BANK OF ENGLAND

Staff Working Paper No. 742

Using online job vacancies to understand the UK labour market from the bottom-up

Arthur Turrell, James Thurgood, David Copple, Jyldyz Djumalieva and Bradley Speigner

July 2018

BANK OF ENGLAND

Staff Working Paper No. 742
Using online job vacancies to understand the UK labour market from the bottom-up

Arthur Turrell,[1] James Thurgood,[2] David Copple,[3] Jyldyz Djumalieva[4] and Bradley Speigner[5]

Abstract

What type of disaggregation should be used to analyse heterogeneous labour markets? How granular should that disaggregation be? Economic theory does not currently tell us; perhaps data can. Analyses typically split labour markets according to top-down classification schema such as sector or occupation. But these may be slow-moving or inaccurate relative to the structure of the labour market as perceived by firms and workers. Using a dataset of 15 million job adverts posted online between 2008 and 2016, we create an empirically driven, 'bottom-up' segmentation of the labour market which cuts across wage, sector, and occupation. Our segmentation is based upon applying machine learning techniques to the demand expressed in the text of job descriptions. This segmentation automatically identifies traditional job roles but also surfaces sub-markets not apparent in current classifications. We show that the segmentation has explanatory power for offered wages. The methodology developed could be deployed to create data-driven taxonomies in conditions of rapidly changing labour markets and demonstrates the potential of unsupervised machine learning in economics.

Key words: Vacancies, classification, disaggregation.

JEL classification: J6, J42, C55.

(1) Bank of England. Email: arthur.turrell@bankofengland.co.uk (corresponding author)
(2) Bank of England. Email: james.thurgood@bankofengland.co.uk
(3) Bank of England. Email: david.copple@bankofengland.co.uk
(4) Nesta. Email: jyldyz.djumalieva@nesta.org.uk
(5) Bank of England. Email: bradley.speigner@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are grateful to James Barker, David Bholat, David Bradnum, Matthew Corder, Rodrigo Guimaraes, Frances Hill, Tomas Key, Ioana Marinescu, Kate Reinold, Paul Robinson, and Ben Sole for comments and suggestions. We would especially like to thank William Abel for his help throughout the project.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Publications and Design Team, Bank of England, Threadneedle Street, London, EC2R 8AH
Telephone +44 (0)20 7601 4030 email publications@bankofengland.co.uk

ISSN 1749-9135 (on-line)

1 Introduction

Large, naturally occurring datasets present big opportunities for understanding the economy in new levels of detail. They can complement more traditional data sources, such as the surveys often used in economics. These data, which may be captured in the course of using a phone or a website for another purpose, can be cheaper, more granular, and on scales in excess of what is practical for a survey.

One way that they can help is by giving new perspectives on how markets are organised. When economists and statisticians assess the labour market in detail, looking at the demand for workers of specific types, they use schema such as occupation or sector. These classifications help to make sense of the millions of jobs in the economy by putting them into similar buckets. They are carefully designed by statisticians. Because these schema are fixed for many years at a time, they allow for analysis with a time dimension.

However, there is a cost to this; it ignores changes in the 'true' set of jobs in the economy. Demand for new jobs may emerge, and older jobs become less important. So, alongside fixed schema, it is useful to have real-time, data-driven methods to classify jobs. Real-time job classifications could be used to understand structural changes in the labour market and also to inform the design of future classification schema.

The demand for labour is richly heterogeneous, but neither the level nor the type of disaggregation which are appropriate to use to study it are readily apparent. This has been widely acknowledged, for instance in Barnichon and Figura (2015), "The appropriate size of a labor market segment, i.e., the definition of the labor market unit, is an open question in the literature", and Petrongolo and Pissarides (2001), "The key problem here is to define the unit of the micro market". By type, we mean distinct ways to organise the labour market, for instance by region, occupation, or sector. By level, we mean the level of disaggregation of the labour market: with region as the type, this could be at the union (e.g. United Kingdom), country, or county level.

Using job adverts posted daily online between 2008 and 2016, we create an empirically driven, 'bottom-up' segmentation of demand in the labour market which cuts across wage, sector, and occupation. Our segmentation is created by applying machine learning techniques to the demand for labour expressed in the text of job descriptions. We assume that differences in firms' demand are the most natural organising structure for the labour market because they are the key determinant of whether a worker can take up a job or not.

We use text analysis techniques from machine learning to group job vacancies based on the similarity of their job descriptions in an attempt to get at similarities in what is truly demanded by firms. We use

latent Dirichlet allocation (Blei, Ng and Jordan, 2003; Pritchard, Stephens and Donnelly, 2000), weighted saliency (Chuang, Manning and Heer, 2012; Goldsmith-Pinkham, Hirtle and Lucca, 2016), silhouette scores (Rousseeuw, 1987) and the K-means algorithm (Lloyd, 1982) to group vacancies. This abstracts from occupation, sector, or region unless those factors have an influence on the skills demanded. Our results show that the clusters can reproduce groups of jobs familiar in existing classification schemes. We also show that these clusters can highlight new careers, not well captured by existing classifications. Furthermore, our clusters, which are created based only upon text data, have explanatory power for offered wages in vacancies – both in absolute and marginal terms (relative to other categories).

We take advantage of categories which appear in both labour force survey microdata and our bottom-up clusters in order to take our clusters to the supply side of the labour market. We use supervised machine learning to label each individual in the survey data with a data-driven cluster classification label. We perform similar tests to those used on labour demand and find the same results; our bottom-up clusters are powerful at capturing well-established groups of workers and contain information which can help to predict accepted real wages.

Our 'bottom-up' approach is one strategy to resolve the question of what level of disaggregation is appropriate for understanding the demand for labour. If estimates of structural parameters are dependent on the type of disaggregation used, this method also provides an organising framework which is more type-neutral than a disaggregation explicitly by, for example, sector. It could also be applied repeatedly in different periods to see how the structure of the labour market changes over time. More broadly, the methodology we introduce to do create a 'natural' disaggregation could be used to understand a range of markets beyond the specific case of the labour market for which it is developed. The strong results we find with these clusters is indicative of the utility of this approach.

Our paper adds to a small but 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 standard occupation classification (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. Grinis (2017) find that skills associated with so-called STEM subjects (Science, Technology, Engineering, and Mathematics) are often demanded in job vacancies for non-STEM occupations – providing some motivation for this work. Turrell et al. (2018) uses the text in job vacancies to apply the usual SOC codes to them, and then looks at counter-factuals for productivity and output growth in the absence of occupational mismatch. Atalay

et al. (2017) follow a similar process for labelling vacancies obtained from newspaper archives with SOC codes but use the processed data to estimate the extent to which task content shifts are within-occupation versus across-occupation in accounting for the aggregate decline of routine tasks.

The rest of the paper is structured as follows: §2 outlines why a data driven segmentation of the labour market might be useful, §3 describes the vacancy data which we use to construct the segments, §4 explains the clustering methodology, §5 describes the clusters as applied to job vacancies, and §6 applies the same clusters to the labour force. §7 concludes.

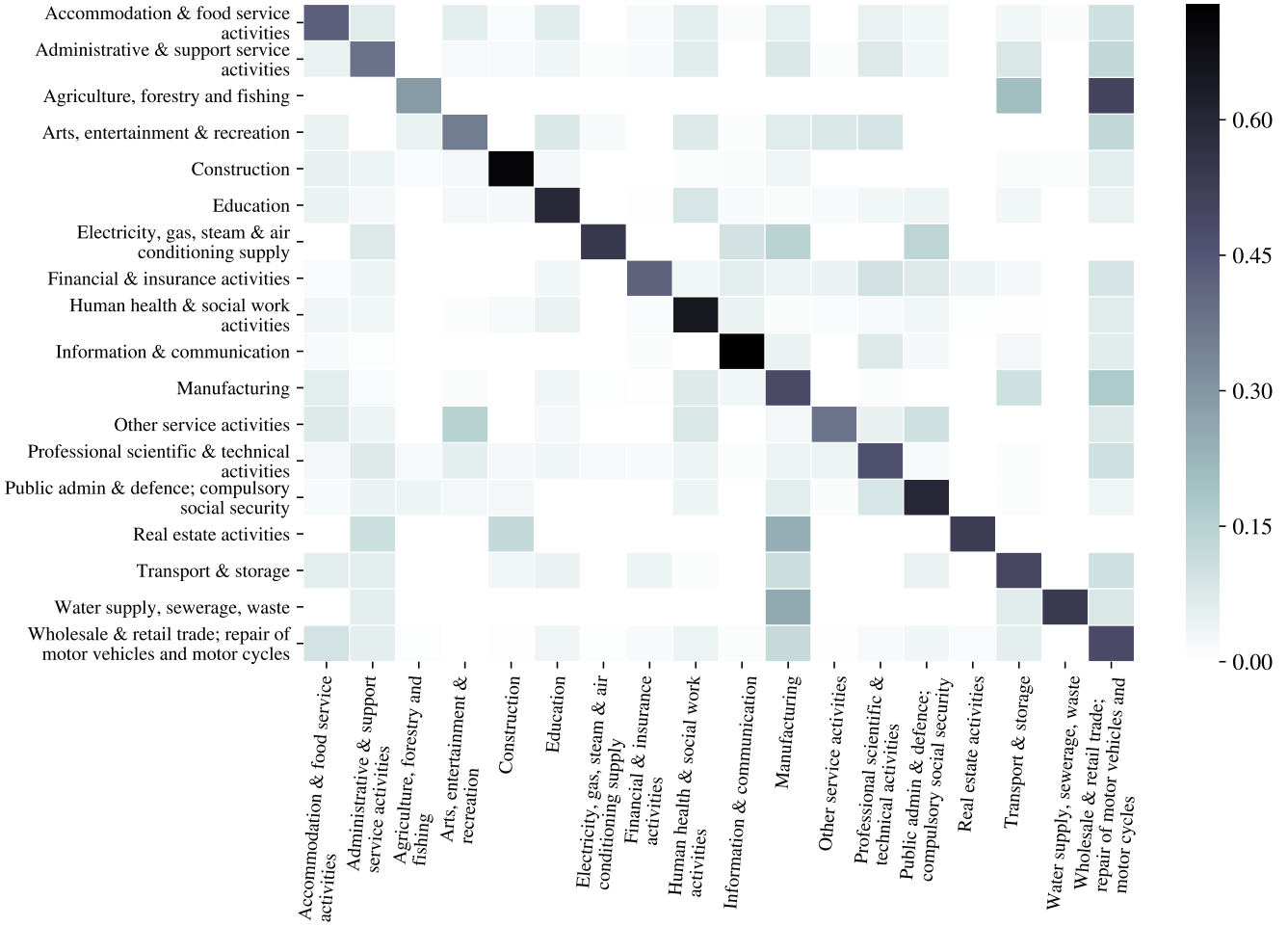

Figure 1: Quarterly job-to-job probability transition matrix from sector (rows) to sector (columns) averaged over 1997Q1 to 2017Q1 and normalised by the number of employed in each sector category in the first quarter. Data from the *Labour Force Survey*.

