BANK OF ENGLAND

Staff Working Paper No. 737

Using job vacancies to understand the effects of labour market mismatch on UK output and productivity

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

July 2018

BANK OF ENGLAND

Staff Working Paper No. 737

Using job vacancies to understand the effects of labour market mismatch on UK output and productivity

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

Abstract

Mismatch in the labour market has been implicated as a driver of the UK's productivity 'puzzle', the phenomenon describing how the growth rate and level of UK productivity have fallen behind their respective pre-Great Financial Crisis trends. Using a new dataset of around 15 million job adverts originally posted online, we examine the extent to which eliminating occupational or regional mismatch would have boosted productivity and output growth in the UK in the post-crisis period. To show how aggregate labour market data hide important heterogeneity, we map the naturally occurring vacancy data into official occupational classifications using a novel application of text analysis. The effects of mismatch on aggregate UK productivity and output are driven by dispersion in regional or occupational productivity, tightness, and matching efficiency. We find, contrary to previous work, that unwinding occupational mismatch would have had a weak effect on growth in the post-crisis period. However, unwinding regional mismatch would have substantially boosted output and productivity relative to their realised paths, bringing them in line with their pre-crisis trends.

Key words: Vacancies, matching, mismatch.

JEL classification: E240, C550, J63.

(1) Bank of England. Email: arthur.turrell@bankofengland.co.uk (corresponding author)
(2) Bank of England. Email: bradley.speigner@bankofengland.co.uk
(3) Nesta. Email: jyldyz.djumalieva@nesta.org.uk
(4) Bank of England. Email: david.copple@bankofengland.co.uk
(5) Bank of England. Email: james.thurgood@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, Emmet Cassidy, Matthew Corder, Rodrigo Guimaraes, Frances Hill, Tomas Key, Graham Logan, Michaela Morris, Michael Osbourne, Kate Reinold, Paul Robinson, Ben Sole, and Vincent Sterk for their comments. We would especially like to thank William Abel for suggestions 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

Since the Great Financial Crisis, UK productivity and output have fallen significantly and mysteriously behind their pre-2008 trend in both levels and growth rates (Haldane, 2017), as shown in Figure 1. This paper studies the effects of labour market mismatch on output and productivity in the post-crisis period, and asks whether eliminating mismatch would have allowed them to continue on their pre-crisis trend paths. Barriers which prevent worker mobility between regions and occupations can lead to misallocation of labour, and hence depress aggregate productivity. We investigate the importance of such barriers using a novel dataset containing detailed information on more than 15 million job vacancies posted online by firms via a recruitment agency. Building on the framework of Şahin et al. (2014), we use these data to quantify how much aggregate productivity and output would rise if mismatch by region or by occupation were eliminated. The most notable contributions relative to the existing literature are the labelling of online vacancies according to existing statistical classifications, an evaluation of matching function parameters on an entirely new dataset, and our results demonstrating the importance of regional heterogeneity in the labour market for productivity and output growth.

There are three main channels we consider for mismatch to affect productivity and output. The first is determined by how vacancies and the unemployed are differently distributed across the different regional and occupation sub-markets, or market segments, within the overall labour market. The second and third depend on the heterogeneity of productivity and of matching efficiency (the 'productivity' of the search-and-match process) across regional and occupational sub-markets.

We make a major innovation in using naturally occurring[1] vacancy data posted online to create time series on vacancies by occupation and region according to the same classifications as existing statistics on the labour market, for instance on unemployment. These data are not available through other means, such as surveys. Data labelling is achieved using a novel algorithm which applies occupational classifications to each individual vacancy based on the text of the job description.[2]

By occupation, we find in our counter-factual simulations for the UK economy that, in contrast to Patterson et al. (2016), mismatch has played only a minor role in weak output and productivity growth since the Great Financial Crisis. The period we study extends beyond that featured in Patterson et al. (2016) and we show that the mismatch effect on productivity disappeared completely after the end of 2012. Our results are robust to the level of disaggregation by occupation.

[1] As opposed to survey data collected for the express purpose of constructing statistics on job vacancies, these data consist of job advertisements posted by real firms looking to hire workers

[2] See `http://github.com/aeturrell/occupationcoder` for the computer code.

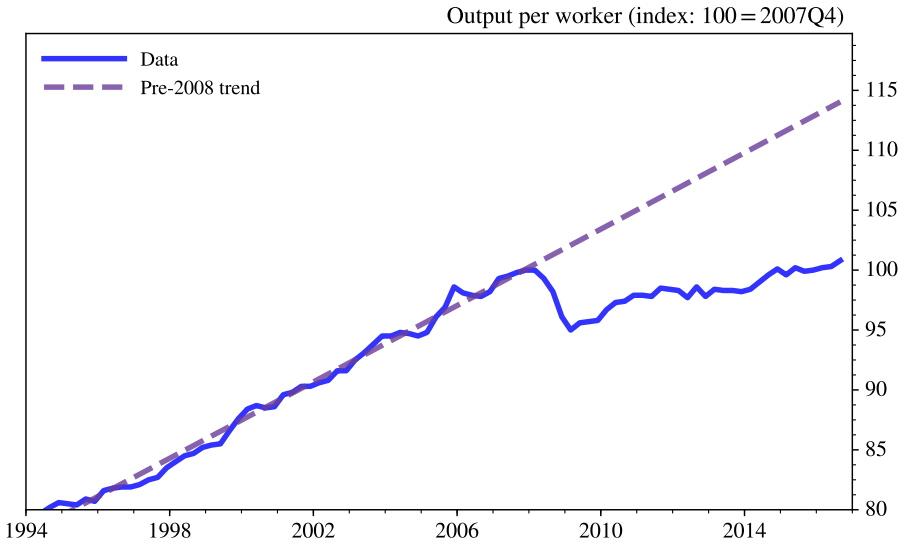


Figure 1: The aggregate output per worker in the UK (seasonally adjusted). Trend lines are fit using data from 2004Q1–2007Q4. Source: ONS, Author calculations.

Relative to the existing literature, we contribute a new analysis of mismatch by region, performing counter-factual simulations of employment, output, and productivity by this disaggregation. Consistent with recent studies on the US, we find very strong effects due to regional heterogeneity. Our main result is that resolving regional mismatch between 2008 and 2015 would have seen the UK at a significantly higher level for both output and productivity than in the realised case; indeed, if this had happened at the start of the period we consider, it would have completely offset the UK's productivity puzzle.

More broadly, we demonstrate that labour market heterogeneity by both occupation and region matter for output growth. Public policies to either lower the barriers to movement across regions, or to equalise productivity, tightness, and rate of job-finding across regions, would reduce the potential for mismatch to affect output. The differences in the rate of job-finding across parts of the labour market that we find mean that the aggregate outflow from unemployment to employment in the UK will change if the distribution of jobs across occupations or regions changes. This could happen if there were changes to the demand for tasks due to, for example, automation or changes to the UK's trading relationships.

The structure of the paper is as follows: §2 sets out previous literature relevant to our question, §3 describes the online job vacancies data and the algorithm which assigns them to official statistical classifications, §4 describes the search and matching theory we use, and §5 presents our results. Results are split into the estimation of the matching function (§5.1), the estimation of matching efficiency and productivity by occupation and region (§5.2), and counter-factual simulations of productivity and output for both occupation (§5.3) and region (§5.4). §6 concludes.

2 Literature

We use the search and matching theory of the labour market in our analysis (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). The basic matching function theory has been extended to take account of workers' search efficiency or search intensity, as in Barnichon and Figura (2015) and Pizzinelli and Speigner (2017). On the other side of the market, both Davis, Faberman and Haltiwanger (2013) and Gavazza, Mongey and Violante (2016) consider extensions which add recruitment intensity. We provide estimates of the matching efficiency at both aggregate and disaggregate levels using our new dataset. These estimates have implications for the shape of the Beveridge curve, a comprehensive review of which may be found in Elsby, Michaels and Ratner (2015).

Our paper is closest to the work of Şahin et al. (2014) and Patterson et al. (2016) who use counter-factual simulations of a social planner's optimal allocation of the unemployed amongst sub-markets to show that mismatch can cause sub-optimal productivity and output growth. This is because barriers between sub-markets prevent job-seekers from finding the most productive jobs. Smith (2012) uses a similar framework to examine unemployment dynamics and, using UK JobCentre Plus data, estimates that around half of the rise in UK unemployment during the crisis was due to mismatch. Patterson et al. (2016) creates a counter-factual path for output based upon a social planner's optimisation which minimises the extent of mismatch by occupation in the labour market. They find that the difference between the counter-factual and realised paths for aggregate output and productivity explain a significant fraction of the UK productivity puzzle between 2007 and 2012, up to two-thirds. A substantial boost in aggregate output is also shown to be due to occupational mismatch.

One of our contributions relative to Patterson et al. (2016) is an analysis of mismatch disaggregated by UK regions. Patterson et al. (2016) looked at regional mismatch by travel-to-work-areas and the regions of England in an early draft of their paper but found it to have had an insignificant effect on unemployment dynamics. The importance of regional heterogeneity in labour markets has recently been highlighted by results suggesting that misallocation lowered aggregate growth in the US by more than 50% between 1964 and 2009 (Hsieh and Moretti, 2017). There are reasons to think that regional mismatch may be important for the UK too; it has similarly constrained housing supply in some areas, large regional variation in house

price to income ratios, and areas of high labour market tightness which are correlated with areas of high productivity. The JobCentre Plus data used in Patterson et al. (2016) and Smith (2012) has also been used to look at regional mismatch; Smith (2012) shows that regional mismatch directly contributed around 2 percentage points to the UK unemployment rate following the Great Financial Crisis. Manning and Petrongolo (2017) examine search and match using highly disaggregated regional JobCentre Plus data over 2004–2006 to show that the attractiveness of jobs to applicants decays strongly with distance.

Our paper adds to a small but growing literature on the analysis of text in job vacancies. Marinescu and Wolthoff (2016) uses 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. One of our contributions is to develop a method, and computer code, to use job title and job description text from vacancies posted online to classify large numbers of vacancies into official categories such as the UK Standard Occupation Classification (SOC) codes. Until recently, the methods which existed to do this were proprietary, limited in the number of searches, or did not make use of the job description field.[3] Our method matches up well to the other approaches with which we have compared it, making use of techniques from machine learning, including the term frequency-inverse document frequency weighting described in Bholat et al. (2015), and fuzzy matching. Our algorithm could be adapted and applied to the SOC classifications of other countries or regions.

We show how eliminating the level of mismatch in the labour market might reduce the impact of the UK's productivity puzzle, the trend-breaking behaviour of both the level and growth rate of productivity (Barnett et al., 2014*b*; Bryson and Forth, 2015). A range of mismeasurement or output revision issues have been offered as explanations for the puzzle including the way intangibles such as research and development are measured (Haskel et al., 2015), and erroneous pre-crisis measurements of the productivity of the finance sector. Of a 16 percentage point puzzle in level terms in 2013Q4, Barnett et al. (2014*a*) suggest that mismeasurement could explain 4 percentage points, while another 6 to 9 could be accounted for by investment in intangibles, high rates of firm survival and impaired resource allocation. Credit constraints on firms following the crisis have also been explored but Riley, Rosazza-Bondibene and Young (2014) did not find strong evidence that this had reduced productivity growth. Goodridge, Haskel and Wallis (2014) point at total factor productivity growth rather than labour or capital productivity being the problem, particularly

[3]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*).

in the oil and gas, and financial services, sectors.

Given that the productivity puzzle is largely a UK phenomenon, some explanations focus more on the structure of the UK labour market. One such explanation is that productivity is pro-cyclical as a consequence of 'labour hoarding', in which firms hold on to workers in the expectation that demand will grow in the near future (Martin and Rowthorn, 2012). Pessoa and Van Reenen (2014) and Blundell, Crawford and Jin (2014) point to wage flexibility in the UK as a cause: the fall in the price of labour coupled with the rise in the cost of capital has meant a fall in the capital to labour ratio and an associated fall in labour productivity. Explanations which include a strong cyclical element struggle to explain the long duration of below trend productivity growth.

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*, the *Annual Survey of Hours and Earnings* (ASHE), and regional and sectoral productivity measures. Our measures of the number of people transitioning from unemployment to employment come from the quarterly longitudinal LFS, while the counts of those currently unemployed and employed come from the cross-sectional LFS. We take the stock variables from the cross-sectional LFS due to data quality issues with the stock variables in the longitudinal LFS.[4]

We use the per worker measure of productivity based on the chained-volume measure of gross value added, G, and the employment counts in the LFS. While data are available on G for Standard Industrial Classification (SIC) codes and for UK Nomenclature of Territorial Units for Statistics (NUTS) codes, they are not available for occupations. To construct productivity by occupation (with occupation labelled here by μ), we use

$$z_{\mu t} = \frac{G_{\mu t}}{E_{\mu t}}; \qquad G_{\mu t} = \sum_{i}^{I} \frac{E_{i\mu t}}{E_{it}} G_{it} \tag{1}$$

so that the value-added by occupation μ is the weighted sum of the value-added of its constituent industries, labelled i, with the weights given by the fraction of employment, E, of μ accounted for by i.

The vacancies data we use consists 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

[4]We use the ONS mapping from Standard Industrial Classification (SIC) 2003 to SIC 2007 to make LFS entries consistently labelled by SIC 2007 code. For SOC, we use fractional mappings from SOC 2000 to SOC 2010 on counts to obtain consistently labelled entries. For transitions, such as 'unemployed' to 'employed', we use the modal mappings from SOC 2000 to SOC 2010. NUTS 2010 is used throughout.

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.

Our work is unusual compared to the recent literature in that it uses data from a job advertisement and employee recruitment firm (a recruiter), Reed, rather than from an aggregator or from a survey. There are two different kinds of website which post job advertisements. 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.[5] 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.

