unemployment." *The American Economic Review*, 104(11): 3529–3564.

Schierholz, Malte, Miriam Gensicke, Nikolai Tschersich, and Frauke Kreuter. 2016. "Occupation coding during the interview." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Smith, Jennifer C. 2012. "Unemployment and Mismatch in the UK."

Taylor, LR. 1961. "Aggregation, variance and the mean." *Nature*, 189(4766): 732–735.

Turrell, Arthur, Bradley Speigner, Jyldyz Djumalieva, David Copple, and James Thurgood. 2018. "Using job vacancies to understand the effects of labour market mismatch on UK output and productivity." Bank of England Staff Working Paper 737.

A Detailed occupation-coding algorithm flow diagram

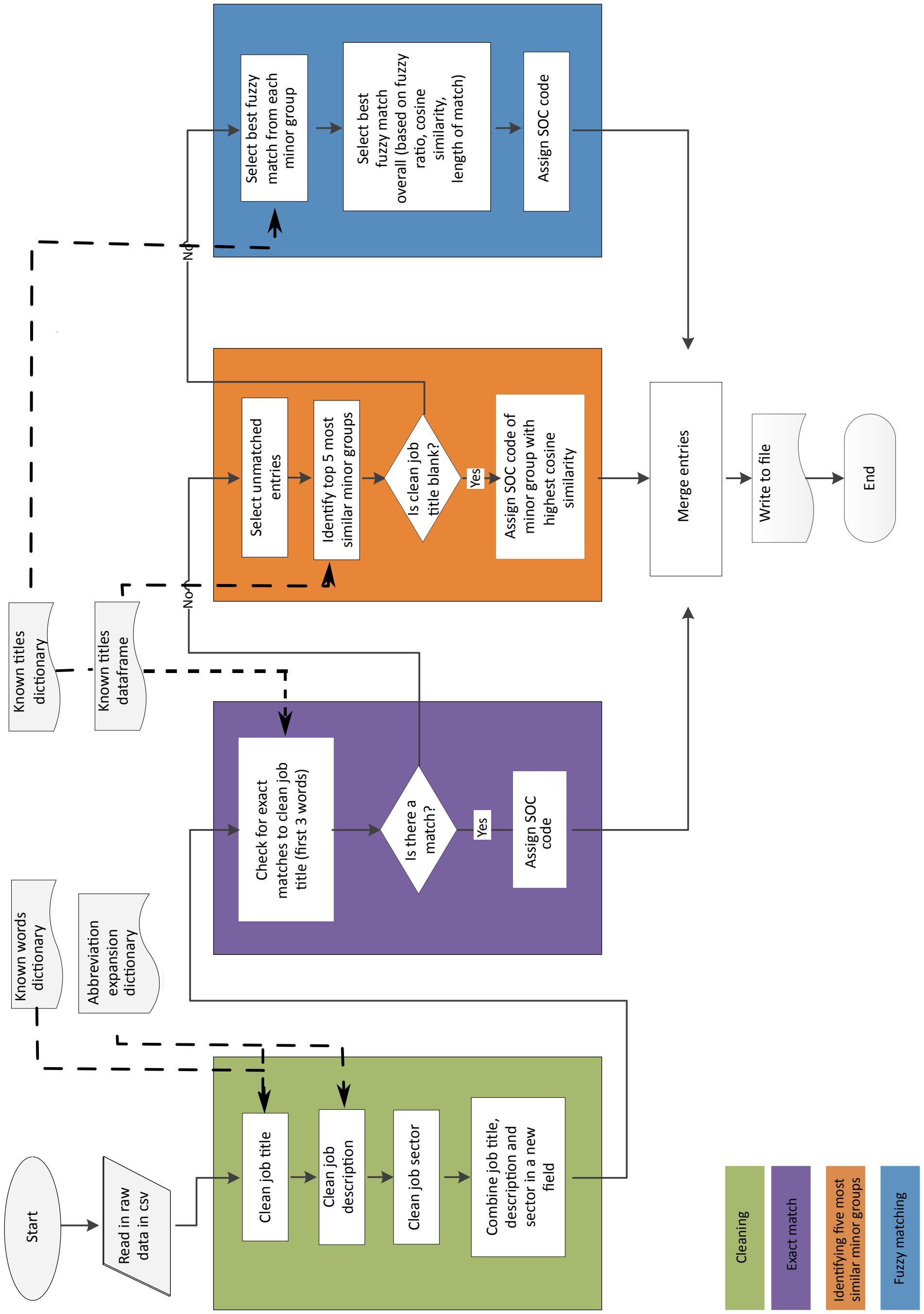
Start
Read in raw data in csv
Clean job title
Clean job description
Clean job sector
Combine job title, description and sector in a new field
Known words dictionary
Abbreviation expansion dictionary
Check for exact matches to clean job title (first 3 words)
Is there a match?
Yes
Assign SOC code
No
Known titles dictionary
Known titles dataframe
Select unmatched entries
Identify top 5 most similar minor groups
Is clean job title blank?
Yes
Assign SOC code of minor group with highest cosine similarity
No
Select best fuzzy match from each minor group
Select best fuzzy match overall (based on fuzzy ratio, cosine similarity, length of match)
Assign SOC code
Merge entries
Write to file
End
Cleaning
Exact match
Identifying five most similar minor groups
Fuzzy matching

Figure 13: A more detailed overview of the algorithm which matches job vacancies to SOC codes at the minor group level (3-digit SOC code).