2 Motivation

Many analyses of labour markets rely upon official classifications which encode regions, sectors, or occupations. However, it is difficult to know what a 'true' segmentation of the labour market should or would look

like. If there is substantial movement of workers across categories, analyses based on official classifications can produce biased results. The estimation of important structural parameters of models can be affected. As specific examples, metrics of labour market 'mismatch' between workers and jobs are increasing in the level of disaggregation for the same type of market for both simple and sophisticated indices (Jackman and Roper, 1987; Şahin et al., 2014), suggesting a guide to an appropriate level of disaggregation may be useful. Turrell et al. (2018) gives an example where these same mismatch indices differ when changing the type of disaggregation. Furthermore, they show empirically that the point estimates of the elasticity parameter of a labour market matching function can be different across both the type and the level of disaggregation.

We show using data from the ONS *Labour Force Survey* (LFS) and *Annual Survey of Hours and Earnings* that workers do regularly transition across sector, as classified by Standard Industrial Classification (SIC) code, across occupation, as classified by Standard Occupational Classification (SOC) code, and across region, as classified by Nomenclature of Territorial Units for Statistics (NUTS) code when changing jobs. These data are shown in Figures 1, 2 and 3 respectively.

Official classifications are a useful means to categorise different types of roles in the labour market. They have the advantage that they are often applied to data in consecutive years or, if they are not, are applied according to a new standard onto which an old standard can usually be mapped. Similar classifications internationally allow for cross-country comparisons. But no official classification is perfect and each involves trade-offs. As noted, they may bias estimation of structural parameters if they do not reflect the underlying moves in the market. Nathan and Rosso (2015) and Hoberg and Phillips (2016) demonstrate that classifications may not be updated quickly enough to reflect either changing markets, or the changing nature of production.

Even if the structure is appropriate, official classifications can be applied inconsistently. Disagreements amongst those who code job titles into occupational classes, hereafter 'coders', can be substantial (Schierholz et al., 2016). 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") and reduces to 70–80% at the 3-digit level. Automated approaches which use job title alone have even lower levels of agreement; in Belloni et al. (2014), 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. Our analysis of vacancies through the lens of the clusters is not sensitive to the official classifications, and so has no bias due to these effects. However, the labour force survey data we use to look at accepted wages does employ official classifications.

Having the means to create market segments that are data driven is useful. The level and type of clas-

Table 1: Correlation matrix of aggregate vacancy 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

sification which emerges can be informative in itself. The methodology presented could be automatically deployed by statistical agencies from year to year as a guide to where new clusters of demand are forming, and to help understand what would be the most useful revisions to official classifications in future. For analysis of the labour market, estimates of structural parameters of models based on data driven classifications can give confidence to those based upon official classifications, if they are similar, or suggest that a re-think is required, if they are not. Counter-intuitively, and because the market segments are data driven with no reliance on official classifications, the methodology presented (but not the actual clusters) could help to resolve differences in point estimates across countries and time (which may necessarily use different official classifications) by producing more comparable results.

Finally, firm demand driven segmentations based upon our methodology could be used to predict worker flows, when combined with salient information about workers which is outside of the scope of firm demand (and therefore not included in our clusters), e.g. current region, age, and level of education.

3 Data

We use several datasets from the UK's Office for National Statistics (ONS), including the *Labour Force Survey* (LFS) (Office for National Statistics, 2017), the *Vacancy Survey*, and the *Annual Survey of Hours and Earnings* (ASHE).[1]

Our vacancies data are obtained from a job advertisement and employee recruiter, Reed.[2] They consist of approximately $15,242,000$ individual jobs posted at daily frequency from January 2008 to December 2016. The fields in the raw data which are typically available for each vacancy include a job posted date, an offered nominal wage, an idiosyncratic sectoral classification, a job location, a job title, and a job description. The value that our data add are that they can give vacancies split by region and occupation, two disaggregations which are not available in the official statistics on vacancies.

[1] We use the ONS mapping from SIC 2003 to SIC 2007 to make entries consistently labelled by SIC 2007 code. For SOC, we use fractional mappings from SOC 2k to SOC 2010 on counts to obtain consistently labelled entries. For transitions, such as 'unemployed' to 'employed', we use the modal mappings from SOC 2k to SOC 2010. NUTS 2010 is used throughout.

[2] These are not publicly available.

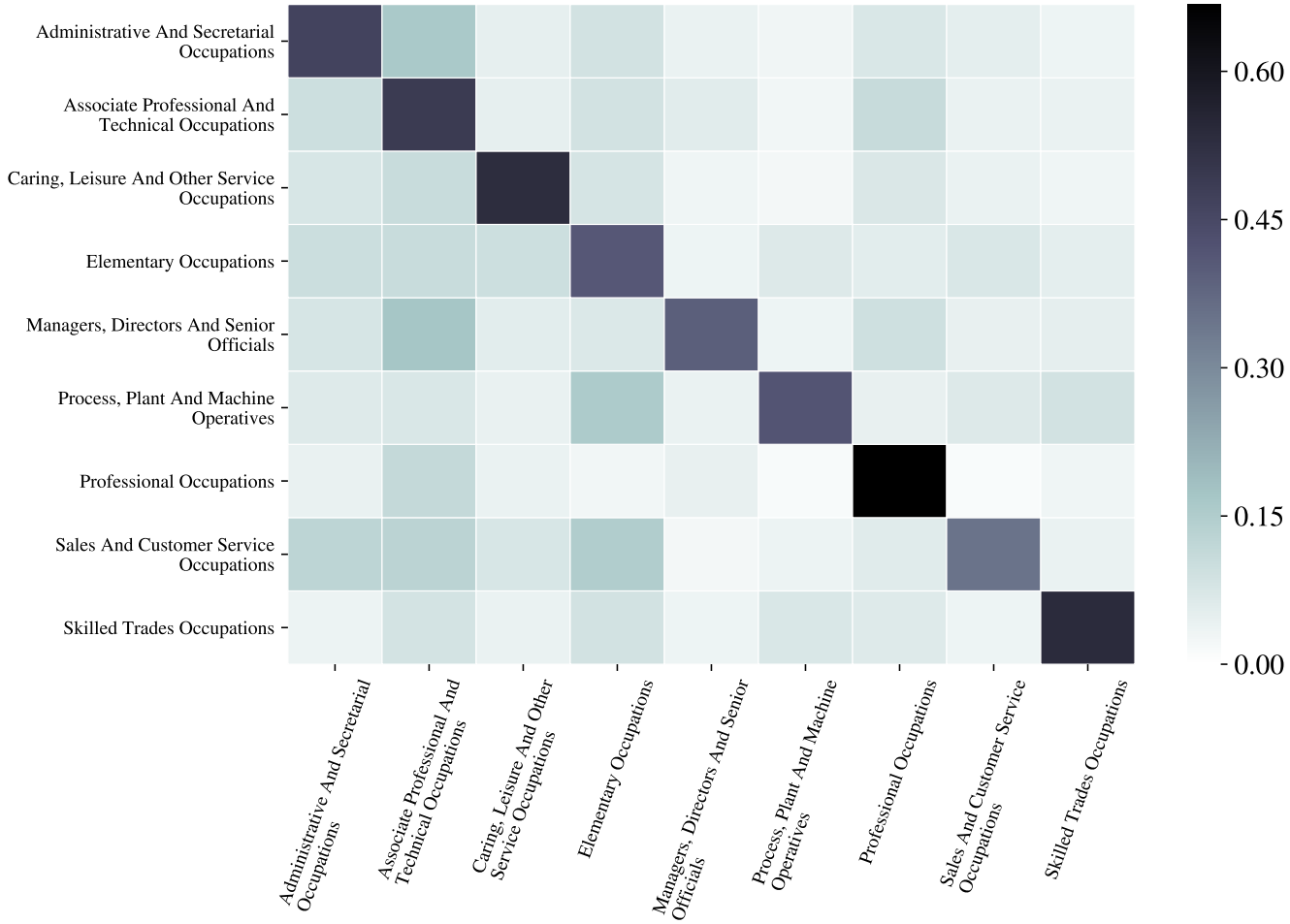

Figure 2: Quarterly job-to-job probability transition matrix from occupation (rows) to occupation (columns) averaged over 2007Q1 to 2017Q1 and normalised by the number of employed individuals in each occupation in the first quarter. Data from the *Labour Force Survey*.