A feature of 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 clearly erroneous. As an example of the latter, there are offered wages which appear too low (as they are not compliant with the minimum wage law) or unrealistically high. We discard these.

The most similar data to ours are the UK Jobcentre Plus statistics, used in Patterson et al. (2016), Smith (2012), and Manning and Petrongolo (2017). These data were collected via UK government offices but 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 JobCentre Plus data consist of vacancies aimed predominantly at those on unemployment benefit and as such are not representative of all vacancies. Patterson et al. (2016) find that the data over represent some sectors. These data were not included the ONS' labour market statistics releases between 2005 and their discontinuation because of concerns over their appropriateness as a labour market indicator (Bentley, 2005).

In the US, the most similar datasets to the vacancy survey and our data are the Job Openings and Labor Turnover Survey and the Conference Board Help Wanted Online series respectively.[6]

Similarly to the JobCentre Plus data, our data are from a website and so do not cover all job vacancies. The raw series accounts for around 40% of UK vacancies annually (see Fig. 2). 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

[5]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).

[6]As explained by Cajner, Ratner et al. (2016), there have been significant discrepancies between the stock of vacancies implied by these two US series which may be caused by changes in the price charged to employers to post online job vacancies. However, there is mixed evidence of whether this is an issue for online job adverts data more generally; Hershbein and Kahn (2016) find that it does not matter much using 'Burning Glass' data.

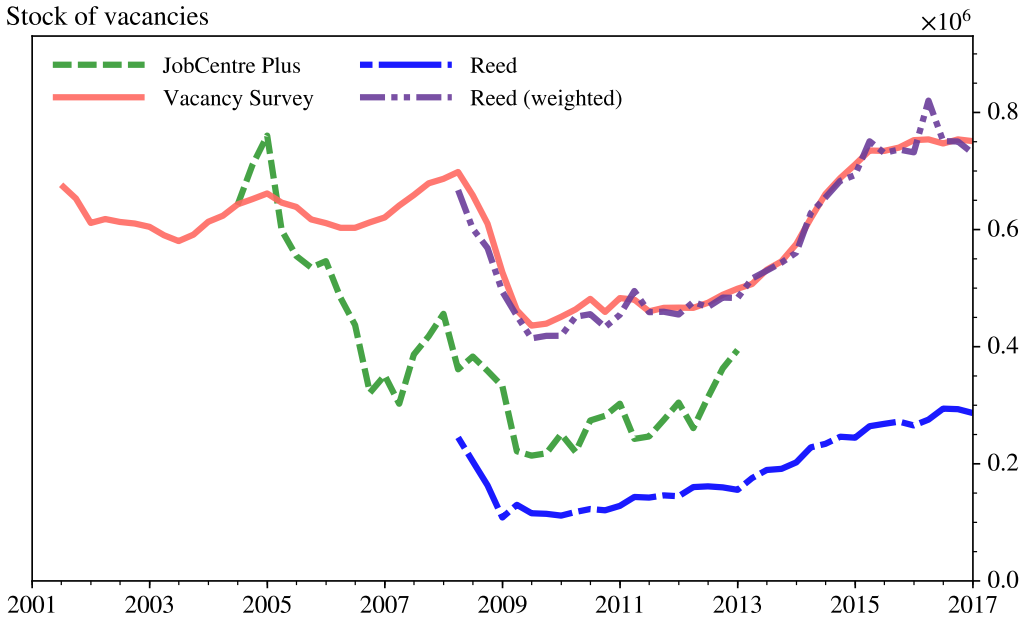


Figure 2: The aggregate stock of vacancies from three data sources. Source: Reed, ONS, National Online Manpower Information System (NOMIS), Author calculations

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. Given this potential for discrepancies between our data and the official UK data, we weight the stock of vacancies from Reed using the ONS Vacancy Survey which contains a cross-sectional category – sector – which is also available in our vacancy data. We use the weighted data to access vacancies according to classifications which do not currently exist in the official statistics.

The Vacancy Survey offers a cross-section either by size of firm or by sector. There is no breakdown by region or occupation, the two schema of interest for this analysis. Therefore our data add a new perspective on vacancies, but can only be compared to the Vacancy Survey on aggregate, or by sector. In the Reed dataset, each individual vacancy is a flow, with entries removed after being on the site for 6 weeks. 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:

$$V_m = V_{m-1} + \sum_{d \in m} \left(\dot{V}_d - \dot{V}_{d-6\times7} \right)$$

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

	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

The aggregate time series of the Vacancy Survey, raw Reed stock of vacancies, and JobCentre Plus Statistics are shown in Fig. 2. Neither of the latter have the same overall level of vacancies as the official statistic. 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. Note that what matters is that the trends in job vacancies as posted online match the trends in job vacancies more broadly.

To overcome some of the bias in the Reed vacancy data, and to ensure that it matches the Vacancy Survey in aggregate more closely, we weight it using the monthly sectoral disaggregation of the Vacancy Survey. The idiosyncratic Reed sectors are mapped into single letter Standard Industrial Classification (SIC) sections. The stock weight of an individual vacancy v in sector i and month m is given by

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

with $V^{\text{vs}}_{i,m}$ the monthly stock of vacancies by sector according to the Vacancy Survey, and $V_{i,m}$ the stock of vacancies from the Reed data. This mechanically ensures that the aggregate time series are very similar and improves the correlation between them. As the weights are applied at the individual vacancy level, aggregations of the re-weighted data by other schema (region or occupation) should more accurately reflect the true levels of UK vacancies; not just those posted online. Weighting the Reed data reduces bias but increases variance, as can be seen in Fig. 2. In subsequent sections, we use the weighted data.

In order to make the data useful in understanding mismatch, it must be cleaned and transformed into classifications which also exist in official data. The online job vacancy data from Reed do not have any official classifications attached. We predominantly use three official classifications for our data; Standard Industrial Classification (SIC) sections (used to reweight the data), up to 3-digit Standard Occupational Classification (SOC), and up to 3 character UK Nomenclature of Territorial Units for Statistics (NUTS) code. Here, we discuss NUTS and SIC mappings; the more involved mapping to SOC code is described in §3.1.

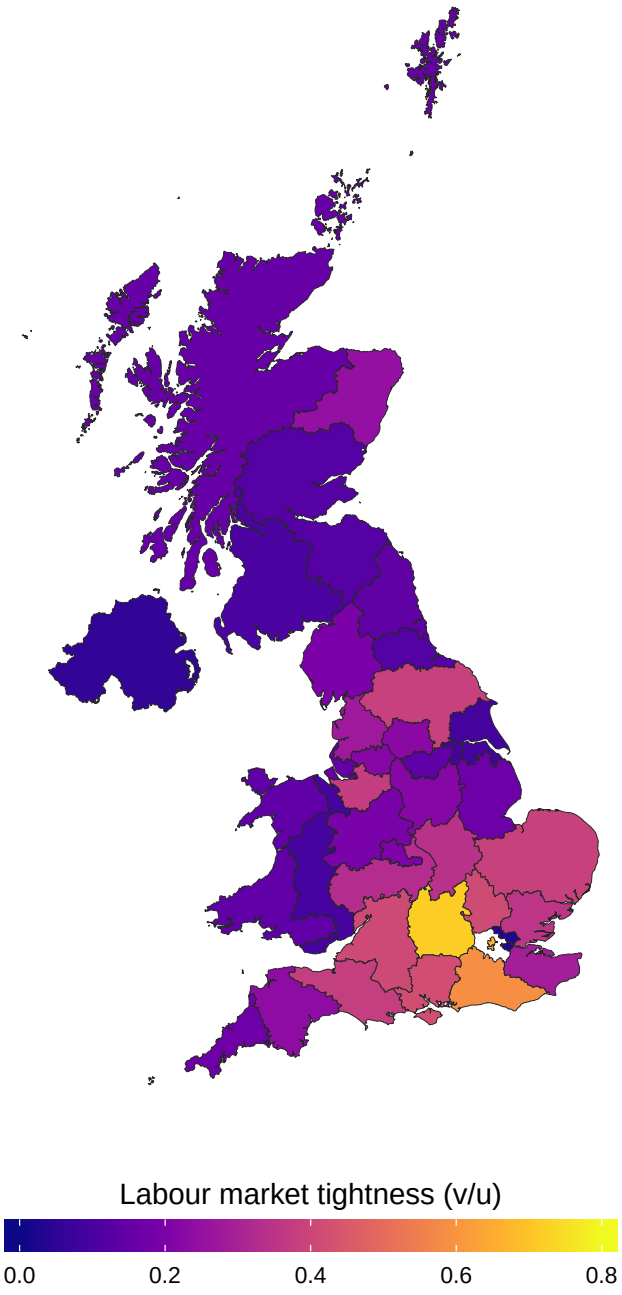


Figure 3: Map of mean UK labour market tightness, $\theta = \left(\frac{v}{u}\right)$, by 2-character NUTS code over the period 2008Q1–2016Q4. Some of the NUTS classifications are different in the ONS data relative to the NUTS2010 standard (EUROSTAT., 2011). This causes problems for London (UKI). We 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.

An idiosyncratic Reed sectoral classification has similarities to 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 the aggregated areas represented by each NUTS code. 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. Due to the way that Reed assign latitude and longitude, analysis is performed only on vacancies grouped up to 2-character codes. Figure 3 shows the labour market tightness, $\theta = \frac{v}{u}$, by 2-character NUTS code. Due to the lack of granular regional classifications in the *Labour Force Survey*, our regional counter-factuals are only carried out at the 1-character UK NUTS level.

3.1 Matching job vacancy text to occupational classifications

The mapping to ONS SOC code is less trivial and one of our major contributions. The data do not have any occupational categories which can be easily mapped into SOC, and so we developed an algorithm to take information from other data fields and use that to determine the SOC code. The SOC system is hierarchical, like NUTS, and has four levels.

We use the job title 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. The procedure is described in Algorithm 1. Additional details, including of the dictionaries used, are in Appendix A.

We use term frequency-inverse document frequency (tf-idf) to represent all possible SOC codes with a matrix with dimension $T \times D$ where t is a term from the text associated with a SOC code. d is a document (in this case a text string corresponding to each 3-digit SOC code) with D documents (the number of unique 3-digit SOC codes) in total. The specific form of tf-idf is

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

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 $\hat{\vec{v}}_d = \frac{\vec{v}_d}{||v_d||}$.

New documents which are not in the corpus can be represented in the vector space by calculating their tf-idf score in the space of terms from the original corpus. For instance, a job vacancy description plus title may be expressed as $\hat{\vec{v}}'$. In our algorithm, each document corresponds to a different 3-digit SOC code and

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 is the cosine similarity score referred to in Algorithm 1. Of the top five matches, 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).

Algorithm 1 SOC coding

```
Create term frequency-inverse document frequency (tf-idf) matrix based on known job titles
for vacancy in vacancies do
procedure CLEAN VACANCY
    clean title with known words dictionary & abbreviation expansion dictionary
    clean description with abbreviation expansion dictionary
    clean Reed sector field
    combine title, description and Reed sector into new field
procedure MATCH TO SOC
    if exact match to known job titles then return SOC code
    else
        express vacancy text as vector in tf-idf space of known job titles
        identify five 3-digit SOC codes closest to vacancy by cosine similarity
        if vacancy title blank then return top SOC code by cosine similarity
        else
            select best fuzzy match between vacancy title and known job titles drawn from top five 3-digit SOC codes
```

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 for SOC codes here (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. 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 try to evaluate the performance of Algorithm 1, we asked the ONS to code a randomly chosen subset of our data using their proprietary algorithm. This algorithm does not use the job description text and is

Table 2: Summary of evaluation of SOC coding algorithm against ONS coding (3-digit level). Source: ONS, Author calculations

	Manually assigned code	Assignment by proprietary algorithm
Sample size	330	67,900
Accuracy	76%	91%

designed for clean job titles, of the kind that appear as 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") and which are confusing to 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. As a consequence, of the 2×10^5 example vacancies submitted, only 68×10^3 were able to be confidently coded by the ONS algorithm. This large disparity is likely to be focused on the most difficult job titles. Also, unlike our method and because of their different domains of application, the ONS approach does not use the job description text – a much richer source of information on the occupation. Our method of coding agreed with the ONS method in 91% of the samples that the ONS algorithm could confidently code.

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 Algorithm 1, with only a match at the 3-digit level being counted as accurate. The results from both are shown in Table 2. The results are similar to the levels of agreement seen between human coders and 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. Finally, we 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

going to more aggregate classifications loses some of the rich heterogeneity which we find in the data. Our estimates of the stocks of vacancies are also likely to be biased by the fact that our data are reweighted by sector, rather than by occupation. If accurate information on the stocks of vacancies were available by occupation there would be no need to carry out this process, so this presents the best endeavour given data availability.

Since Algorithm 1 was created, we became aware of similar approaches which have recently been developed. 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 two respects: our model is trained on the official job category descriptions, while theirs is trained on 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. 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. 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. Future work could usefully compare or combine all of these methods on the same SOC matching problem.

4 Theoretical framework

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. These barriers could reflect segment-specific capital, relocation costs, or preferences. Mismatch can take the form of vacancies and unemployment which co-exist within a sub-market and which combine to form hires only slowly, or vacancies and unemployment which are in different sub-markets and therefore cannot combine at all. Mismatch lowers the overall efficiency of the labour market. At the aggregate level, let vacancy and unemployment rates at time t be given by $v_t = \frac{V_t}{L_t}$ and $u_t = \frac{U_t}{L_t}$ respectively, with L_t the size of the labour force. Denote hires out of unemployment from period $t-1$ to period t as h_t. Let there be I labour market segments indexed by $i \in (1, 2, \ldots, I)$ such that the rate of unemployment and of vacancies in each sub-market is given by u_{it} and v_{it} respectively at time t. Aggregate labour market tightness is given by $\theta_t = \frac{V_t}{U_t} = \frac{v_t}{u_t}$, while the tightness of segment i is given by $\theta_{it} = \frac{v_{it}}{u_{it}}$.