Our data were originally posted online, and so do not cover all job vacancies. The raw aggregate series accounts for around 40% of UK vacancies annually. Previous work has found that online job vacancy postings can give a good indication of the trends in aggregate vacancies (Hershbein and Kahn, 2016). 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. Vacancies posted online are also unlikely to be representative of all vacancies advertised in the economy, introducing a potential source of bias.

We show the extent to which our data are representative by using the ONS Vacancy Survey as a comparator. The Vacancy Survey is disaggregated by firm size, and by sector. Our vacancy data have a

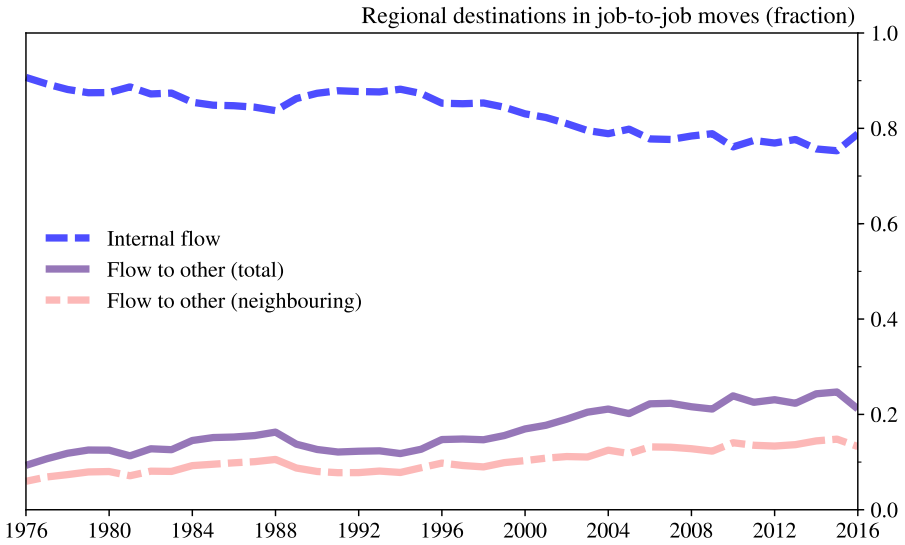

Figure 3: Regional job-to-job flows between 1-character UK NUTS regions. Data from the *Annual Survey of Hours and Earnings*.

sector field but no firm (or firm size) information, and so we use the appearance of sectoral counts in our data, and in the Vacancy survey, as a point of comparison to our data. These are shown as box and swarm plots over the mean annual ratios in Figure 4. Noting that the aggregate series under-represents all vacancies, the average ratio is greater than unity as expected, although sectors representing professionals, administrative staff, and other service activities are well captured by the data. The most clearly underrepresented sectors are public administration and manufacturing, together accounting for around 9% of vacancies in the last quarter of 2016 according to the Vacancy Survey. Around 64% of vacancies have an annual ratio with a median of less than 5. However, given the discrepancy between our data and the official UK data, we calculate weights for the stock of vacancies in the Reed data to make it more representative of UK vacancies as a whole. These weights are given by the reciprocals of the ratios shown in Figure 4.

In order to calculate the weights (equivalently, the ratios) some processing is necessary. In the Reed dataset, each individual vacancy is a flow, with entries removed after being on the site for 6 weeks. We transform this to be in terms of stocks.[3] The correlations, shown in Table 1, show that the aggregate, unweighted Reed vacancy time series is better correlated with the Vacancy Survey measure than the Job-Centre Plus data. To overcome the bias in the Reed vacancy data, and to ensure that it matches the

[3]In discrete time, this flow is $\dot{V}_d$ with d referring to a day. Therefore, to retrieve stocks, the data are transformed as follows where the time index refers to monthly frequency:

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

This aggregates the daily flow into monthly stocks.

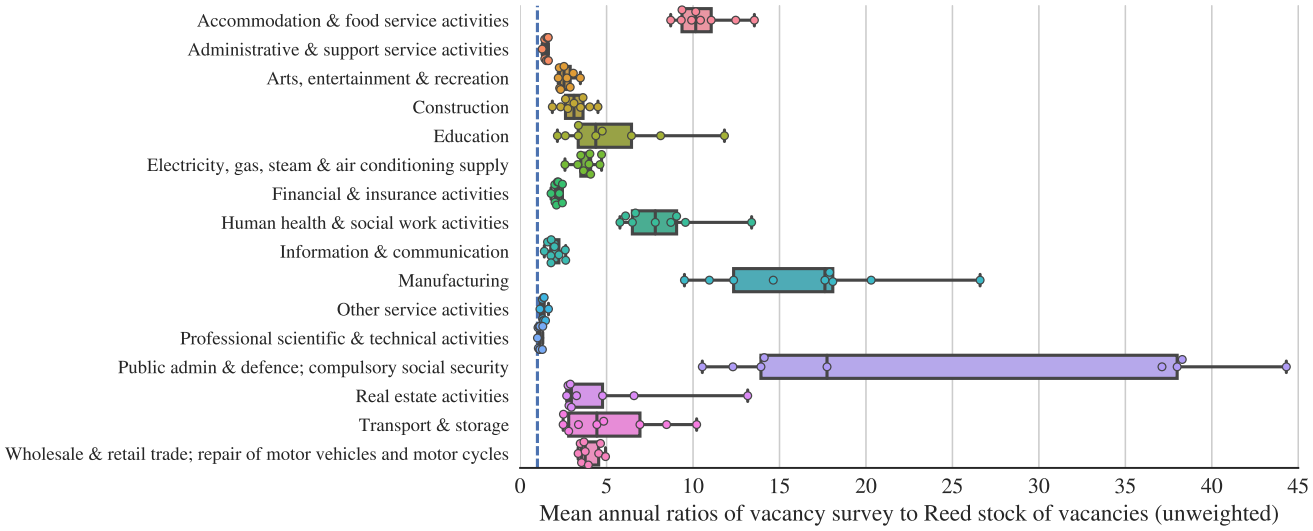

Figure 4: Box and swarm plots of the distribution of the mean annual ratio of the stocks of vacancies between the Vacancy Survey and the Reed data. Points along the dotted line indicate a ratio of unity for that sector.

Vacancy Survey in aggregate even more closely, we weight it using the monthly sectoral disaggregation of the Vacancy Survey which improves the correlation to the aggregate Vacancy Survey time series.[4] The vacancy data are labelled with official classifications following the procedures in Turrell et al. (2018), though these are not used to derive the market segmentation. We use the weighted version of the Reed vacancy data in two applications in this paper: to create vacancy time series, and to use as sampling weights for the vacancies which we use as inputs for the topic model in §4.

4 An empirically driven segmentation of the labour market

We assume that the skills demanded by firms are the most natural way to aggregate the labour market because this is a key determinant of whether a worker can move job or not. An accurate grouping of vacancies and employment would have predominantly diagonal entries in its equivalent of Figures 1 and 2 (if the clusters were also used on the supply of labour). To try and capture the bundles of skills that make up firms' demand for labour, our proxy variable is the text expressing the demand for labour in job vacancy descriptions in the Reed data.

Job descriptions do not provide a perfect or noiseless measure of demand. For example, some of the

[4]The bespoke Reed sectors are mapped into single letter Standard Industrial Classification (SIC) sections. The weight of a vacancy v in sector i and month m is given by

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

with VS the vacancy survey. This improves the match to the aggregate time series. Weighting the Reed data reduces bias but increases variance.

text in them is designed to prompt job-seekers to apply by extolling the virtues of the firm, or by explaining how exciting the role is. In the following, we explain how we abstract from some of this noise and only retain text features which do express the demand for labour.

To transform the demand expressed in job vacancies into mutually exclusive market segments which are data driven we follow a four step process. For computational reasons, only a third of vacancies are initially used to create the market segments. We sample these using the weights from the ONS Vacancy Survey (see §3).

In the first step, the text associated with each job vacancy is cleaned and the title and job description are combined into a single 'document' per vacancy. In the second step, a topic model is run over the entire longitudinal corpus, creating N topics which help determine the *type* of disaggregation which is the most appropriate given the input data. Our type will be a distribution over the usual official classifications. We also use an algorithm to choose N, thus helping to determining the best *level* of disaggregation in topics. In the usual classifications, this would reflect the granularity of the classification, for instance a single street, a village, a county, or a country. In our case, having more levels might result in a cluster which represents education and teaching as a whole being split into teachers and teaching assistants, or primary and secondary school teachers. *A priori*, we do not know how the algorithm will decide to split clusters at higher levels of disaggregation.

In the third step, each job vacancy is represented in the N-dimensional space of all topics and is then grouped into one of K market segments which are our final *types* of market segment. We interchangeably call them clusters as they draw together many jobs into groups. The intuition behind having both topics and clusters is that topics can represent many different skills, tasks, or aspects of a job separately, while clusters are larger groups of these skills. More practically, this step is needed because the objective is to obtain a classification which has discrete, mutually exclusive membership. We use the K-means algorithm (Lloyd, 1982) for this, which allows for a vacancy represented in a vector space to either be classed as k or k', but not as both. The representation of vacancies in the space of all topics is not suitable for this purpose as it is continuous, rather than discrete, and it is N-dimensional. Figure 5 is a schematic of how a set of points (vacancies) in the topic space are assigned to K mutually exclusive clusters. In keeping with the 'bottom-up' philosophy, we determine both K and N, on which K is dependent, using data-driven algorithms. The value for K then gives the final say on what the *level* of disaggregation is. With K fixed, these clusters define our 'natural' view of the UK labour market.

In the fourth and final step, we use machine learning on this cluster-labelled dataset to train a model

which can map unseen vacancies to a cluster. We use this trained model to apply cluster labels to the remaining 2/3 of the data.

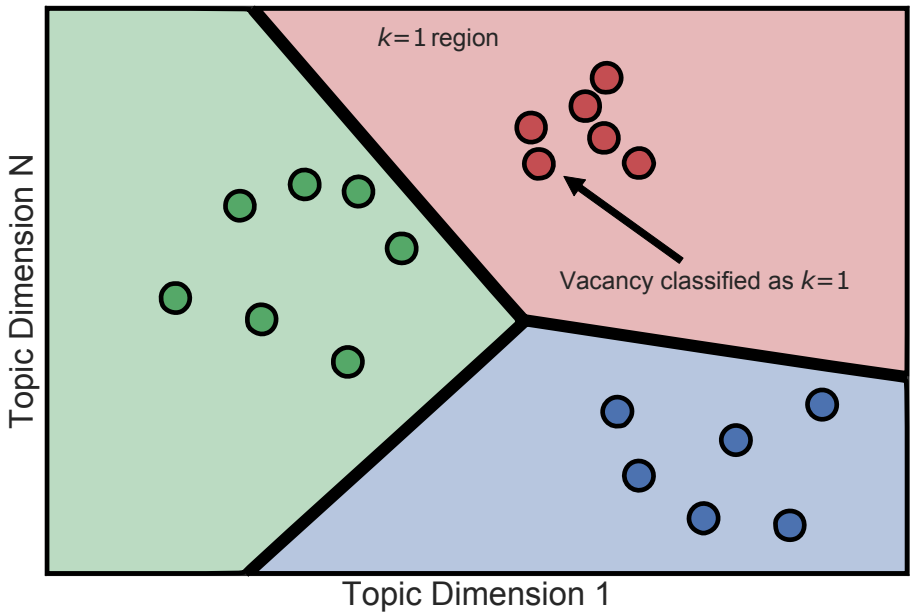

Figure 5: Schematic example of how vacancies, represented by points, are located in the topic vector space and assigned to clusters $k \in \{0, \ldots, K\}$. The first and Nth dimensions of the topic vector space are projected onto a 2D plane. Also shown are three example K-means areas, divided by classification boundaries. The K-means boundaries are for illustrative purposes only; see Lloyd (1982) for a full description of how K-means assigns points to each distinct k value.

4.1 Step 1: Cleaning documents

In the cleaning process, each document is created as a combination of the text of the job description and title with all punctuation removed. Stopwords, such as 'the', 'to', 'as', are removed, as are any digits or single letters. We also impose an upper and lower bound on the frequency of the number of occurrences of a word in each document. The maximum threshold, 5×10^5 occurrences across documents, is applied to remove very common words such as 'experience', 'work' and 'team', which provide little useful information about the job. The lower threshold is 10^5, and this removes words that do not appear commonly across job descriptions; some of these are mispelt words. This forms the cleaned corpus.

4.2 Step 2: Creating topics based on demand

In the second step, the topic model we use is Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2003) as implemented in the GENSIM Python package (Řehůřek and Sojka, 2010). This uses the online variational Bayes algorithm (Hoffman, Bach and Blei, 2010). The LDA model is fed a weighted random sample of 5

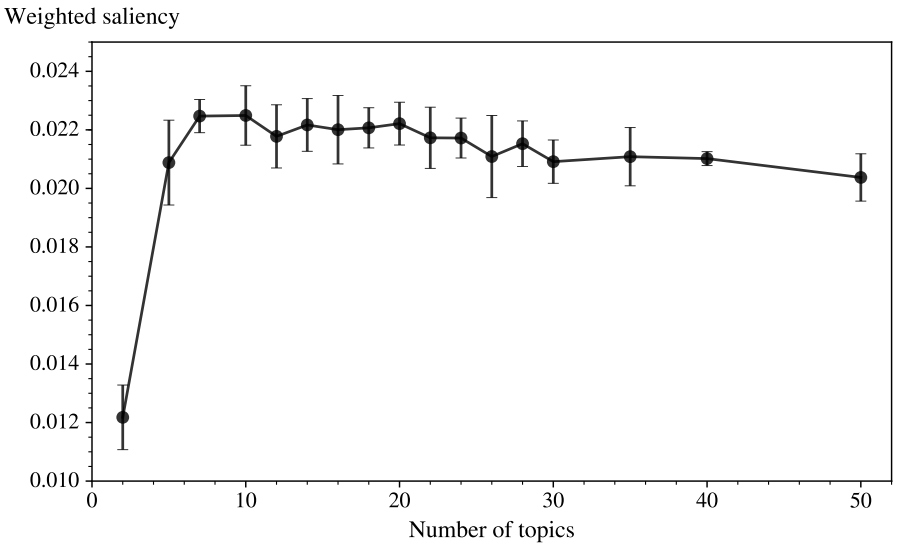

Figure 6: Weighted saliency score, defined in equation (2) and labelled ζ, is used to determine the optimal number of topics. The error bars are from 5 different Monte Carlo runs of the algorithm, each on a random sample of 10^6 vacancies from the dataset.

million job vacancies from the total dataset. Computational limitations necessitate the use of a sample. The reasoning behind using a random sample weighted with the weights from §3 is that this gives clusters representative of the labour market as a whole; if the weights had been omitted the topics would have been dominated by finer details from jobs which are very common in this dataset, such as in sales occupations. Once selected for inclusion, the LDA algorithm treats each individual vacancy in the same way.

In LDA, each document d is assumed to be a mixture of latent topics, and each topic is a distribution over words appearing in the corpus as a whole. Take w_{ld} to be the lth word in document d, $\boldsymbol{\theta}$ to be the distribution of topic weights over documents, and $\boldsymbol{\beta}$ to be the distribution of word weights over topics. LDA may be thought of as a probabilistic factorisation of the matrix of word counts into the matrix of topic weights $\boldsymbol{\theta}$ of dimension $D \times N$ and the dictionary of topics $\boldsymbol{\beta}$ of dimension $L \times N$ in which $P(w_{ld}) = \sum_n \theta_{dn}\beta_{ln}$. It is assumed that $\theta_d \sim \text{Dirichlet}(\kappa)$ for the distribution of topic weights over each document d, that $\beta_n \sim \text{Dirichlet}(\eta)$ for the distribution of word weights over each topic n, that topic weight $n_{ld} \sim \text{Categorical}(\theta_d)$ and that word $w_{ld} \sim \text{Categorical}(\beta_{n_{ld}})$. LDA is used to analyse documents via the posterior distributions of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and topic assignments n: that is, it finds $P(n, \theta, \beta | w, \kappa, \eta)$. The words w_{ld} are the only observables; there are priors for the parameters κ and η for which we use the GENSIM defaults. LDA tells us nothing about the number of topics to choose; that is N is an input of the algorithm rather than an output. In the limit of large N, there are as many topics as terms and, in the limit of very small N, topics will be very broad and difficult to interpret.

To determine an optimal value for N using the data, we use the 'weighted saliency' as our criterion function (Goldsmith-Pinkham, Hirtle and Lucca, 2016). Alternatives to this approach include maximising the coherence of topics (Röder, Both and Hinneburg, 2015) or maximising the topic stability (Greene, O'Callaghan and Cunningham, 2014). The approach we follow relies on earlier work which defines the 'saliency' of words within the corpus (Chuang, Manning and Heer, 2012) as

$$s(w|n) = P(w) \times d(w|n) \tag{1}$$

with $P(w)$ the probability of choosing a word at random from our corpus and $d(w|n)$ the distinctiveness of a word given a topic n. $d(w|n)$ is given by

$$d(w|n) = P(n|w) \ln \left(\frac{P(n|w)}{P(n)} \right)$$

Distinctiveness, $d(w|n)$, determines how likely a word is associated with topic n; in statistical mechanics terms it is the relative entropy of $P(n|w)$ with respect to $P(n)$, or equivalently the point-wise Kullback-Leibler divergence. $P(n|w)$ may be calculated using that $P(w|n) \equiv \beta$ and Bayes' theorem, i.e. $P(n|w) = P(w|n)P(n)/P(w)$. The distinctiveness drives the saliency measure in (1) by giving low weight to words which are very prevalent within the corpus, i.e. high $P(w)$, but which are associated with many topics. This allows words mentioned infrequently but associated with a single topic to be identified. $P(n)$ is the likelihood of topic n controlling for document length,

$$P(n) = \frac{\sum_{d=1}^{D} \theta_{n,d} L_d}{\sum_{n=1}^{N} \sum_{d=1}^{D} \theta_{n,d} L_d}$$

We use code from the pyLDAviz package to implement this (Sievert and Shirley, 2014). The weighted saliency (Goldsmith-Pinkham, Hirtle and Lucca, 2016) sums the saliency score of the top five words given each topic with weights given by a measure of the average load of topic n across documents. More rigorously, define

$$l_{i+1,n} := \{s(w_l|n) \in s(w|n)_i : s(w_l|n) \geq s(w_{l'}|n) \forall s(w_{l'}|n) \in s(w|n)_i\}$$

as the set of the top $i+1$ values of $s(w|n)$ for unique words, and the set $s(w|n)_{i+1} := s(w|n)_i \setminus l_{i+1,n}$ as

the set of $s(w|n)$ with those entries removed. Then the weighted saliency is given by

$$\zeta(N) = \sum_{n=1}^{N} \lambda_n \sum_{x \in l_{5,n}} x \tag{2}$$

The average load of topic n across documents is given by

$$\lambda_n = \frac{1}{D} \sum_{d=1}^{D} \theta_{n,d}$$

Maximising $\zeta(N)$ with respect to N determines a set of topics which both contains salient words, which help distinguish between topics via the term in salient words in equation (2), and which are important across documents, through the term in λ_n. Figure 6 shows ζ for our corpus, which does not demonstrate a clear and unambiguous peak. There is a plateau of high, statistically indistinguishable values of weighted saliency between $N = 8$ and $N = 20$. In the next step, the documents will be clustered based upon how similar the topics expressed in documents are. Noting that the decrease in ζ between $N = 8$ and $N = 20$ is within the error bars, and that the clustering is designed to produce an efficient representation of the data (discarding topic dimensions which do not carry useful information), we opt for the maximum number of topics on the plateau, $N = 20$. For comparison, official classifications use 90 3-digit SOC codes, 25 2-digit SOC codes, and 9 1-digit SOC codes, or NUTS codes of which there are 12, 40, and 174 at the 1, 2, and 3 character UK levels respectively. There are 21 SIC 'sections'.