If there is mismatch across sub-markets, unemployed workers search in a segment with too few vacancies for them, while vacancies exist in a segment with too few searching workers to take them up. The period

during, and directly following, a shock in which employee-employer pairs are broken differently across industries, or in which there is a rapid change in the kind of tasks demanded of labour, can exacerbate mismatch. The Great Financial Crisis is an example of the former kind of shock; as Smith (2012) shows, mismatch across both sectors increased sharply for the UK during the crisis. Rapidly changing distributions of vacancies across sectors or occupations could be induced by automation, as analysed by Acemoglu and Restrepo (2017) and Autor, Levy and Murnane (2003). Similarly, sudden changes in trading relationships which changed the composition of tasks demanded in production could increase mismatch. Mismatch reduces the aggregate rate of job finding, $\frac{h}{U}$, and job filling, $\frac{h}{V}$, for a given aggregate level of U and V.

Due to the concavity of the matching function (discussed below), dispersion in labour market tightness across sub-markets lowers the rate at which job seekers are matched to vacancies on aggregate. This is one channel for mismatch to affect output. A basic version of the Diamond-Mortensen-Pissarides (Diamond, 1982; Mortensen and Pissarides, 1994) search-and-match theory of the labour market reveals other channels. For brevity, the time index is implicit in many of the following definitions. 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 are given by h_i. Define output per worker in segment i by z_i, which we take to be exogenous. Hires based upon the theoretical matching function and a segment-specific matching efficiency are given by

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

According to this matching function theory, there are several channels through which mismatch can affect output in addition to unbalanced θ_i across i. Mismatch-based output effects can also be caused by heterogeneity in ϕ_i or z_i, or both. For example, if all unemployment and vacancies are in a market with low z_i, then the output from any hires will be lower than if they were in a sub-market j with $z_j > z_i$. Differences in ϕ cause the flow of new hires to differ given the levels of U and V within the sub-market because the rate of job finding satisfies $\frac{h_i}{U_i} \propto \phi_i$. The search-and-match theory behind these channels has been extended

to account for factors such as recruitment intensity (Davis, Faberman and Haltiwanger, 2013) and search intensity (Pizzinelli and Speigner, 2017) but we do not consider these effects here.

To estimate the effect of mismatch, we use the framework developed by Şahin et al. (2014) and used by Patterson et al. (2016) and Smith (2012). Given I market segments, this model gives a counter-factual, optimal path for output by imagining a social planner that assigns the unemployed to different market segments. Let $\boldsymbol{\Xi}_t$ be a set of parameters representing known constants in discrete time labelled by t such that

$$\boldsymbol{\Xi}_t = (\boldsymbol{z}_t, \boldsymbol{V}_t, \boldsymbol{\phi}_t, \xi_t)$$

Each vector is of length I and they represent productivity, the stock of vacancies, matching efficiency, and job destruction rate across sub-markets respectively. Let u_t be unemployment and $\boldsymbol{e}_t$ be the vector of employment by market segment. In each time period, the social planner operates as follows; firstly, $\boldsymbol{\Xi}_t$ are observed. Then $\boldsymbol{e}_t$ is given, determining u_t, the aggregate unemployment rate. Next, unemployed workers searching in segment i, labelled in percentage terms by u_i, are matched so that there are $h_i = \phi_i M(U_i, V_i)$ new hires in segment i within period t. Production occurs in the existing matches given by $\boldsymbol{e}_t$ and the new hires given by $\boldsymbol{h}_t$, though new hires are assumed to be a fraction $\gamma < 1$ less productive than existing ones. Following previous work, we set $\gamma = 2/3$. Job destruction occurs, determining the next period's employment $\boldsymbol{e}_{t+1}$. At this point, the social planner chooses the division of searchers for the next period, that is they choose $\boldsymbol{u}_t$. Once determined, L_{t+1} (next period labour force size) and the next period stock of employed, $e_{t+1} = \sum_i e_{i,t+1}$, together set the next period stock of unemployed workers u_{t+1}.

The planner chooses $\boldsymbol{u}_t$ to maximise output, a problem which is given by

$$V(u_t, \boldsymbol{e}_t; \boldsymbol{\Xi}_t) = \max_{\{u_{i,t}\}} \left\{ \sum_i z_{i,t}(e_{i,t} + \gamma h_{i,t}) - \xi u_t + \beta \mathbb{E}\left[V(u_{t+1}, \boldsymbol{e}_{t+1}; \boldsymbol{\Xi}_{t+1})\right] \right\}$$

such that $\sum_i u_{i,t} \leq u_t$ where $e_{i,t+1} = (1 - \xi_t)(e_{i,t} + h_{i,t})$ and $u_{t+1} = L_{t+1} - \sum_i e_{i,t+1}$. The full solution for $\boldsymbol{u}_t$ is given in Appendix B, and is an increasing function of $\boldsymbol{z}$, $\boldsymbol{\phi}$, and $\boldsymbol{\theta}$.

This allocation allows for the construction of counter-factual output at each time period t via

$$Y_t^* = \sum_i^I z_{it} e_{it}^* + y_t^* \tag{3}$$

where $e_{it}^* = (1 - \xi_{t-1})e_{i,t-1}^* + h_{it}(v_{it}, u_{it}^*)$. Output per worker (our measure of productivity) in the realised and counter-factual cases is given by Y_t/e_t and Y_t^*/e_t^* respectively.

5 Results

5.1 Estimation of the matching function

In this section, we examine the implications of our vacancies data for empirical estimates of the matching function, the theoretical foundations of which were described in §4. 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.

There is a well-developed empirical literature on the estimation of the matching function spanning a number of datasets. There is widespread accord on the fundamental properties of the matching function, including constant returns to scale (Petrongolo and Pissarides, 2001). Parameter estimates do nevertheless vary to some degree across data samples. We add our own evidence to the wider literature on disaggregated empirical matching functions. Two early contributions are Coles and Smith (1996) and Bennett and Pinto (1994), who employ regional data to estimate matching functions and find that there is not a large bias introduced by aggregation.

The novel angle which our dataset permits is an examination of how the levels of vacancies and unemployment affect transitions into employment at the occupational level over the period of the Great Recession. Patterson et al. (2016) is a notable recent example of a study which also estimates an occupational-level matching function using administrative data on job vacancies, though from a different source, and it is instructive to compare results.

One advantage of our algorithmic approach to assigning conventional SOC codes to online job adverts is that it enables us to make use of longitudinal survey data on flows from unemployment to employment as the dependent variable in the matching regressions. This is worth emphasising since the choice of the dependent variable can influence the matching function parameter estimates (Petrongolo and Pissarides, 2001). Rather than use transitions from the LFS directly as the dependent variable, Patterson et al. (2016) use the average of vacancy off-flows (from JobCentre Plus data) and claimant unemployment off-flows (from the National Online Manpower Information Service, or NOMIS) as the dependent variable, which are a proxy for labour market transitions. The longitudinal LFS is a weighted and representative sample of employment flows at the UK national level though it is at quarterly frequency and is known to have non-response and response error biases (Jenkins and Chandler, 2010).

Table 3: Matching function parameter estimates. All results are significant at the 1% level. Source: Reed, ONS, Author calculations.

	1-digit SOC	2-digit SOC	3-digit SOC	1-digit NUTS	Aggregate data
Elasticity parameter (α)					
Point estimate (least squares)	.396	.427	.431	.254	.367
Standard error	.075	.050	.037	.020	.030
Point estimate (IV)	.392	.442	.371	.275	.350
Standard error	.073	.061	.048	.026	.031
Cross-sections	9	25	90	12	-
Observations	324	852	2120	423	35

We adopt a matching regression specification which assumes the segmented labour market discussed in §4, with segments indexed by i. There is no interaction among the different sub-markets. Gross flows from unemployment to employment for the ith occupation at time t are given by the matching function

$$h_{i,t} = \phi_i V_{i,t-1}^{\alpha} U_{i,t-1}^{1-\alpha}$$

The baseline empirical matching regression is

$$\ln\left(\frac{h_{i,t}}{U_{i,t-1}}\right) = \ln\phi_i + \alpha \ln\left(\frac{V_{i,t-1}}{U_{i,t-1}}\right) + \epsilon_{i,t} + d_t \tag{4}$$

where ϕ_i capture cross-section fixed effects and $d_t \in \{d_2, d_3, d_4\}$ represents a set of three quarterly dummy variables. As a baseline assumption, as in Patterson et al. (2016), we impose a constant elasticity α across all occupational sub-markets and over time, as well as constant returns to scale in matching. Ordinary least squares with cross-section clustered standard errors is then applied to the pooled data. In the next section we report our baseline results from the estimation of equation (4) and we also discuss a few simple extensions to the baseline model.

We report matching function estimates for data disaggregated to the 1-, 2- and 3-digit SOC level and the 1-digit NUTS level. In addition, we also report results from a matching regression estimated on the aggregated data. The regression results using the pooled data suggest fairly consistent results for matching elasticities centred around 0.4, which is in the middle of the range described as 'plausible' by Petrongolo and Pissarides (2001). These estimates are close to the range of 0.45 to 0.56 reported by Patterson et al. (2016) and very close to the OLS and GMM estimates of Barnichon and Figura (2015) based on US data. The matching elasticity appears to be increasing somewhat in the level of disaggregation, with the 3-digit elasticity a little higher than the 1-digit estimate, but the effect is not large. The point estimate based on

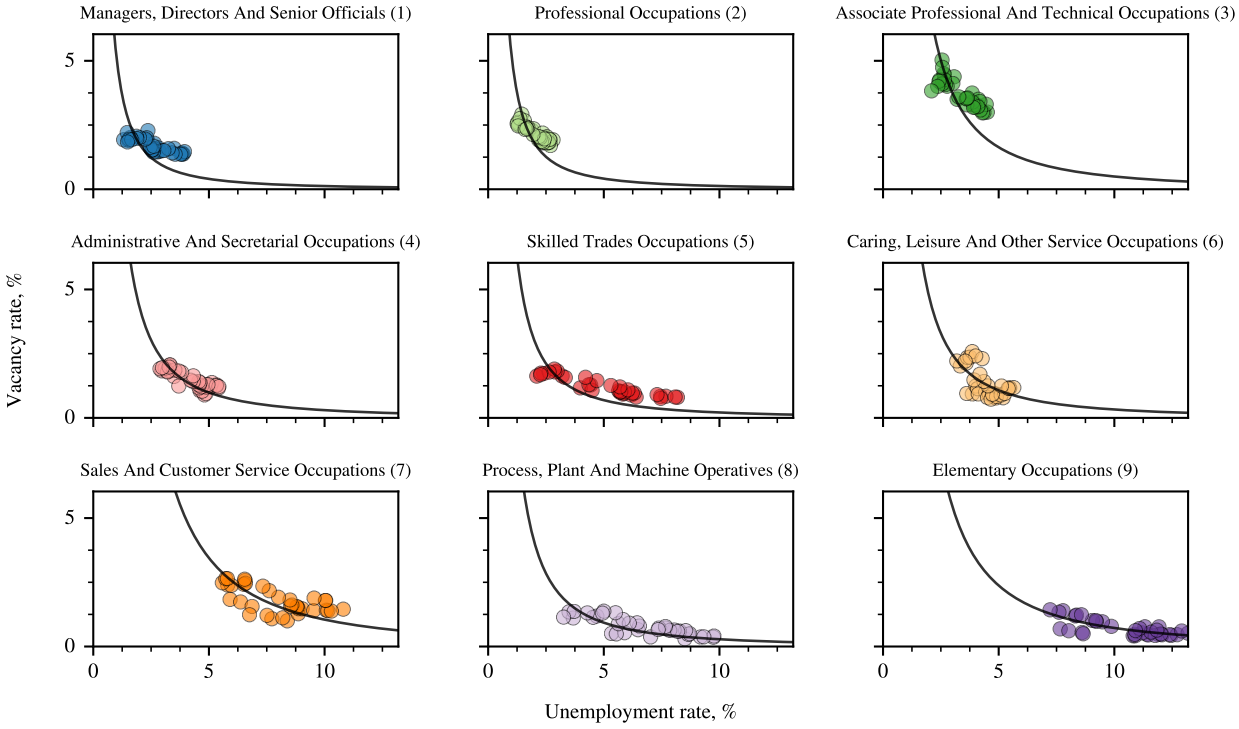


Figure 4: Beveridge curves (lines) using estimates of the parameters in equation (4) and data (points) in u-v space for each 1-digit SOC code at quarterly frequency. Source: Reed, ONS, Author calculations

the NUTS-1 sample is lower, around 0.25, suggesting that congestion on the demand side of the market does not rise as quickly when regional labour markets tighten as when occupational job markets tighten.

It has long been recognised that matching regressions are likely to be affected by simultaneity bias (Blanchard and Diamond, 1989), and Borowczyk-Martins, Jolivet and Postel-Vinay (2013) report evidence that this bias can be quantitatively significant. In recognition of this, Table 3 also reports parameter estimates from an instrumental variables regression in which we instrument for labour market tightness with a single lag. While the point estimate for the 3-digit SOC level is moderately lower under the instrumental variables specification, in general the results do not differ substantially from the ordinary least squares estimates.

Figure 4 plots Beveridge curves and quarterly u-v points for each 1-digit SOC code. These curves use the matching efficiency estimates plotted in §5.2. 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.[7] 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; generally, the more highly skilled the occupation, the less volatile its movement along the Beveridge curve. The driver of the variation relative to the curve is also different; 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.