Figure 7 shows the topics created by the LDA algorithm along with their topic weights over the entire corpus. Within each topic, the size of the words shown is given by β_n, the distribution of word weights over that topic. In order to protect the information in individual vacancies, only words which appear in several job titles are shown in the bottom right-hand panel. Some topics clearly contain words related to a particular occupation, for instance $n = 4$ contains 'teachers', 'school', and 'education'. $n = 14$ is clearly related to the preparation of food. Other topics are much less easy to interpret in terms of traditional jobs, for instance $n = 0$ contains 'project', 'planning', and 'operations' but do not seem to relate to a particular occupation or industry. Similarly, $n = 2$ seems to be about compliance and safety standards rather than a specific type of job. In runs with a larger number of topics (larger N), the marginal topics became increasingly uninterpretable. Eventually topics were added which were clearly not expressing demand.

We do not know *ex ante* whether our topics are the best grouping of different themes expressed in the corpus. Particularly, as N increases, the number of topics which are not useful in understanding distinct roles also increases. For example, at larger N, a diversity topic, which advertised firms' commitment to

Figure 7: The topics output by the Latent Dirichlet Allocation algorithm. Each word's size corresponds to its weighting within the topic, i.e. the words are scaled according to $\boldsymbol{\theta}$. Only words which appear in numerous vacancy descriptions are shown.

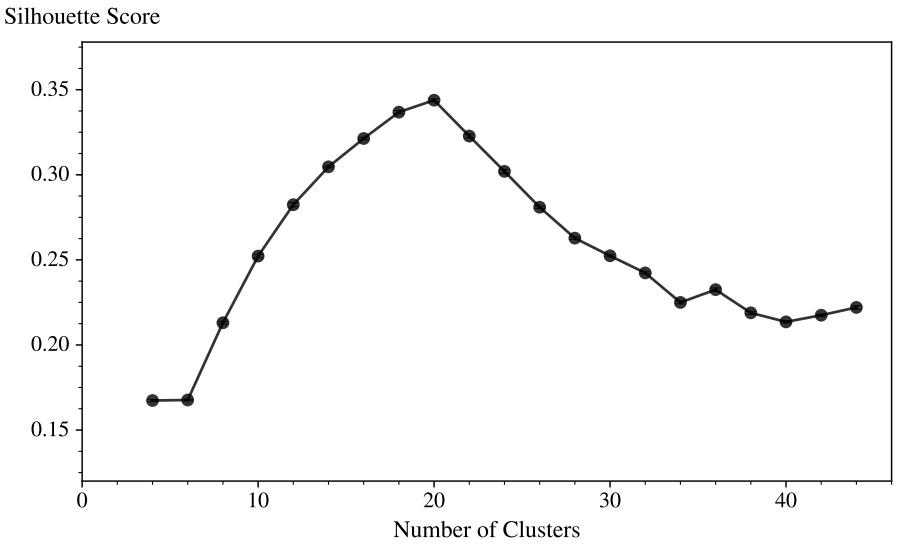

Figure 8: Mean silhouette score, $\bar{S}$, for a range of values of K. Equation (3) gives the definition of silhouette score for each total cluster number K and document d. The mean silhouette scores shown are from ten repeats using 5×10^4 randomly sampled vacancies in each repeat.

best practice in hiring but which was not useful in distinguishing between types of labour, began to appear.

4.3 Step 3: Clustering similar jobs in mutually exclusive groups

To retain only the useful information contained in topics, but discard that which says little about how vacancies might be (dis-)similar, we turn to the third step, using 'K-means' (Lloyd, 1982). Each vacancy is a vector of topics, $\boldsymbol{\theta}_d$, embedded within the vector space defined by the topics. K-means clusters documents (vacancies) which are close in this space into K mutually exclusive groups using the Euclidean norm to find a partition of the D documents into the $K < D$ clusters which minimises the within cluster sum of squares. K is an input into the K-means algorithm.

Following our strategy of letting the data speak, we use the silhouette score to determine K (Rousseeuw, 1987). This score is maximised by maximising the inter-cluster distance while minimising the intra-cluster distance; that is, documents representing vacancies in the topic vector space should be in regions that are both well-separated tightly defined. Let $\bar{a}$ be the mean distance between a document d and every other document within the same cluster, and let $\bar{b}$ be the mean distance between document d and every point in the next nearest cluster. Then,

$$S_{K,d} = \frac{\bar{b} - \bar{a}}{\max(\bar{a}, \bar{b})} \tag{3}$$

To choose K, we find the maximum of the mean silhouette score, $\bar{S}$ as a function of K. Note that the computation of this metric is intensive, with $\mathcal{O}(\max\{K\} \cdot d^2)$ operations. Given this, the average of ten

random samples of 1% of documents is taken. The mean silhouette score for a range of values of K is shown in Figure 8, and exhibits a distinct peak at $K = 20$.

The clustering process is a way of discarding topics which apply to all vacancies but which are not useful in discriminating between different parts of the labour market. As a robustness check, we verified that as N gets larger, a $K < N$ is chosen. Not knowing that $N = K$ *ex ante* in this case provides one justification for performing the clustering step. In calculations not shown, we also found that $\bar{S}$ is a function of N, giving further motivation for determining N through the weighted saliency measure before determining K.

5 Labour market demand from the bottom-up

Having determined $N = 20$ and $K = 20$ for the corpus, the vacancies as represented in the vector space defined by topics are grouped into clusters. Table 2 shows the top three words common to the job vacancies assigned to each cluster. Some roles are very clear and distinct amongst them, for instance $k = 18$ is the legal profession, $k = 6$ is software development, and $k = 19$ are sales assistants. This is remarkable given that at no point were these different roles explicitly signalled to the algorithm.

Job vacancies have local idiosyncrasies; the word 'class', appearing in $k = 11$ refers to the different classes of light or heavy goods vehicles licences available in the UK. With 'driver' as the other most common word, this suggests that $k = 11$ captures drivers of goods vehicles.

Some of the clusters are very directly related to the original topic dimensions. Cluster $k = 2$ is represented by 'teaching', 'school', and 'teacher' and strongly overlaps with words associated with topic $n = 4$. The specific characteristics of this cluster are shown in Figure 9, which shows, starting from the top left hand panel and proceeding clockwise, the most common words from across the corpus of this cluster's documents (including titles), the number of jobs within this cluster which have each 1-digit SOC code, the count of vacancies within this cluster by region per unit of labour force in that region, the top 10 ONS sectors (SIC) by count of vacancies within this cluster, the most common trigrams drawn from the job titles within this cluster, the top 10 occupations by 3-digit SOC code within this cluster, the count of permanent versus temporary jobs within this cluster, and the nominal offered salary per annum as a probability density function. Note that these jobs are almost completely concentrated in the 'education' sector, and most are in a single 3-digit SOC code. Even weighted by the number of people in the labour force in each region, the jobs are very highly concentrated in London. There are a substantial number of topics which have a strong overlap with a specific cluster because they directly relate to an area of labour market demand. $n = 14$ to $k = 7$, 'chef', 'food', and 'restaurant', is another.

Cluster	1st most common word	2nd most common word	3rd most common word
0	field	account	executive
1	consultant	commission	graduate
2	teaching	school	teacher
3	property	temporary	estate
4	accountant	accounts	finance
5	project	planning	projects
6	software	design	engineer
7	chef	food	restaurant
8	home	nurse	nursing
9	marketing	media	digital
10	production	maintenance	engineer
11	drivers	driver	class
12	worker	children	social
13	administrator	administration	telephone
14	hr	employee	payroll
15	telesales	account	executive
16	procedures	appropriate	safety
17	centre	insurance	advisor
18	firm	practice	legal
19	assistant	retail	store

Table 2: The top three words common to the job vacancies assigned to each cluster.