The data strongly suggest that steady state unemployment rates differ by occupation (see equation (5) of Appendix B). Beveridge curves by region present a similarly heterogeneous picture (see Appendix C). According to both types of segmentation, there is behaviour consistent with genuinely distinct sub-markets. The different sub-markets imply different rates of outflow from unemployment to employment. Compositional changes in jobs will have an effect on the aggregate rate of outflow, affecting the amount of slack in the economy.

5.2 Productivity and matching efficiency

As noted in §4, dispersion in productivity and matching efficiency exacerbate the effects of mismatch. There is significant heterogeneity in productivity across a range of official classifications. For instance, there is strong evidence of wide heterogeneity in firm-level productivity performance in Field and Franklin

[7] 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.

(2013), both within sectors and across sectors (Barnett et al., 2014*b*; Broadbent, 2012; Haldane, 2017); see Appendix C for more evidence of productivity dispersion across official classifications.

Figure 5 shows that matching efficiency estimates, found according to the regressions in §5.1, have a wide distribution across occupations at the 1-digit SOC code level. There is no consistent pattern of matching efficiency across occupational categories, but on average higher skill sub-markets do appear to have lower levels of matching efficiency. Indeed, the highest measured matching efficiency level is for elementary occupations, and the lowest is for managers, directors and senior officials. This finding appears to be consistent with Şahin et al. (2014) who find that matching efficiency is lowest for management, professional and related occupations. This is consistent with the average number of years of education of workers in lower numbered 1-digit SOC codes being higher, implying a higher level of specialisation.

Figure 5 also shows estimates of productivity by occupation. The lower matching efficiency occurs in the most productive occupations, meaning that it takes longer to match unemployed individuals with more productive jobs. Matching efficiency broadly falls with increasing skill level, as opposed to productivity which broadly rises.

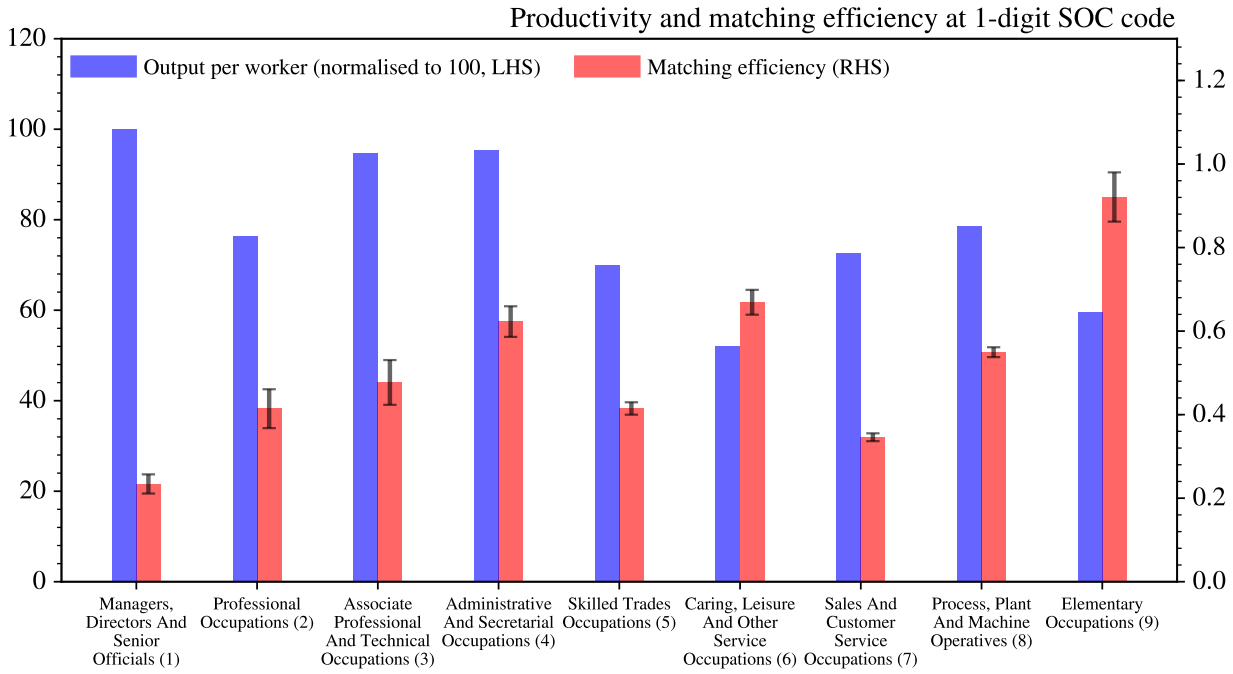


Figure 5: Estimates of productivity (left-hand y-axis) and of the matching efficiency (right-hand y-axis) by 1-digit SOC code. Standard errors are shown for the estimates of the matching efficiency. Source: ONS, Reed, Author calculations

Figure 6 shows the breakdown of estimates of productivity and matching efficiencies by 1-character UK NUTS code as estimated in §5.1. As with occupation, productivity and matching efficiency have some

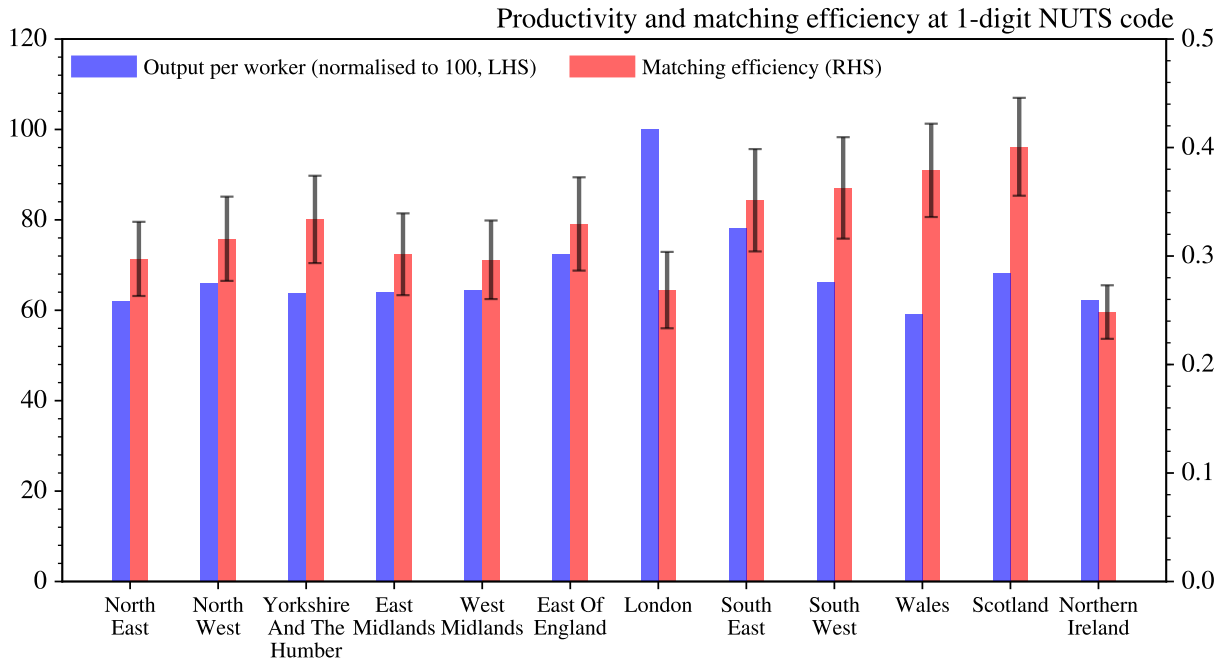


Figure 6: Estimates of productivity (left-hand y-axis) and of the matching efficiency (right-hand y-axis) by 1-character NUTS code. Standard errors are shown for the estimates of the matching efficiency. Source: ONS, Reed, Author calculations

anti-correlation but it is not as strong, and there are some regions, most notably Northern Ireland and the South East, where they are positively correlated. Note also that these areas of correlated productivity and matching efficiency are also correlated with the areas of high mean labour market tightness as shown in Figure 3. From §5.1, the point estimate for the matching elasticity, α, for NUTS-1 region is lower than for occupation, i.e. tight regional labour markets do not cause as many new hires as tight occupational labour markets do.

5.3 Occupational mismatch, productivity and output

We run simulations of counter-factual paths for employment, output and productivity using the matching theory described in §4 at 1-, 2- and 3-digit SOC codes. The matching theory is designed to model the flows between employment and unemployment, and vice versa, not the flows into, and out of, the labour force. Therefore, in our matching function estimation we take the definition of job destructions and hires to be flows out of, and in to, employment.

Because our matching function does not recognise changes to employment and unemployment due to changes in the size of the labour force, the counter-factuals do not straightforwardly accord with the true paths taken by employment, output and output per worker. Even if they did, it is well-known that the

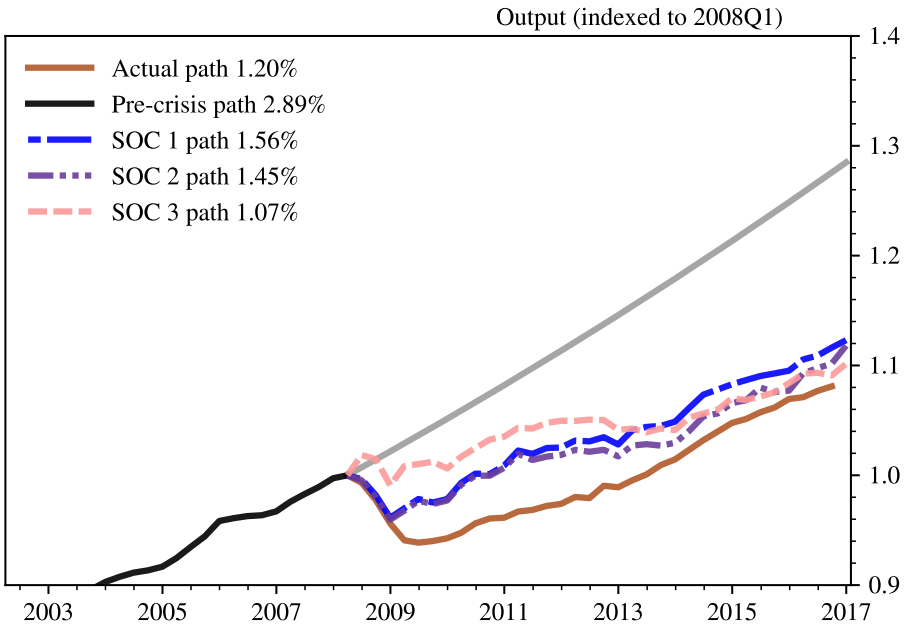


Figure 7: Realised and counter-factual paths for output. Simulations are run at different SOC code levels. Legend growth rates refer to the mean year-on-year growth rates measured quarterly. Source: ONS, Reed, Author calculations

longitudinal flows data in the LFS do not reproduce the stocks in the aggregate data due to response error, non-response bias, and that the longitudinal data form only a subset of the aggregate LFS data (Chandler, 2017; Jenkins and Chandler, 2010).

In light of this, we begin the counter-factual paths for employment, output and productivity from the same level as the aggregate paths in 2008 Q1 and map the evolution of their quarter-on-quarter growth rate from that start point using the growth rate in e^*_{it}. Figure 7 shows the social planner's optimal path for output. Counter-factual paths for employment and output would be around 4 and 2 percentage points higher respectively than the actual paths at the end of the simulation period. Given how tight the UK labour market was in general at the end of 2017, this implies a perhaps implausibly large reduction in the unemployment rate, with the almost complete elimination of structural unemployment due to mismatch. Taking this into account, the estimates of output and productivity growth should be seen as upper bounds, and the conclusion that even with the most optimistic scenarios for employment, the extent to which output can be raised is small. The year-on-year growth rates for output are around half the estimates of those in Patterson et al. (2016) – though theirs were over the period 2006 to 2012.

The surprise is in the path for productivity, where the counter-factual suggests that maximisation of output would entail *lower*, or similar, output per worker than has been realised in the data, as shown in Figure 8. Although there was a gap it closed in 2013, and reversed thereafter. The gap we find in output by

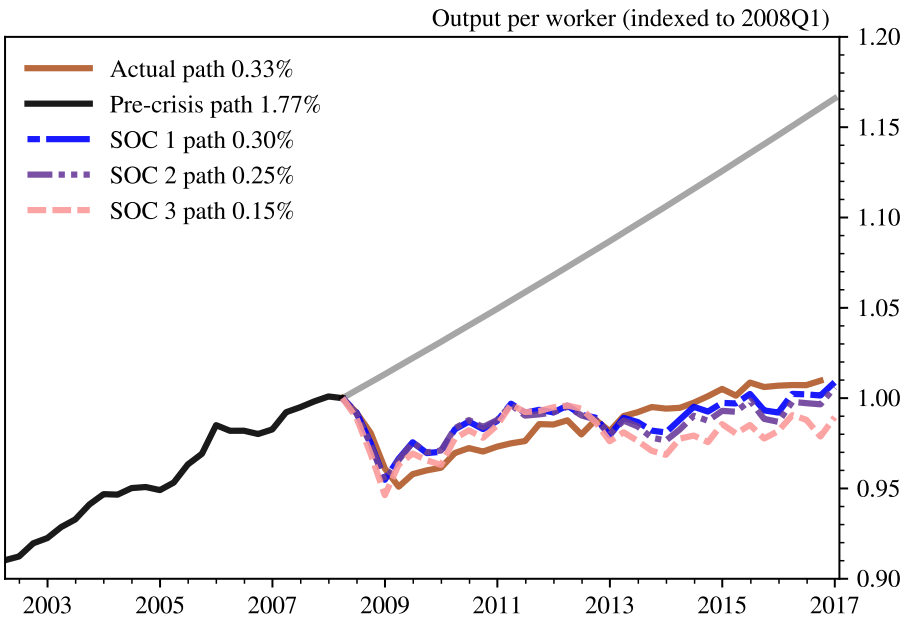


Figure 8: Realised and counter-factual paths for productivity. Simulations are run at different SOC code levels. Legend growth rates refer to the mean year-on-year growth rates measured quarterly. Source: ONS, Reed, Author calculations

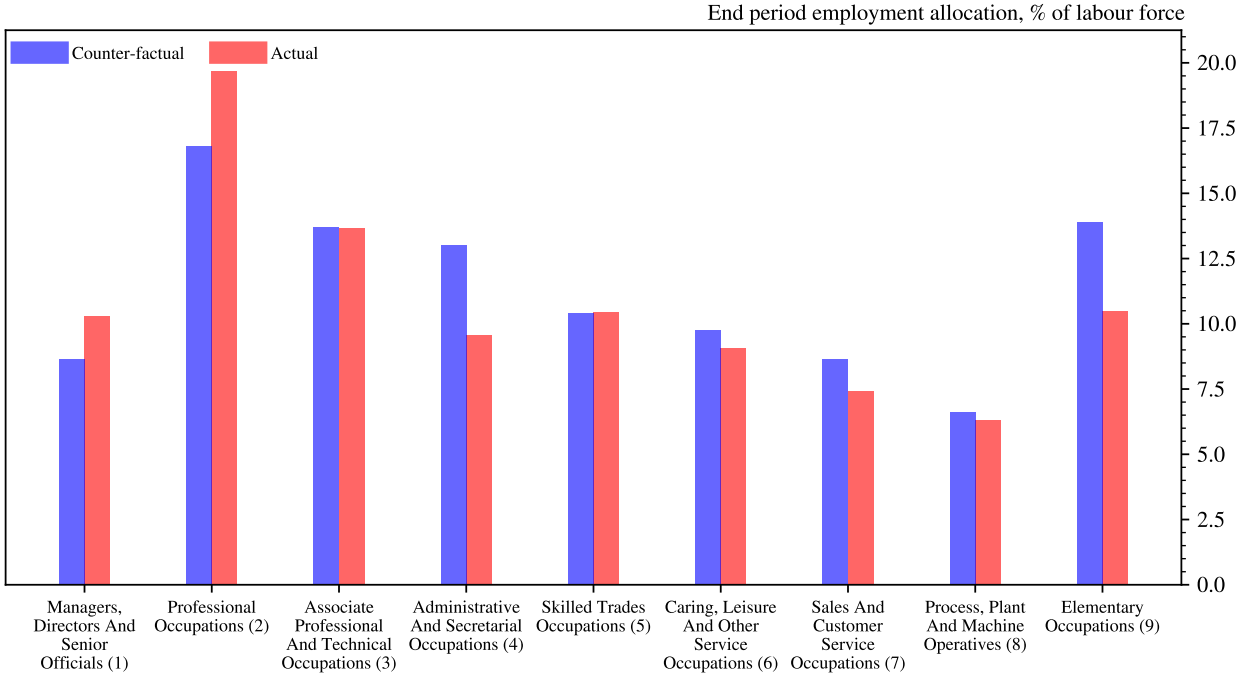


Figure 9: Realised and counter-factual final period employment rates by 1-digit SOC code. Source: ONS, Reed, Author calculations

occupation is substantially smaller than the one found in Patterson et al. (2016), with our counter-factual implying a year-on-year productivity growth rate which is less than a sixth of theirs. This counter-factual seems entirely counter-intuitive. But the social planner takes account of the weighted sum of the product of matching efficiency and productivity (see $\mathcal{M}_{xt}$, defined in equation (10) of Appendix B). As Figure 5 shows, though higher skilled jobs tend to have higher productivity, they tend to have considerably worse matching efficiency. In order to maximise output, the social planner chooses to have unemployed workers searching amongst lower productivity jobs. The underlying differences in the employment by market segment confirm this: at the end of the counter-factual simulation by 1-digit SOC code, although employment overall is higher, the SOC category which includes managers (1) accounts for proportionally fewer employees than those in so-called elementary occupations (9) in the actual distribution of employment, as shown in Figure 9. Further simulations, in Appendix D, show that the lower growth in output per worker is a consequence of optimising in favour of aggregate output, and this is driven by heterogeneity in matching efficiency.

The difference between our results and those of Patterson et al. (2016) are most likely to have their origin in the different datasets, which are likely to have different biases as noted in §3. Driven by this, we estimate significant variation in matching efficiency by occupation, whereas they do not. As shown in Figures 21 and 22 of Appendix D, estimates of matching efficiency can make a material difference to outcomes. Although our assumptions are broadly the same, there are some other differences. For their matching function estimation they use, as h_{it}, the average of the reduction in the stock of live vacancies and the off-flow of unemployment benefit claimants, whereas we use the flow from unemployment to employment by sub-market from the longitudinal LFS. However, our point estimate for α by 2-digit SOC code is only slightly lower, at 0.427, than theirs, which is 0.463. The period of study is also different; 2006–2012 versus 2008–2016 here. Finally, our SOC coding algorithm may have differences compared to the algorithm which assigned codes for the JobCentre Plus data. Using the occupational coding algorithm in §3.1, other online vacancy datasets could be processed in accordance with ours and used to provide further estimates.

Our data do not include a long enough time series to document whether there has been an overall decline in matching efficiency, or whether the extent of heterogeneity in ϕ has increased, using the occupational view of the labour market. Both of these have the potential to increase $\bar{u}$, the steady state of unemployment, given fixed job destruction rates and tightness. Previous results, referred to in Petrongolo and Pissarides (2001), suggest that the matching efficiency declines over time across countries. Hall and Schulhofer-Wohl (2015) attribute the decline in the US matching efficiency to changes in the composition of jobseekers. Pizzinelli and Speigner (2017) meanwhile find that the composition of unemployed jobseekers masked a 10% fall in the

matching efficiency between 1995 and 2016 in the UK. Lower matching efficiencies directly translate to worse outcomes for output and employment. Evidence on the changing heterogeneity of matching efficiencies is scarce, but the phenomenon of job polarisation (Goos and Manning, 2007) is likely to exacerbate differences in matching efficiency. As Figure 8 shows, an increase in the heterogeneity of ϕ_i can push down on $\overline{\frac{\mathrm{d}z}{\mathrm{d}t}}$.

Our simulations suggest that direct occupational mismatch cannot explain the UK's current productivity puzzle. As the simulated paths for output per worker show, it may have contributed to some of the level difference in the period up to 2012 but its effects then washed out. Additionally, realised output has only been marginally lower than its optimal path in the absence of any mismatch, suggesting that this effect has barely weighed down on output growth. Although the gap in actual and counter-factual output was substantial in 2011, the paths had shown signs of convergence by the end of 2017. And, given the very optimistic path for employment, these simulations represent an upper-bound on output and productivity growth rates in the absence of occupational mismatch.

However, this analysis does suggest that a shock to the higher skilled employment categories, for instance the four top 1-digit SOC codes, would be very damaging to short-term output because of their combination of high z and low ϕ. Based upon matching efficiency alone, those unemployed in managerial occupations (SOC code 1) will take around three times as long to find new work than those in elementary occupations (SOC code 9), and those in the former are estimated to be 2/3 more productive than those in the latter. At a more aggregate level, forecasts of output growth should perhaps factor in differences in unemployment by former occupation. The analysis also suggests that training aimed at increasing the rate of matching to the most productive jobs would be beneficial for the level of output, especially following a shock of the aforementioned kind.

5.4 Regional mismatch, productivity, and output

Regional sub-markets offer a different perspective on the effects of mismatch. There is regional heterogeneity in the UK on many dimensions; house price to income ratios, tightness (see Figure 3), and productivity and matching efficiency (see Figure 6). The differences in regional UK productivity are stark. Differences in productivity across regions can create a direct output level difference as compared to a counter-factual scenario with uniform productivity.

As with occupation, we simulate counter-factual paths for employment, output, and output per worker in the case where the social planner is free to move unemployed workers to more productive regions. In common with the previous simulations and literature, we assume that newly hired workers' marginal

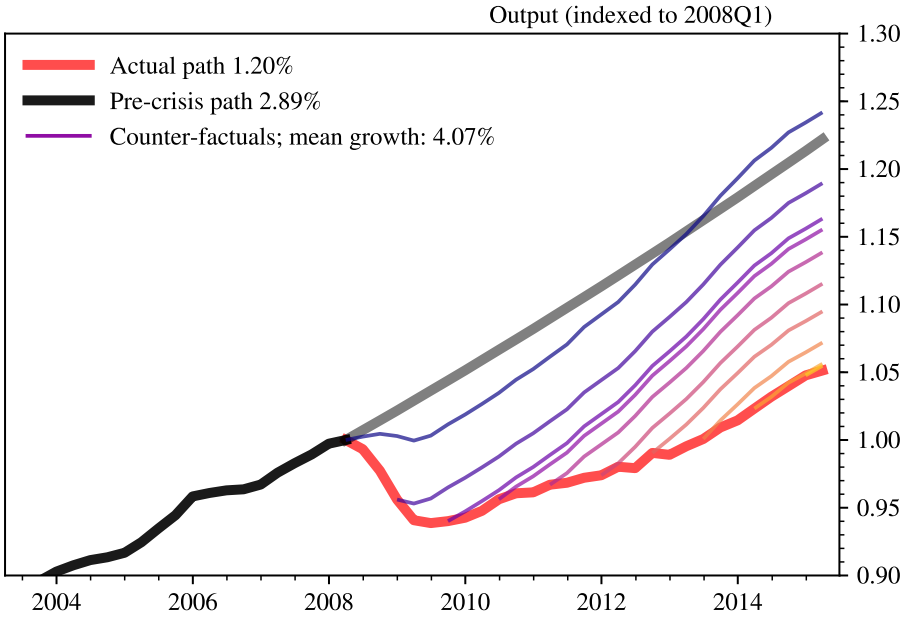


Figure 10: Realised and counter-factual paths for employment by 1-digit NUTS code. Legend growth rates refer to the mean year-on-year growth rates measured quarterly.
Source: ONS, Reed, Author calculations

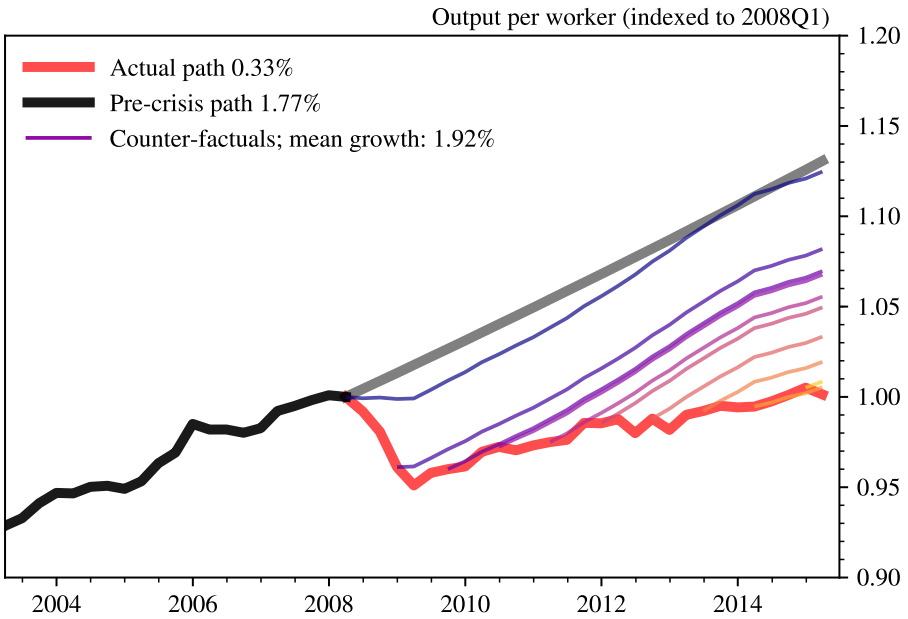


Figure 11: Realised and counter-factual paths for employment by 1-digit NUTS code. Legend growth rates refer to the mean year-on-year growth rates measured quarterly.
Source: ONS, Reed, Author calculations

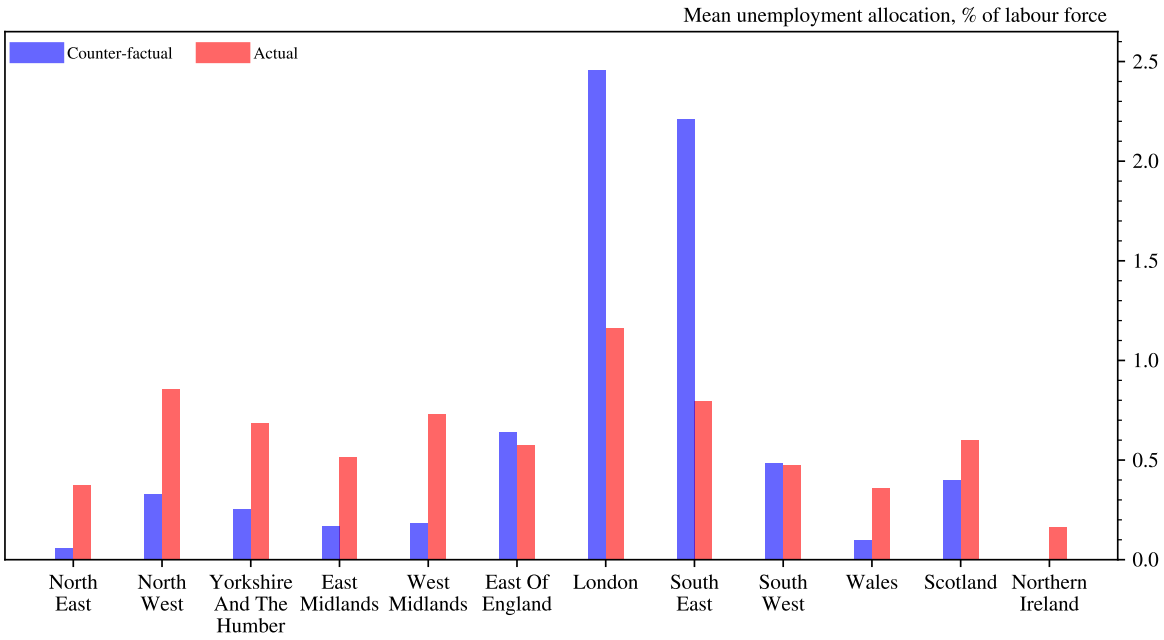


Figure 12: Realised and counter-factual mean rates of unemployment by 1-digit NUTS code. Source: ONS, Reed, Author calculations

productivity is equal to the mean productivity within the destination region. In the limit of all of the labour force being moved to a single region this is nonsensical but if the flows between regions represent only a small increase relative to reality then it is plausible.