Cluster 2

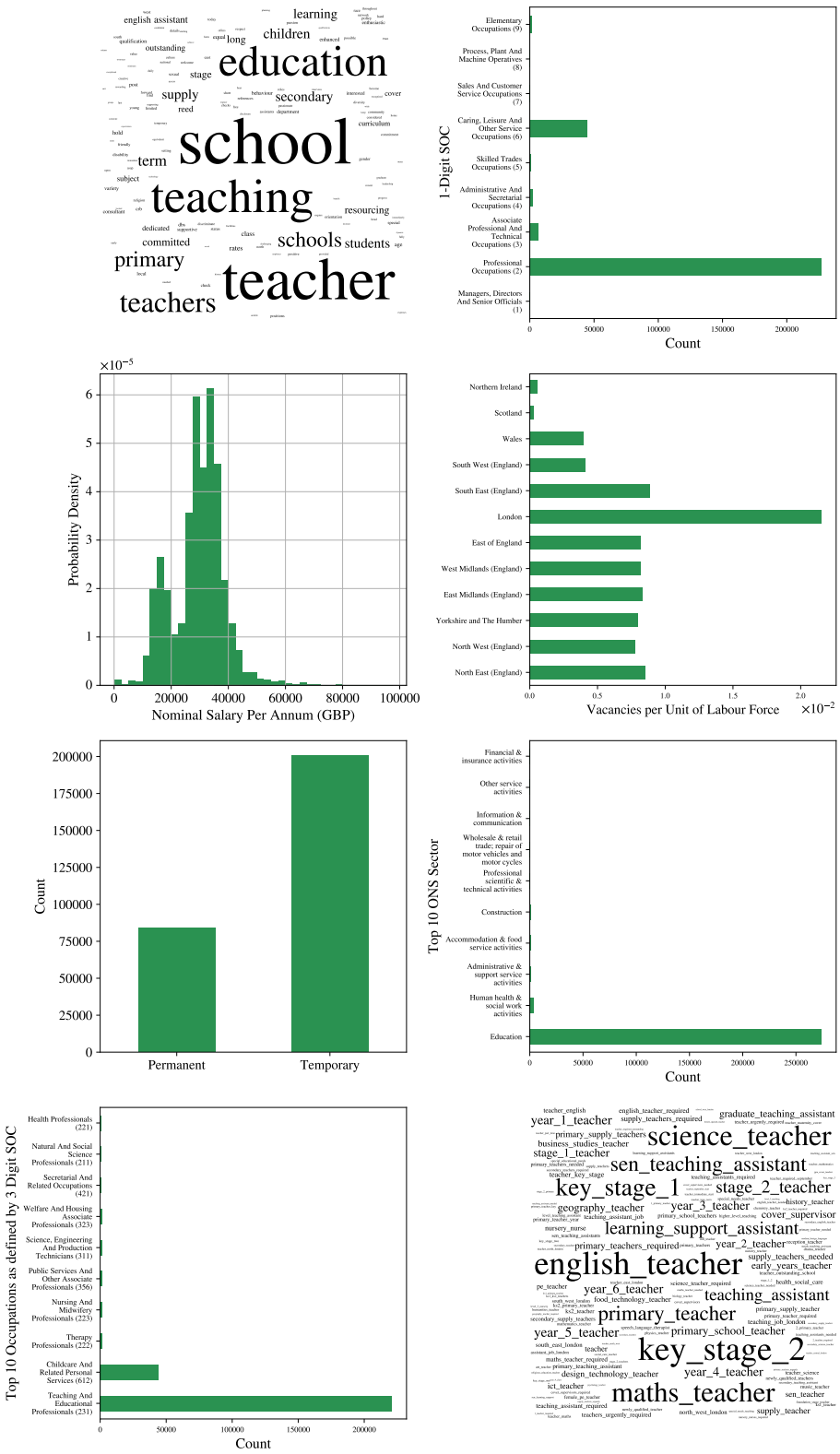

Figure 9: The properties of the vacancies which are sorted into cluster 2, which show a strong association with teaching. Descriptions of each panel are in the text.

Cluster 5

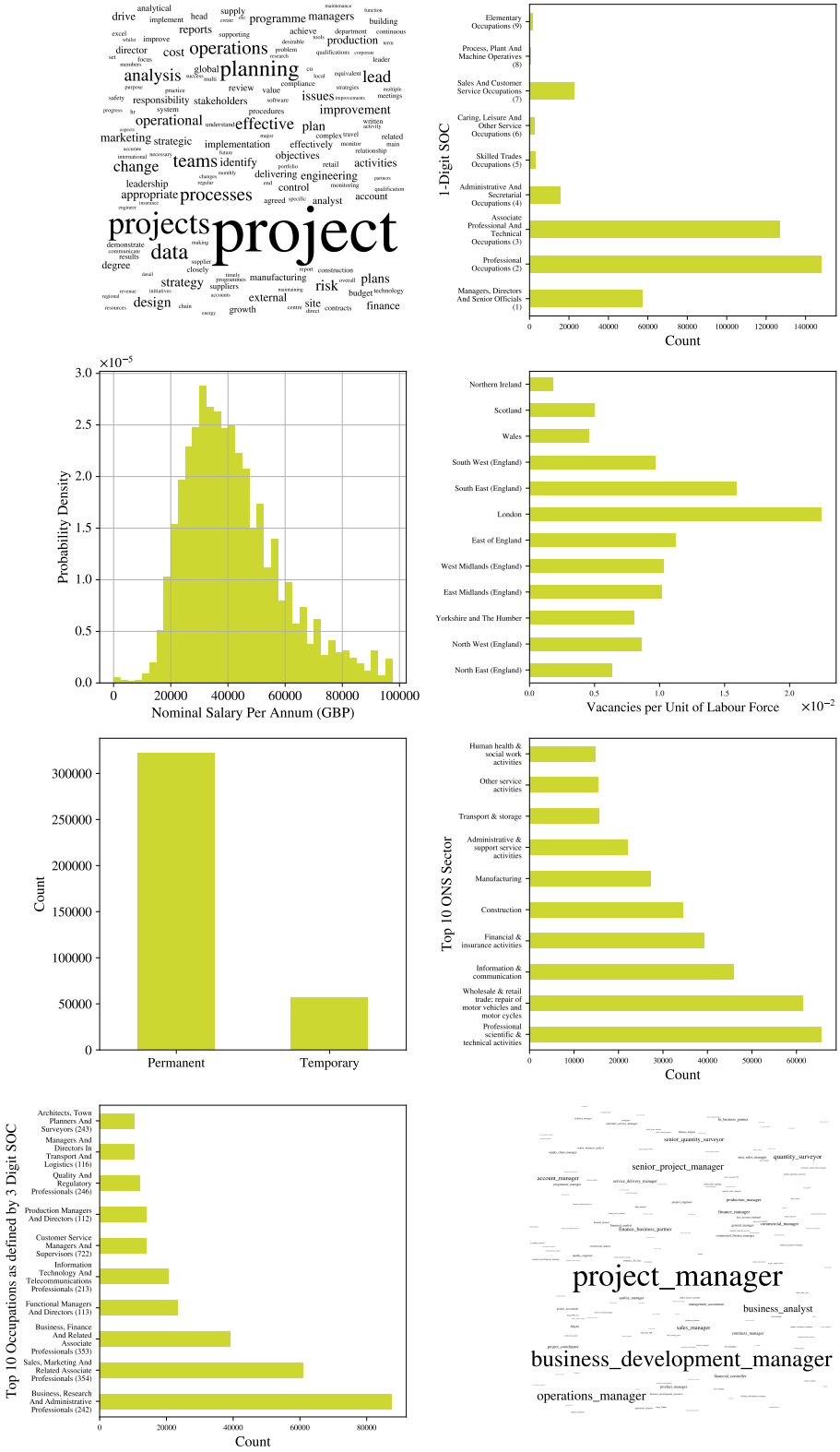

Figure 10: The properties of the vacancies which are sorted into cluster 5, which are found across many different SOC, NUTS, and SIC codes. Descriptions of each panel are in the text.

These traditional jobs are easy to classify and the extra utility of the bottom-up process is low; these jobs are already well captured by official classifications. However, that these topics are cleanly mapped into the relevant clusters gives confidence in the algorithm: the clusters are aggregating truly similar jobs rather than similar features of advertisements for different types of job. They are created without supervision.

What of the topics which were less obviously associated with a traditional role? Topics $n = 0$ and $n = 2$ are examples. For example, job vacancies with a strong score in topic $n = 0$ were most likely to be assigned to cluster $k = 5$. As shown in Figure 10, $k = 5$ is much less well described by existing classifications. It is comparatively strongly driven by n-grams (up to trigrams) such as 'business development manager' and 'project manager'. These types of job can be in a range of industries (SIC codes) or occupations (SOC codes) as shown in the sectoral and occupational breakdown panels in Figure 10. Cluster $k = 16$, which mainly draws on topic $n = 2$ ('safety', 'plan', 'risk', 'compliance'), is similarly split across SIC and SOC codes.

It is feasible that workers could move between official classifications in order to take up these types of job. This demonstrates the power of the bottom-up approach; it details roles which break the barriers between traditional labour market segments. Relatively recent changes in the nature of work in some industries are also reflected in the cluster-related words shown in Table 2; while marketing is not a new career, it appears in cluster $k = 9$ alongside 'digital'.

5.1 Step 4: Mapping remaining vacancies into clusters

For computational reasons, only a third of vacancies (sampled by the weights created in §3) were used to create the clusters. In order to apply these clusters to the other 2/3 of our data, we created a simplified version of the 20 dimensional space defined by the topics. The vector space was simplified by only using the rows of $\boldsymbol{\beta}$ corresponding to the top 200 words by weight for each topic. As might be anticipated from Zipf's law, there is a long tail of low weight words for each topic and using the top 200 words vastly simplifies the computational time required to turn vacancies into vectors in the topic space. On a training set of 80,000 vacancies labelled with clusters, we applied several supervised classification methods and checked their out-of-sample accuracy on 20,000 vacancies with known labels. The k-Nearest-Neighbours classification algorithm performed the best, and was subsequently trained on 10^6 labelled vacancies (test set of 10^5, out-of-sample accuracy 79%) before being used as the model to label all $\sim 10^7$ vacancies which were not initially assigned to a cluster.

Table 3: Correlation of selected cluster time series with official classification time series.