Given that mismatch by region goes back at least to the 1970s, we produce counter-factuals from a range of start times beginning from the start of our vacancy data in 2008 Q1. We do not suggest that regional mismatch is behind the productivity puzzle; the simulations only show the extent to which eliminating mismatch would boost output and productivity, and in so doing work against the productivity puzzle. The choice to have multiple start dates is also motivated by the path for unemployment. The UK unemployment rate began to increase from 5.3% in 2008 to a high of 8.5% in 2011 Q3. As the social planner described in §4 can only allocate the unemployed, a high u means that the planner can make many re-allocations with potentially very large benefits. The series end date is limited by the available data on productivity by NUTS region, the latest release of which extends to 2015.

By running simulations beginning from every third quarter between 2008 Q1 and just prior to the end of the series, we give a fairer impression of the boost to output which might result from unwinding regional mismatch. These paths are shown for the 1-digit NUTS level in Figures 10, and 11 for output and productivity respectively. The NUTS average annual growth rates (across all simulated counter-factual paths) are also shown. These results suggest a very strong role for mismatch as an inhibitor of output and

productivity growth in the UK, no matter when the origin date. Starting from 2008Q1, unwinding regional mismatch could have almost entirely pushed out the productivity puzzle, and would have seen output at a higher level than the pre-crisis growth rate implies.

The driver of these results is made clear by looking at the social planner's mean unemployment allocations compared to the actual mean distribution of the unemployed across 1-digit NUTS regions, shown in Figure 12. The social planner shifts the unemployed from searching in many UK regions to searching predominantly in just three geographically contiguous ones: London, the South East, and the East of England. Note that this counter-factual implies only a doubling of the number of unemployed searching for work in London, rather than a doubling of the labour force size. The social planner allocates, on average, many more of those searching for work to the areas in which there is a confluence of favourable labour market conditions as given by high values of $\{\phi, \theta, z\}$. The benefits to output, and perhaps to productivity, of the social planner's allocation are especially large if there are regions which are very heterogeneous in $\{\phi, \theta, z\}$. From Figure 6 and Figure 3 it is apparent that such favourable regions exist; South East England has an average $\theta \approx 0.5$, much higher than most regions, and both high ϕ and z. London, while having a lower matching efficiency, has an unusually high productivity and typically a very tight labour market. Meanwhile, regions more similar to Northern Ireland have low ϕ, θ, and z. The allocations of the social planner reflect this.

The paths shown in Figures 10 and 11 seem high. By design, they do not take into account the very real barriers to moving, either housing costs or differences in the skills or qualifications required for jobs in different regions. However, as the North West and Yorkshire combined account for nearly 25% of the UK's output difference relative to the pre-crisis trend, and there are regions like London which are high in $\{\phi, \theta, z\}$, the regional results are plausible. If some of the differences between regions could be harmonised (either by reducing the heterogeneity in $\{\phi, \theta, z\}$, or by reducing moving costs for firms or workers), our results indicate that there may be a substantial boost to output and productivity. The results also imply consequences for wage growth, which has been very weak over the period of the recovery given the tightness of the labour market. Productivity is a primary determinant of wages, and so the re-allocation of the unemployed, some of whom will go on to become workers, to more productive and tighter regional submarkets would likely push up on wage growth too. One drawback of the analysis is that it only takes into account regional productivity; it does not factor in other differences between regions, such as the distribution of jobs by occupation or the average level of education.[8]

[8]These differences are large; the percentage of 25-64 year olds with tertiary education is around twice as high in London as in Northern Ireland. However, the causality of the relationship between these figures and the higher productivity level in

6 Conclusion

We have shown how to take naturally occurring vacancy data and map them into official occupational classifications in order to understand the effects of mismatch in the labour market on productivity and output according to two market segmentations: occupation and region. The tools we have developed can be deployed on other vacancy data, and could easily be adapted to other types of unstructured data – perhaps individual level surveys with free text fields for job descriptions. With this rich data, new estimates of the structural parameters of the theory of the matching function have been calculated.

By occupation, the effects of mismatch on productivity and output are small, and do not account for the productivity puzzle, even with an unemployment rate very close to zero in an output-optimising counter-factual scenario. However, we have highlighted the sensitivity of output to shocks to the top occupational groups, and one which could be brought on by rapid automation and technological change, or by a re-distribution of the demand for tasks brought about as a result of changing trading relationships. It is clear in this case that micro-level data are required to understand the macroeconomic picture: an aggregate analysis which used a mean matching efficiency would miss the important role that heterogeneous matching efficiency is playing in determining (counter-factual) output per worker.

By region, our results are dramatic but in-line with similarly strong results with respect to misallocation by region in the US. Unwinding the significant level of regional mismatch in the UK in 2008Q1 could have allowed output and productivity to grow at roughly equivalent rates to the pre-crisis period, both of which are considerably higher than the actual growth paths. Regional mismatch cannot explain the productivity puzzle, but its elimination at that time would have offset it. Counter-factuals beginning from dates between 2008 and 2015 do not bring the level of productivity and output back to the pre-crisis trend immediately, but do suggest that unwinding regional mismatch could have begun to raise the levels relative to the realised paths. Given the significant size of these effects, calculations of potential output should perhaps factor in the extent of regional dynamism. Future work could produce detailed estimates of the implied costs of moving, similar to Artuç, Chaudhuri and McLaren (2010), and incorporate them into the model used here. Similarly to the recent studies in the US, we have shown that regional differences can have a large impact on aggregate output, and differences in regional productivity growth a large impact on aggregate output growth.

All of our results suggest that close attention should be paid to the heterogeneous aspects of the labour market. Different matching efficiencies across occupations and regions result in different outflows from

London is complex.

unemployment to employment, meaning that the composition of jobs is important not only for determining output and productivity but also for determining unemployment dynamics. We have highlighted the inequalities of the matching process across regions, caused by the correlation of productivity, matching efficiency, and tightness within regions and the need for more work on the drivers and costs of these regional differences.

These conclusions demonstrate the power of large datasets to improve our understanding of macroeconomic phenomena, and drive home how important heterogeneity – in multiple dimensions – is to understand the aggregate picture of the economy. Micro-data matter for macro, and using rich, granular data sources can tease out exactly how.

References

Acemoglu, Daron, and Pascual Restrepo. 2017. "Robots and Jobs: Evidence from US labor markets."

Artuç, Erhan, Shubham Chaudhuri, and John McLaren. 2010. "Trade Shocks and Labor Adjustment: A Structural Empirical Approach." *American Economic Review*, 100(3): 1008–45.

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

Autor, David H, Frank Levy, and Richard J Murnane. 2003. "The skill content of recent technological change: An empirical exploration." *The Quarterly journal of economics*, 118(4): 1279–1333.

Barnett, Alina, Adrian Chiu, Jeremy Franklin, and María Sebastiá-Barriel. 2014*a*. "The productivity puzzle: a firm-level investigation into employment behaviour and resource allocation over the crisis." *Bank of England Quarterly Bulletin*, 54(2): 246.

Barnett, Alina, Sandra Batten, Adrian Chiu, Jeremy Franklin, Maria Sebastia-Barriel, et al. 2014*b*. "The UK productivity puzzle." *Bank of England Quarterly Bulletin*, 54(2): 114–128.

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

Battu, Harminder, Ada Ma, and Euan Phimister. 2008. "Housing tenure, job mobility and unemployment in the UK." *The Economic Journal*, 118(527): 311–328.

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.

Bennett, Robert J, and Ricardo R Pinto. 1994. "The hiring function in local labour markets in Britain." *Environment and Planning A*, 26(12): 1957–1974.

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

Bholat, David, Stephen Hansen, Pedro Santos, and Cheryl Schonhardt-Bailey. 2015. *Text mining for central banks. Handbooks*, Centre for Central Banking Studies, Bank of England.

Blanchard, Olivier Jean, and Peter A Diamond. 1989. "The aggregate matching function." National Bureau of Economic Research.

Blundell, Richard, Claire Crawford, and Wenchao Jin. 2014. "What can wages and employment tell us about the UK's productivity puzzle?" *The Economic Journal*, 124(576): 377–407.

Borowczyk-Martins, Daniel, Grégory Jolivet, and Fabien Postel-Vinay. 2013. "Accounting for endogeneity in matching function estimation." *Review of Economic Dynamics*, 16(3): 440–451.

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.

Broadbent, Ben. 2012. "Productivity and the allocation of resources." *Speech given at Durham Business School*, 12.

Bryson, Alex, and John Forth. 2015. "The UK's Productivity Puzzle." CEPREMAP Working Papers (Docweb) 1511.

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*.

Chandler, M. 2017. "Labour market flows: May 2017." *ONS Reports*, Labour Market Releases.

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

Davis, Steven J, R Jason Faberman, and John C Haltiwanger. 2013. "The establishment-level behavior of vacancies and hiring." *The Quarterly Journal of Economics*, 128(2): 581–622.

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.

Diamond, Peter A. 1982. "Wage determination and efficiency in search equilibrium." *The Review of Economic Studies*, 49(2): 217–227.

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."

Field, Simon, and Mark Franklin. 2013. "Micro-data perspectives on the UK productivity conundrum – an update." *ONS Reports*.

Gavazza, Alessandro, Simon Mongey, and Giovanni L Violante. 2016. "Aggregate recruiting intensity." National Bureau of Economic Research.

Goodridge, Peter, Jonathan Haskel, and Gavin Wallis. 2014. "The UK productivity puzzle is a TFP puzzle: current data and future predictions." Imperial College, London, Imperial College Business School Working Papers.

Goos, Maarten, and Alan Manning. 2007. "Lousy and lovely jobs: The rising polarization of work in Britain." *The review of economics and statistics*, 89(1): 118–133.

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.

Haldane, A. G. 2017. "Productivity Puzzles." Bank of England speech given at the London School of Economics.

Hall, Robert E, and Sam Schulhofer-Wohl. 2015. "Measuring job-finding rates and matching efficiency with heterogeneous jobseekers." National Bureau of Economic Research.

Haskel, J, P Goodridge, A Hughes, and G Wallis. 2015. "The contribution of public and private R&D to UK productivity growth." Imperial College, London, Imperial College Business School Working Papers 21171.

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

Hsieh, Chang-Tai, and Enrico Moretti. 2017. "Housing Constraints and Spatial Misallocation."

Jäger, Kirsten. 2017. "EU KLEMS Growth and Productivity Accounts 2017 Release, Statistical Module1."

Jenkins, K, and M Chandler. 2010. "Labour market gross flows data from the Labour Force Survey." *ONS Reports*, Economic and Labour Market Review.

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

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.

Martin, Bill, and Robert Rowthorn. 2012. "Is the British economy supply constrained II? A renewed critique of productivity pessimism." *ONS Reports*.

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.

Pessoa, João Paulo, and John Van Reenen. 2014. "The UK productivity and jobs puzzle: does the answer lie in wage flexibility?" *The Economic Journal*, 124(576): 433–452.

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.

Pizzinelli, Carlo, and Bradley Speigner. 2017. "Matching efficiency and labour market heterogeneity in the United Kingdom." *Staff Working Paper No. 667*.

Rabe, Birgitta, and Mark P Taylor. 2012. "Differences in opportunities? Wage, employment and house-price effects on migration." *Oxford Bulletin of Economics and Statistics*, 74(6): 831–855.

Riley, Rebecca, Chiara Rosazza-Bondibene, and Garry Young. 2014. "The financial crisis, bank lending and UK productivity: sectoral and firm-level evidence." *National Institute Economic Review*, 228(1): R17–R34.

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)*.

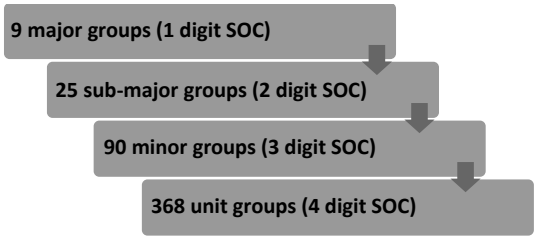
Smith, Jennifer C. 2012. "Unemployment and Mismatch in the UK."

A Labelling job vacancies with SOC codes

Algorithm 1, in §3.1, matches job vacancies up to SOC codes at the 3-digit level, or 'minor group' (see Figure 13). The algorithm has four main stages; cleaning, exact matching, identification of similar minor groups, and fuzzy matching. The flow of the algorithm is shown in Figure 14. The algorithm uses three

look-up dictionaries. The dictionaries are the *known titles dictionary*, *known words dictionary*, and *acronym expansion dictionary*. These dictionaries are available as part of the release of the coding algorithm.

Figure 13: Schematic of SOC code hierarchy.



Before being matched, the text associated with each job vacancy is cleaned. The cleaning process for text converts plural forms to singular forms (with some exceptions), expands abbreviations, removes stopwords[9], and removes digits, punctuation, and extra spaces. In the first step of the algorithm, the *known titles dictionary* is used to create a vector space of known job titles and their associated SOC codes. This is achieved by creating a single string for each 3-digit SOC code which combines all of the known job titles and a short official job description of that SOC code. We use term frequency-inverse document frequency (tf-idf) to represent each vacancy as a matrix with dimension $T \times D$ where t is a term. Our terms are comprised of all 1–3-grams. An n-gram is all combinations of words with n words, so all 1–3-grams is all combinations of words with a length less than or equal to three words. d is a document (in this case a text string corresponding to each 3-digit SOC code) with D documents (the number of unique 3-digit SOC codes) in total. We use the SCIKIT-LEARN Python package to do this.

The known 3-digit SOC code job titles dictionary (*known titles dictionary*) is an index which links around 10^4 alternative job titles to SOC codes. The *known titles dictionary* is used for identifying exact job title matches and for fuzzy matching. Publicly available ONS resources were used to generate the dictionary; the ONS Standard Occupational Classification, an extract from which may be seen in Table 4, and the Standard Occupational Classification 2010 Index, an extract from which is shown in Table 5. As shown in Figure 13, 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 6.

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

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

The *acronym expansion dictionary* is used in 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 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 7 shows an extract from the *acronym expansion dictionary*.

The *known words dictionary* 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. 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. 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 (described further in the document), for exact and fuzzy job title matching, it was decided to use a stricter list of stopwords in order to increase the

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 5: 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 6: 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 7: An extract from the *acronym expansion dictionary.*

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'. Secondly, there are 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.

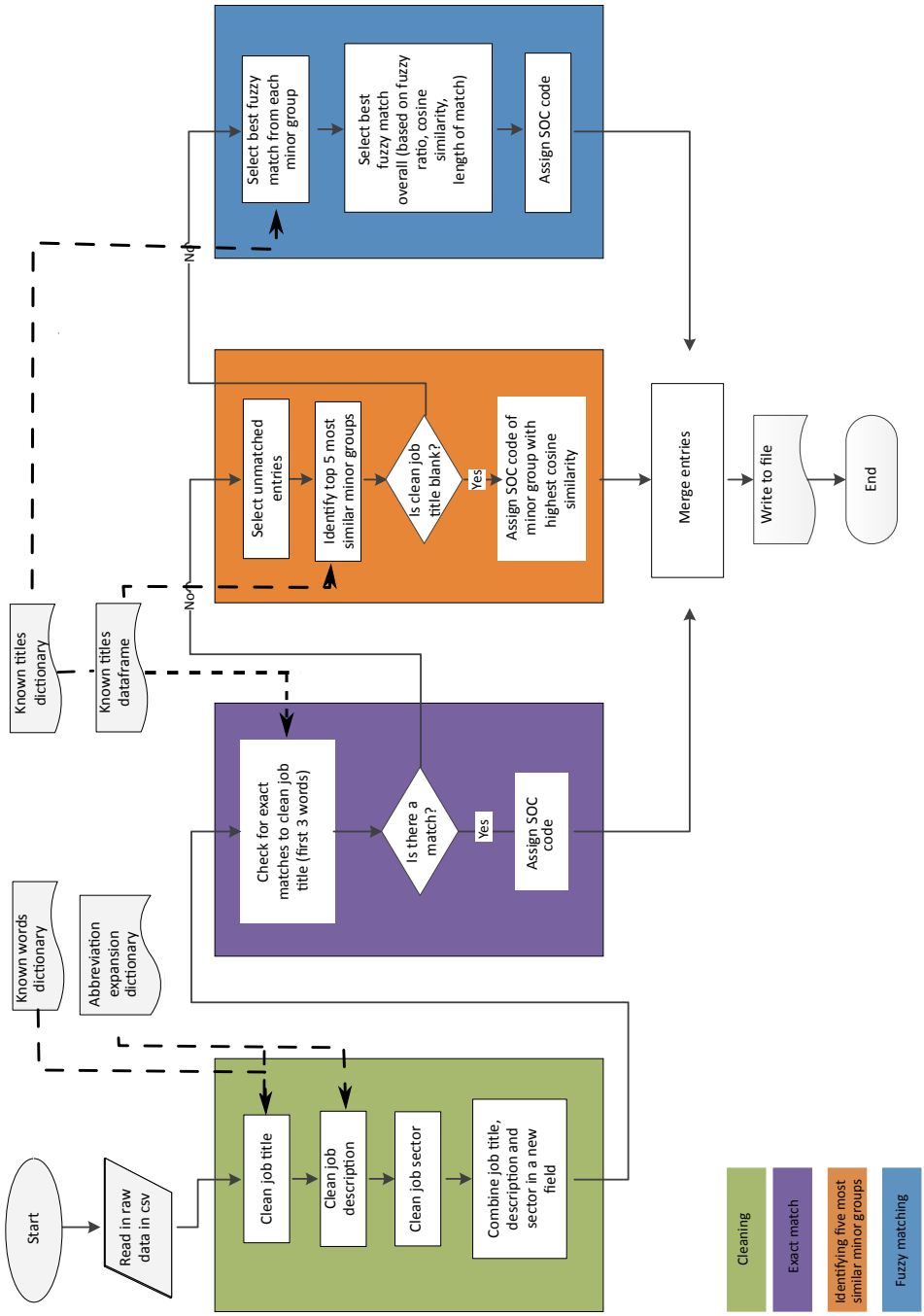


Figure 14: An overview of the algorithm which matches job vacancies to SOC codes at the minor group level (3-digit SOC code).

B Theory

In our model, the joint dynamics of unemployment and vacancies are given by

$$\frac{\mathrm{d}U}{\mathrm{d}t} = \xi(L-U) - h(U,V)$$
$$\frac{\mathrm{d}V}{\mathrm{d}t} = \Gamma - h(U,V)$$

with ξ the job destruction rate and Γ the flow of newly created vacancies. This neglects labour force entry and exit, and job-to-job flows. The Beveridge curve is the locus of points in U-V space such that $\dot{U} = 0$, so that $\xi(L-U) = h(U,V)$ and (under constant returns to scale)

$$\xi = h\left(\frac{u}{1-u}, \frac{v}{1-u}\right)$$

Given h, u, v, and ξ, a Beveridge curve can be traced out (see §5.1 for plots).

In discrete time, $\Delta u_{t+1} = \xi_t(1-u_t) - h_t$ and the steady state unemployment rate is

$$\bar{u} = \frac{\xi_t}{\xi_t + h_t} \tag{5}$$

Next period unemployment is determined by the rate of convergence to the steady state unemployment rate so that

$$u_{t+1} = (h_t + \xi_t)\,\bar{u} + [1 - (h_t + \xi_t)]\,u_t \tag{6}$$

A decline in matching efficiency ϕ has two effects on unemployment in the next period: as h_t falls, it raises $\bar{u}_t$, and so the steady state unemployment rate is higher; and it also increases the persistence of unemployment through the second term in equation (6) because the rate of convergence to the steady state is $(h_t + \xi_t)$. A lower matching efficiency means fewer hires, which means less output over time.

From the *Labour Force Survey*, the hires and job destruction rate in each market segment can be calculated. Let $p(\mu,\nu)$ denote a specific individual who, from quarter $t-1$ to quarter t, transitions from status μ to status ν, where μ and ν can take values of e or u for employed or unemployed respectively. The job destruction rate for a segment of the labour market i is given by

$$\xi_{it} = \frac{\sum_p p(e,u)_{it}}{\sum_p p(e,e)_i + p(u,e)_{it}} \tag{7}$$

while hires into i are given by

$$h_{it} = \sum_{p} p(u, e)_{it} \tag{8}$$

Equations (7) and (8) are used, respectively, to define the flow out of, and into, employment. Similarly, when calculating the counter-factual path for employment,

$$e^*_{it} = (1 - \xi_{i,t-1})e^*_{i,t-1} + \frac{h_{it}}{1 - \mathcal{M}_{xt}}$$

equation (7) is used for the job destruction rate (which is taken to be exogenous) and the model implied number of hires h_{it}, given by equation (2), are used.

To estimate the effect of mismatch, we use the search-and-matching framework developed by Şahin et al. (2014) and used by Patterson et al. (2016) and Smith (2012), where solutions to more generalised versions of the problem may be found. Given I market segments, this model gives a counter-factual and optimal path for output by imagining a social planner that assigns the unemployed to different market segments. It is solved here with a homogeneous job destruction rate, ξ_t. Let $\boldsymbol{\Xi}_t$ be a set of parameters representing known constants in discrete time labelled by t such that

$$\boldsymbol{\Xi}_t = (\boldsymbol{z}_t, \boldsymbol{V}_t, \boldsymbol{\phi}_t, \xi_t)$$

where the vectors are in bold fonts. The vectors are of length I and represent productivity, the stock of vacancies, and matching efficiency across sub-markets respectively. ξ is the cross-market job destruction rate. Let u_t be unemployment and $\boldsymbol{e}_t$ be the vector of employment by market segment. The social planner operates as follows; firstly, $\boldsymbol{\Xi}_t$ are observed. Then $\boldsymbol{e}_t$ is given, determining u_t. Next, unemployed workers searching in u_i are matched so that there are

$$h_i = \phi_i M(U_i, V_i)$$

new hires in segment i within period t. Production occurs in the existing matches given by $\boldsymbol{e}_t$ and the new hires given by $\boldsymbol{h}_t$, though new hires are assumed to be a fraction $\gamma < 1$ less productive. Job destruction occurs, determining the next period's employment $\boldsymbol{e}_{t+1}$. Then the planner chooses the division of searchers for the next period. With that determined, L_{t+1} (next period labour force size) and the next period stock of employed, $e_{t+1} = \sum_i e_{i,t+1}$, set the next period stock of unemployed workers u_{t+1}.

The planner's problem is given by

$$V(u_t, \boldsymbol{e}_t; \boldsymbol{\Xi}_t) = \max_{\{u_{i,t}\}} \left\{ \sum_i z_{i,t}(e_{i,t} + \gamma h_{i,t}) - \xi_t u_t + \beta \mathbb{E}\left[V(u_{t+1}, \boldsymbol{e}_{t+1}; \boldsymbol{\Xi}_{t+1})\right] \right\}$$

such that $\sum_i u_{i,t} \leq u_t$. Also note that

$$e_{i,t+1} = (1 - \xi_t)(e_{i,t} + h_{i,t})$$

$$u_{t+1} = L_{t+1} - \sum_i e_{i,t+1}$$

The Lagrangian for the problem is

$$\mathcal{L} = \max_{\{u_{i,t}\}} \{V(u_t, \boldsymbol{e}_t; \boldsymbol{\Xi}_t)\} - \mu \left(\sum_i u_{i,t} - u_t \right)$$

The first order condition is

$$\frac{\partial \mathcal{L}}{\partial u_{i,t}} = \frac{\partial f}{\partial u_{i,t}} - \mu = 0$$

so that

$$\gamma z_{i,t} \phi_{i,t} \frac{\partial M}{\partial u_{i,t}} + \beta \mathbb{E}\left[\frac{\partial V_{t+1}}{\partial u_{i,t}}\right] = \mu$$

where

$$\frac{\partial V_{t+1}}{\partial u_{i,t}} = \frac{\partial V_{t+1}}{\partial e_{j,t+1}} \frac{\partial e_{j,t+1}}{\partial u_{i,t}} + \frac{\partial V_{t+1}}{\partial u_{t+1}} \frac{\partial u_{t+1}}{\partial e_{j,t+1}} \frac{\partial e_{j,t+1}}{\partial u_{i,t}}$$

with

$$\frac{\partial e_{j,t+1}}{\partial u_{i,t}} = (1 - \xi_t)\phi_j \frac{\partial M}{\partial u_{i,t}} \delta_{ij}$$

and δ_{ij} the Kronecker delta. Then

$$\frac{\partial u_{t+1}}{\partial e_{j,t+1}} = -\sum_k \delta_{jk}$$

so that

$$\gamma z_{i,t} \phi_{i,t} \frac{\partial M}{\partial u_{i,t}} + (1 - \xi_t)\phi_{i,t} \frac{\partial M}{\partial u_{i,t}} \beta \mathbb{E}\left[\frac{\partial V_{t+1}}{\partial e_{j,t+1}} - \frac{\partial V_{t+1}}{\partial u_{t+1}}\right] = \mu$$

The envelope theorem gives that

$$\frac{\partial V_t}{\partial u_t} = \frac{\partial \mathcal{L}_t}{\partial u_t} = \mu - \xi_t$$

and

$$\frac{\partial V_t}{\partial e_{i,t}} = \frac{\partial \mathcal{L}_t}{\partial e_{i,t}} = z_{i,t} + \beta(1-\xi_t)\mathbb{E}\left[\frac{\partial V_{t+1}}{\partial e_{j,t+1}} - \frac{\partial V_{t+1}}{\partial u_{t+1}}\right]$$

The optimal decision for the labour force size in the next period, L_{t+1} , is such that $\mathbb{E}\left[\frac{\partial V_{t+1}}{\partial u_{t+1}}\right] = 0$. With this, and the assumption that $\boldsymbol{z}_t$ and ξ_t are martingales, the second envelope condition can be iterated forward to give

$$\mathbb{E}\left[\frac{\partial V_{t+1}}{\partial e_{j,t+1}}\right] = \frac{z_i}{1-\beta(1-\xi)}$$

Now the first order condition is

$$\gamma z_{i,t}\phi_{i,t}M_{u_{i,t}} + \frac{\beta(1-\xi)}{1-\beta(1-\xi)}z_{i,t}\phi_{i,t}M_{u_{i,t}} = \mu$$

The matching function is assumed to be a smooth and positive increasing function of its arguments in the Cobb-Douglas form and with constant returns to scale such that its derivative is a function of the ratio of its arguments only, i.e.

$$\frac{\partial M}{\partial u_{i,t}} = M_{u_{i,t}}\left(\frac{v_i}{u_i}\right)$$

For fixed $v_{i,t}$, this means that $M_{u_{i,t}}$ is a positive decreasing function of $u_{i,t}$. The first order condition now implies that

$$\gamma z_{i,t}\phi_{i,t}M_{u_{i,t}} + \frac{\beta(1-\xi)}{1-\beta(1-\xi)}z_{i,t}\phi_{i,t}M_{u_{i,t}}$$

The social planner therefore tries to equalise

$$z_i\phi_i\frac{\partial M\left(\frac{V_i}{U_i^*}\right)}{\partial u_{i,t}}$$

across all sub-markets i.