Cluster	Description	Compared to		Correlation			
				Vacancies		Employed	
		SOC name	SIC section	SOC	SIC	SOC	SIC
2	School, Teacher, Education	Teaching and Educational Professionals (231)	Education	0.940	0.873	0.978	0.927
7	Chef, Restaurant, Food	Food Preparation and Hospitality Trades (543)	Accommodation & food service activities	0.965	0.970	0.265 (0.824)	0.984
8	Nurse, Home, Nursing	Nursing and Midwifery Professionals (223)	Human health & social work activities	0.954	0.975	-0.224 (0.794)	0.904

5.2 Testing the demand-based vacancy clusters

If these clusters capture the demand for labour quantitatively, then the time series of the stocks of vacancies in clusters which clearly capture careers well described by existing classifications will be very strongly correlated to the time series of those same careers using the stock of vacancies as counted according to the official classifications. Example of such careers include teachers, nurses, and chefs. In Table 3 we present the correlation of the time series of these easily identified types of cluster (e.g. teachers) with their closest known SIC code (e.g. Education) and SOC code (e.g. Teaching and educational professionals). The correlations are shown under the 'Vacancies' heading. These correlations are calculated from the quarterly time series over the entire, fully labelled dataset of vacancies broken down by the relevant cluster group, SIC code, and SOC code. Vacancies labelled by SOC code are not available from the ONS and are drawn from Turrell et al. (2018). The cluster description column in Table 3 features the three most common words associated with that cluster. These vacancy cluster time series are very strongly correlated with the relevant vacancy time series using the official classifications.

To determine whether the clusters are informative for the description of the demand for labour more broadly, we performed ordinary least squares regressions of the logarithm of the real offered wage with dummies for each month, and then varied the categories used as explanatory variables. This again used the full dataset only excluding the 5% of entries on either end of the real wage distribution. As Table 4 shows with 'Yes' indicating which categories are used as explanatory variables, the bottom-up clusters had more power than the commonly used disaggregation of the labour market by sector, and also by region, but did not predict wages as strongly as occupation. All models have a joint F-test significance level of $p < 0.01$ as indicated by $***$. Note that there are

$$p = 40(\text{NUTS 2 regions}) + 25(\text{SOC 2 codes}) + 21(\text{SIC sections}) + 20(\text{clusters}) = 106$$

explanatory variables, but 10^7 observations so that Adjusted $R^2 \approx R^2$.

Table 4: Regressions of offered wage on classification fixed effects.

Dependent variable: ln (real offered wage)	I	II	III	IV	V	VI
Region (NUTS2)	Yes				Yes	Yes
Sector (SIC section)		Yes			Yes	Yes
Occupation (SOC2)			Yes		Yes	Yes
Cluster				**Yes**		**Yes**
F-test significance	***	***	***	***	***	***
Observations	1.4×10^7	1.4×10^7	1.4×10^7	1.4×10^7	1.4×10^7	1.4×10^7
R^2	0.055	0.128	0.304	0.213	0.385	0.429
Adjusted R^2	0.055	0.128	0.304	0.213	0.385	0.429
AIC	1.2×10^7	1.2×10^7	8.3×10^6	1.1×10^7	6.5×10^6	5.5×10^6

We speculate that occupations are a better predictor of offered wages because our clusters 'collapse' the different levels of the same career into a single sub-market, thus combining a range of wage levels. For instance, for teaching this could be classroom assistant, teacher, and head teacher. These would be given different occupational codes, but would all sit in the same cluster – and descriptions of them in the vacancies are likely to be similar. This could suggest a role for a 'seniority' dimension of the classification to discriminate rungs of the same type of job. As regression VI shows compared to V, the clusters also provide extra power relative to the three other categories.

6 Labour market supply from the bottom-up

Thus far, the data which created these clusters (the Reed vacancy data) have also been used to evaluate their performance. Now we attempt to apply the same clusters to the supply side of the labour market (the labour force). Ideally, we would do this by using these clusters to predict the transitions of workers – that is, if these clusters truly capture the careers available to workers then the within cluster job-to-job transition rates should dominate the without cluster job-to-job transition rates. A formal model which did this would also include information on workers which is outside of the scope of our demand driven clusters, e.g. on the workers' current region, age, and level of education. With all of that information included, transition rates using the clusters as a predictor could be compared to transition rates based on sectoral or occupational classifications.

Although we attempted to obtain the data on hires which would allow these calculations to be performed, they are not feasible with the LFS in the form in which it is currently available. To create the hires data by cluster, it would be necessary to first match, record-by-record, the longitudinal and cross-sectional LFS. The matching is needed as the longitudinal LFS does not contain all of the variables necessary for the

machine learning algorithm to be run. We tried several methods of creating a unique match variable, but none yielded reasonable time series for the stocks or flows of employment or unemployment in the matched data. This prevented us from creating a cluster equivalent of the job-to-job flows of Figures 1 and 2. This would be an important check on the method of creating the clusters in future work, if the required data can be obtained.

Instead, we apply our cluster labels to the cross-sectional LFS alone, and test their performance via their explanatory power for wages, and in the degree to which the time series of similar clusters, occupations, and sectors correlate. We use the LFS data from 2008 to 2015 inclusive, close to the same time period as our vacancy data (which differs in that it also includes 2016).

Applying the natural market clusters to the labour force is constrained by the set of information which is shared both by our vacancy data and by the labour force survey data (the latter do not come with the job descriptions which we used to label the former). Let each cluster be labelled k, with $k \in \{0, \ldots, K-1\}$. The problem is to predict k for each i using $\vec{x}_i$ a vector of information on each individual in the cross-sectional LFS. We seek a function G such that

$$\vec{y} = G(X)$$

where $\vec{y}$ is a vector of values of $k \in \{0, \ldots, K-1\}$ for all $i \in I$. There is no job description in the survey data, so the approach used in §5 is not applicable. We solve this classification problem by using the data fields which exist in both the vacancy data and in the fields related to individuals in the LFS. We train the model on the data fields in the vacancy data for which we have a known cluster, and then apply the trained model to the same fields in the LFS. The information set on an individual in the LFS which overlaps with the information in each job vacancy is given by

$$\vec{x}'_i = (\text{SOC}, \text{NUTS}, \text{SIC}, \ln \text{real wage})_i \tag{4}$$

where these variables pertain to the job occupied by those employed and the job sought or job last occupied for those unemployed. The first three variables are discrete, with the fourth continuous and only relevant in the case of employed individuals. The SOC code is at the 3-digit level, the NUTS code at 1-digit, and the SIC codes are at the section level.

To build a model which will associate each individual i at each time t with a cluster k, we train a support vector classification algorithm[5] in estimating $\hat{G}$ on the problem $y_v = G(\vec{x}_v)$ with $\vec{x}_v$ the same information

[5]We tried several algorithms at smaller scales for this problem, with support vector classification consistently providing the best performance.

Table 5: Null (false) and valid (true) combinations representing more than 0.1% of classifications for employed individuals in the *Labour Force Survey*.

NUTS	SOC	SIC	ln(real wage)	% of all cases
FALSE	FALSE	FALSE	FALSE	53.5
FALSE	TRUE	TRUE	FALSE	0.19
TRUE	TRUE	FALSE	FALSE	5.00
TRUE	TRUE	FALSE	TRUE	1.65
TRUE	TRUE	TRUE	FALSE	29.9
TRUE	TRUE	TRUE	TRUE	9.63

set as in equation (4) but for a vacancy v. We use 10^6 of the originally labelled vacancies, holding out 20% of these as a test set.

Any difference between the vacancy data and the LFS could create problems for the accuracy of the trained support vector classification algorithm when applied to the LFS. There are some significant differences. Firstly, if there exist some combinations of valid and null entries across the dimensions of X in the vacancy data (on which it is trained) which are not present in the matched LFS data (on which it is deployed), the model is unlikely to label these combinations with high accuracy. The combinations of null and valid entries are just the most simple type of difference; if the two datasets have different distributions over the valid entries in each classification, and there are many possible combinations of these valid entries. As the LFS data represent supply and filled jobs, while the vacancy data represent demand, the expectation would be that these distributions are different in general. Due to the nature of machine learning, this issue may be a bigger factor for continuous variables than discrete ones. The model is trained on offered wages, which reflect the demand for labour, while the wages of those working are the equilibrium price paid for labour, and these are not necessarily the same, though we necessarily treat them as such for the purposes of the application of cluster labels to the LFS. Finally, if real wages change considerably across the period captured by the classification, then this may reduce the performance of the algorithm. However, average aggregate UK real wage growth has been close to zero for much of 2008 to 2016 inclusive.

We try to correct for the first problem in which some entries in $\vec{x}_i$ are null (i.e. missing) for some individuals i, but the same combination of null entries never occurs in $\vec{x}_v$ for any v. We do this by artificially flipping some of the valid entries in the vacancy data to be null. For the overlapping data columns in X, the vacancies are cleaner than the LFS in the sense that there are far fewer missing entries apart from 3.3% which have missing offered wages. The LFS data have many more null entries. Out of a possible 2^4 combinations of having categories with null entries or not, the LFS microdata we use exhibits 14 of them, of which 6 are substantial (see Table 5). Note that in Table 5, a considerable percentage of entries do not have any of the $\vec{x}_i$ data attached at all. In this case, a model trained on known cases with only

Table 6: Null (false) and valid (true) combinations of classifications in the re-balanced vacancy data.

NUTS	SOC	SIC	ln(real wage)	% of all cases
FALSE	FALSE	FALSE	FALSE	16.7
FALSE	TRUE	TRUE	FALSE	16.7
TRUE	TRUE	FALSE	FALSE	17.2
TRUE	TRUE	FALSE	TRUE	16.1
TRUE	TRUE	TRUE	FALSE	17.2
TRUE	TRUE	TRUE	TRUE	16.1

null values in x_v will most likely revert to picking a cluster with the probability implied by the frequency of appearance of that cluster in the vacancy data. This makes any application of the clusters as applied to the LFS likely to be noisy and error-prone.

We flip entries in the vacancy data used for training to ensure a balanced set in which the percentage of occurrences of each substantial case in Table 5 is set to be equal. That is, we flip some entries to null for the vacancy test and train data so that it has an almost equal representation of each combination of null and valid classifications in the LFS, as shown in Table 6. Training on a balanced set ensures that the trained model is not biased toward any particular case of the six highlighted in Table 5. Both the train and test vacancy sets are re-balanced in this way. With this configuration, the out-of-sample accuracy of the support vector classification on the vacancy test set was 42.0%.

There is no equivalent in- or out-of-sample test for the accuracy of the algorithm which applies the clusters to the LFS. As in §5, we test the correlations of cluster time series with official classification time series, but using the data on those who are employed, and we test the extent to which the cluster classifications explain real wages. Because of the noted problems with applying the clusters to the LFS, we prefer correlations of time series to comparisons of the levels. The 'Employed' column of Table 3 shows the correlation between the clusters representing teachers, chefs, and nurses with their closest equivalent SOC and SIC categories. As with the vacancy time series, the correlation is very strong for teaching (cluster 2). For the other two clusters, chefs (cluster 7) and nurses (cluster 8), the correlation with SIC sections is very good, above 0.9 in both cases. However, by SOC code, the correlations perform much less well. There is a positive correlation for cluster 7 and the 'Food preparation and hospitality trades' SOC code, but it is much smaller. For cluster 8, the correlation with 'Nursing and midwifery professionals' is negative. There may be reasons for the poorer performance by SOC code. As noted, 53.5% of all cases do not have *any* information, but the fraction of invalid entries is actually worse for SIC than for SOC, as shown in Table 5. The analysis is carried out at the 3-digit SOC level, introducing a higher level of noise and difficulty for the algorithm due to the 90 categories versus just 21 for SIC sections. The SOC classification was revised in 2010, and

Table 7: Regressions of accepted wage on classification fixed effects.

Dependent variable: ln (real wage)	I	II	III	IV	V	VI
Region (NUTS1)	Yes				Yes	Yes
Sector (SIC section)		Yes			Yes	Yes
Occupation (SOC2)			Yes		Yes	Yes
Cluster				**Yes**		**Yes**
F-test significance	***	***	***	***	***	***
Observations	3.4×10^5	3.4×10^5	3.4×10^5	3.4×10^5	3.4×10^5	3.4×10^5
R^2	0.196	0.260	0.445	0.355	0.474	0.496
Adjusted R^2	0.196	0.260	0.445	0.355	0.474	0.496
AIC	5.6×10^5	5.3×10^5	4.3×10^5	4.8×10^5	4.2×10^5	4.0×10^5

our data from before this time have to be mapped modally into the new version of the SOC standard. This may be having an effect on the apparent performance of the clustering. The correlation of these time series from Q1 2011 onwards is shown in brackets by the two employed SOC correlation entries for clusters 7 and 8, and these indicate a better match for this post-classification change time period. Indeed, 223 did not exist in the SOC2000 standard and 543 has a significant number of non-zero diagonal entries in the best available fractional mapping from the SOC2000 standard to the SOC2010 standard.

Using the cluster-labelled LFS data, we run regressions with, separately, fixed effects for region, sector, occupation, and cluster, as well as all of these together. As in §5, we perform ordinary least squares regressions of the logarithm of the real wage with dummies for each month. The results are shown in Table 7. All regressions have additional controls for quarter, age, age^2, year, gender, and job type. Again, the 5% on the ends of the real wage distribution is removed. As before, $n \gg p$ so that Adjusted $R^2 \approx R^2$. The clusters have a higher R^2 than the regressions on region or sector, and a lower value than the regression for occupation.

There is an important difference between Model IV of Table 7 and Model IV of Table 4; in the latter, the clusters do not rely on any information about wages but do have explanatory power for wages. The relationship is not so clean in Model IV of Table 7, as the machine learned function $G(\vec{x}_i)$, which assigns an individual represented by $\vec{x}_i' = (\text{SOC}, \text{NUTS}, \text{SIC}, \ln \text{real wage})_i$ to a cluster, uses information on real wages. This makes the interpretation of the non-zero R^2 value in Model IV of Table 7 different too; it demonstrates that G finds real wage information useful in assigning individuals to clusters which were initially created without any wage information. Although the interpretation differs to that of Table 4, the non-zero R^2 is still a strong indication that the clusters carry useful information about accepted real wages.

There is another subtlety in the regression table, in Model VI. If the function $G(\vec{x}_i)$ were linear, then

regressing on clusters in addition to region, sector, and occupation would not add any explanatory power. This is because variation due to cluster would be entirely accounted for by x_i. However, support vector regression, which determines our estimate of the function G, $\hat{G}$, is a non-linear machine learning method and as such the cluster fixed effect can increase the value of R^2 beyond the other three classifications.

Overall, Table 7 confirms the results of Table 4; the bottom-up clusters have more explanatory power than region or sector but less than occupation, and they provide a small amount of additional power in combination with the other three types of category. Note that this table necessarily only includes employed individuals who disclosed their wage. All models have a joint F-test significance level of $p < 0.01$ as indicated by $***$. We speculate that, just as for offered wage, the clusters are limited in their explanatory power for accepted wages relative to occupations because they collapse different levels of the same career.

7 Conclusion

Using the text in the job descriptions of job vacancies is a promising way to better understand different groupings in the labour market, and one which both sidesteps and complements the usual top-down classifications. We have shown that machine learning tools provide empirical methods to choose both the type and level of disaggregation used for analysis. Analysis of this kind could help to inform future taxonomies of jobs and gives an indication of what the 'true' structure of the job market is.

As applied to job vacancies, the clusters which are created align well to traditional roles where appropriate but also suggest types of sub-market which are not obviously present in current classification schemes based on region, sector, or occupation alone. The clusters have explanatory power for the offered wages associated with job vacancies. Though applying the same clusters to the labour force is challenging due to a lack of appropriate information, we have also shown that the clusters contain some information relating to accepted wages, and that the correlation between the clusters which capture clearly defined roles, such as teacher, and the equivalent time series using official classifications is high.

The extent to which the derived clusters can currently be applied to the labour force is inhibited by the availability of data which overlaps with vacancy data. Future improvements in the labour force data could change this, and allow for a richer analysis of whether this demand-driven methodology is more successful at describing job-to-job flows than other classifications.

Given the expected rapid disruption to the demand for tasks, and therefore occupations, from automation, this methodology for a data driven taxonomy of jobs could be useful for statistical agencies and labour market economists. The clustering approach could provide a useful method for understanding sub-markets

in, for instance, vacancies that exist within a wider economic area spanning several countries that each have their own distinct statistical classifications.

There are many potential extensions and modifications to this work. Our topic model, which informs our labour market segments, is based upon the text in job descriptions alone but structural topic models would allow other information to be included in the creation of the topics, for instance on offered wages or using time as an input variable. The wage regressions undertaken suggest that adding a second dimension to our clusters which would capture the seniority of a position within a market segment could boost the explanatory power of our bottom-up method for wages. It would be in principle possible to do this using text analysis techniques applied to job descriptions.

While the empirical possibilities have been made clear, there is not as yet any theory which might explain, endogenously, what determines the number of sub-markets. Our approach has implicitly assumed that the labour market segments we find are mutually exclusive, but some workers will always transition across different classification codes in any mutually exclusive classification. This suggests that a more sophisticated future approach might see jobs assigned to a fuzzy set of sub-markets rather than being bivalently assigned to clusters.

More broadly, combining naturally occurring, semi-unstructured data and highly structured survey data is likely to be a recurring challenge. There is typically little control over the information set covered by naturally occurring data, but if the information gathered by surveys can be more closely aligned with the information types found in naturally occurring data, the benefits of combining both could be realised more often.

References

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

Barnichon, Regis, and Andrew Figura. 2015. "Labor market heterogeneity and the aggregate matching function." *American Economic Journal: Macroeconomics*, 7(4): 222–249.

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.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research*, 3(Jan): 993–1022.

Chuang, Jason, Christopher D Manning, and Jeffrey Heer. 2012. "Termite: Visualization techniques for assessing textual topic models." 74–77, ACM.

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.

Goldsmith-Pinkham, Paul, Beverly Hirtle, and David Lucca. 2016. "Parsing the content of bank supervision." Federal Reserve Bank of New York Staff Reports 770.

Greene, Derek, Derek O'Callaghan, and Padraig Cunningham. 2014. "How many topics? Stability analysis for topic models." 498–513, Springer.

Grinis, Inna. 2017. "The STEM Requirements of 'Non-STEM' Jobs: Evidence from UK Online Vacancy Postings and Implications for Skills & Knowledge Shortages." London School of Economics SRC Discussion Paper 69.

Hershbein, Brad, and Lisa B Kahn. 2016. "Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings." National Bureau of Economic Research.

Hoberg, Gerard, and Gordon Phillips. 2016. "Text-based network industries and endogenous product differentiation." *Journal of Political Economy*, 124(5): 1423–1465.

Hoffman, Matthew, Francis R Bach, and David M Blei. 2010. "Online learning for latent dirichlet allocation." 856–864.

Jackman, Richard, and Stephen Roper. 1987. "Structural unemployment." *Oxford bulletin of economics and statistics*, 49(1): 9–36.

Lloyd, Stuart. 1982. "Least squares quantization in PCM." *IEEE transactions on information theory*, 28(2): 129–137.

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

Nathan, Max, and Anna Rosso. 2015. "Mapping digital businesses with big data: Some early findings from the UK." *Research Policy*, 44(9): 1714–1733.

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.

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.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. "Inference of population structure using multilocus genotype data." *Genetics*, 155(2): 945–959.

Řehůřek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." 45–50. Valletta, Malta:ELRA. http://is.muni.cz/publication/884893/en.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. "Exploring the space of topic coherence measures." 399–408, ACM.

Rousseeuw, Peter J. 1987. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics*, 20: 53–65.

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)*.

Sievert, Carson, and Kenneth E Shirley. 2014. "LDAvis: A method for visualizing and interpreting topics." 63–70.

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 No. 737*.