Defining $\chi_{it} = z_{it}\phi_{it}$, the social planner chooses starred values such that

$$\frac{V_{jt}}{U_{jt}^*} = \left(\frac{\chi_{it}}{\chi_{jt}}\right)^{\frac{1}{\alpha}}\frac{V_{it}}{U_{it}^*}$$

The sum over j gives

$$U_{it}^* = \chi_{it}^{\frac{1}{\alpha}}\left(\frac{V_{it}}{\sum_j \chi_{jt}^{\frac{1}{\alpha}}v_{jt}}\right)\frac{1}{U_t}$$

and the output from new hires following the social planner's optimum allocation is

$$y_t^* = \gamma \sum_i z_{it} V_{it}^{\alpha} \left(U_{it}^*\right)^{1-\alpha}$$

Using the expression for U_{it}^* and defining

$$X_t = \left[\sum_i^I (\chi_{it})^{\frac{1}{\alpha}} \left(\frac{v_{it}}{v_t}\right)\right]^{\alpha}$$

as a constant elasticity of substitution aggregator of segment-specific matching and productivity weighted by vacancy shares, then

$$y_t^* = \gamma V_t^{\alpha} U_t^{1-\alpha} X_t$$

is the counter-factual path for output due to new hires. The output from new hires given by the econometric estimation of the data is

$$y_t = \gamma V_t^{\alpha} U_t^{1-\alpha} \left[\sum_{i=1}^I \left(\frac{\chi_{it}}{X_t}\right) \left(\frac{v_{it}}{v_t}\right)^{\alpha} \left(\frac{u_{it}}{u_t}\right)^{1-\alpha}\right]$$

By comparing the output from new hires, y_t, given the path taken by unemployment, u_t, in reality with the path chosen for output by the social planner, y_t^*, an index of the aggregate output loss due to new hires can be constructed:

$$\mathcal{M}_{yt} = \frac{y_t^* - y_t}{y_t^*} = 1 - \sum_{i=1}^I \left(\frac{z_{it}\phi_{it}}{X_t}\right) \left(\frac{v_{it}}{v_t}\right)^{\alpha} \left(\frac{u_{it}}{u_t}\right)^{1-\alpha} \tag{9}$$

which is bounded between 0 and 1, with maximal mismatch given by unity.

Given the counter-factual output due to new hires, y_t^*, the counter-factual total output, employment, and productivity can be estimated. Şahin et al. (2014) showed that the counter-factual hires are given by $h_{it}^* = h_{it}/(1 - \mathcal{M}_{xt})$ where

$$\mathcal{M}_{xt} = 1 - \sum_i^I \left(\frac{\phi_{it}}{\varphi_t}\right) \left(\frac{v_{it}}{v_t}\right)^{\alpha} \left(\frac{u_{it}}{u_t}\right)^{1-\alpha} \tag{10}$$

with

$$\varphi_t = \sum_i^I \phi_{it} \left(\frac{z_{it}\phi_{it}}{X_t}\right)^{\frac{1-\alpha}{\alpha}} \left(\frac{v_{it}}{v_t}\right)$$

Counter-factual output is then

$$Y_t^* = \sum_i^I z_{it} e_{it}^* + y_t^* \tag{11}$$

where $e_{it}^* = (1 - \xi_{t-1}) e_{i,t-1}^* + h_{it}^*$. The same relationship applies to unstarred values, with $h_{it} = \phi_{it} V_{it}^{\alpha} U_{it}^{1-\alpha}$. Output per worker in the realised and counter-factual cases is given by Y_t/e_t and Y_t^*/e_t^* respectively.

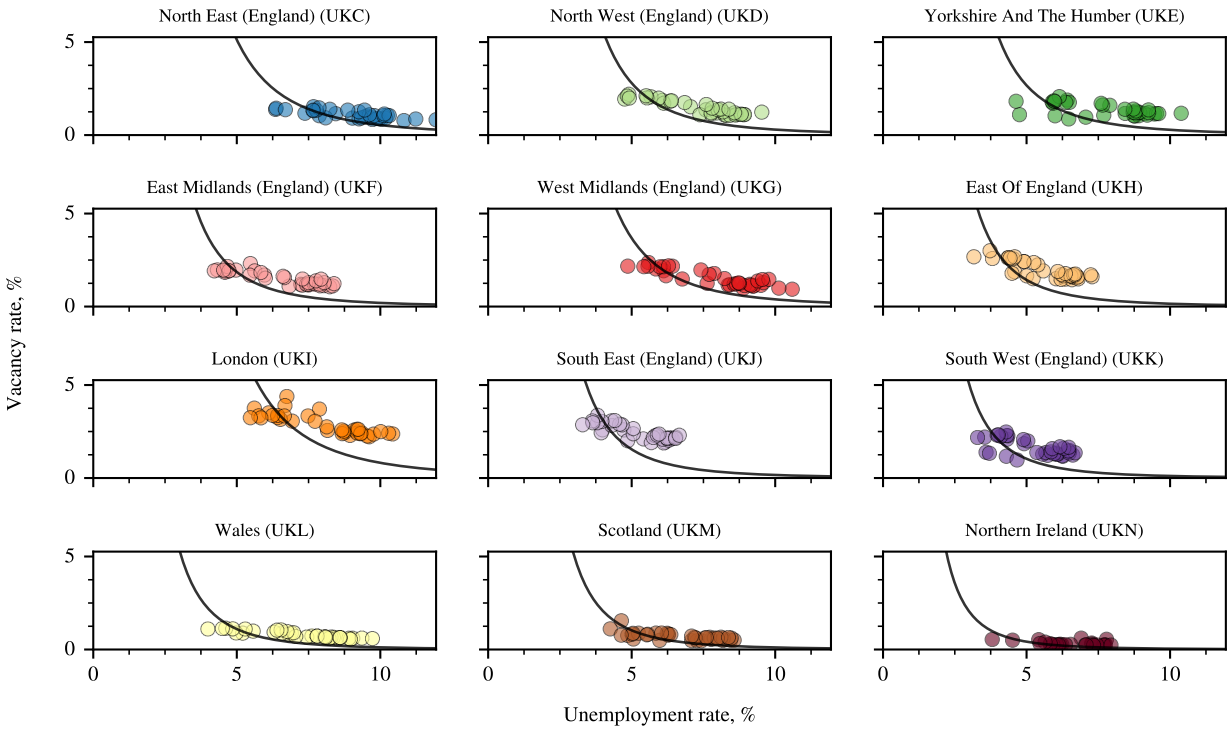


Figure 15: Beveridge curves (lines) using estimates of the parametersin equation (4) and data (points) in u-v space for each 1-character NUTS code at quarterly frequency. Source: ONS, Reed, Author calculations

C Heterogeneity in the UK labour market

There is evidence that, across both regions and occupations, there is significant heterogeneity in the UK labour market. This matters for the simulations in §5.3 and §5.4; greater differences in ϕ, z, and θ will tend to mean more re-allocation by the social planner, especially if the differences are correlated.

Beveridge curves by region show this; see Figure 15. There is less variation in the volatility compared to the equivalent set of Beveridge curves by SOC code. The relative shifts in the curves in Figure 15 along the unemployment axis are particularly stark, for instance between Wales (UKL) and London (UKI).

Figure 16 shows how the standard deviation of productivity across SIC sections grew dramatically before the crisis and remained elevated at the end of 2017. This carries over into SOC3; Figure 18 shows that large and weakly increasing heterogeneity across occupational groups.

Figure 17 shows that this pattern is repeated across NUTS3 groups too, although the standard deviation has grown much less dramatically than across sectors. There was an increase in the standard deviation of productivity per worker across 1-character UK NUTS regions of 21% from 2002 to 2016. 18 percentage points of this occurred following the financial crisis. Differences by region are correlated with other dispar-

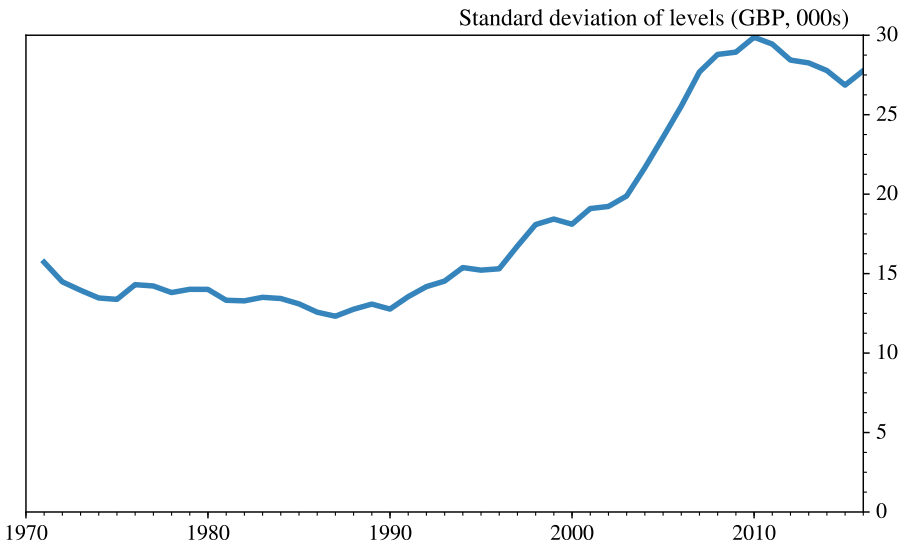


Figure 16: The standard deviation of productivity across SIC sections has grown Haldane (2017); Jäger (2017). Excludes the mining & extraction, energy, and real estate industries. Source: EUKlems productivity database, ONS and Author calculations

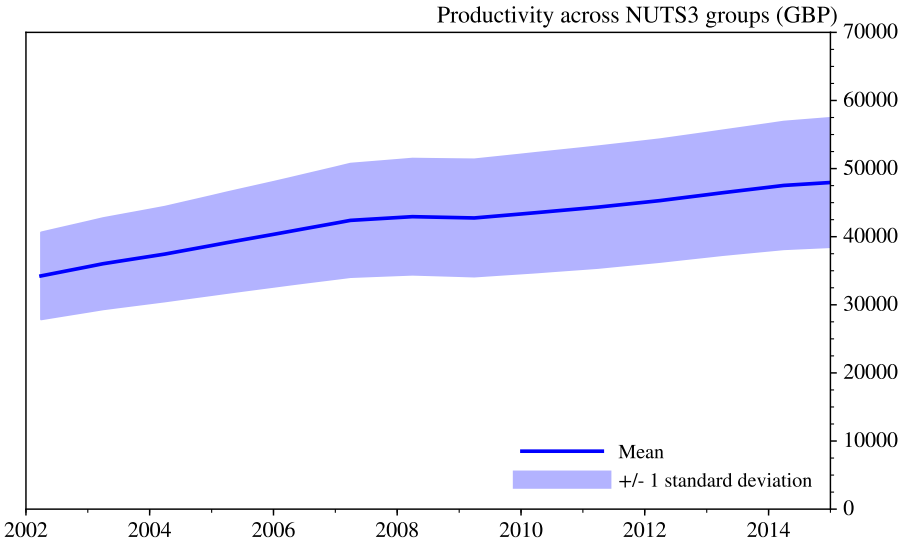


Figure 17: Mean productivity growth across NUTS groups has been gradual, while the standard deviation has increased. Uses the chained-volume measure. Source: ONS

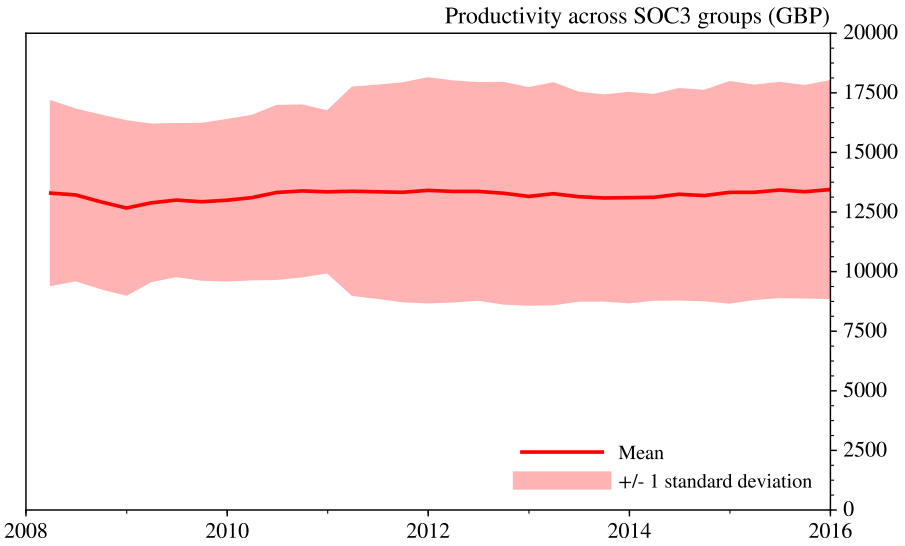


Figure 18: Mean productivity growth across SOC groups is stagnant, while the standard deviation has increased. Uses the chained-volume measure by sector converted into occupations via the equations given in (1). Source: ONS, Reed, Author calculations.

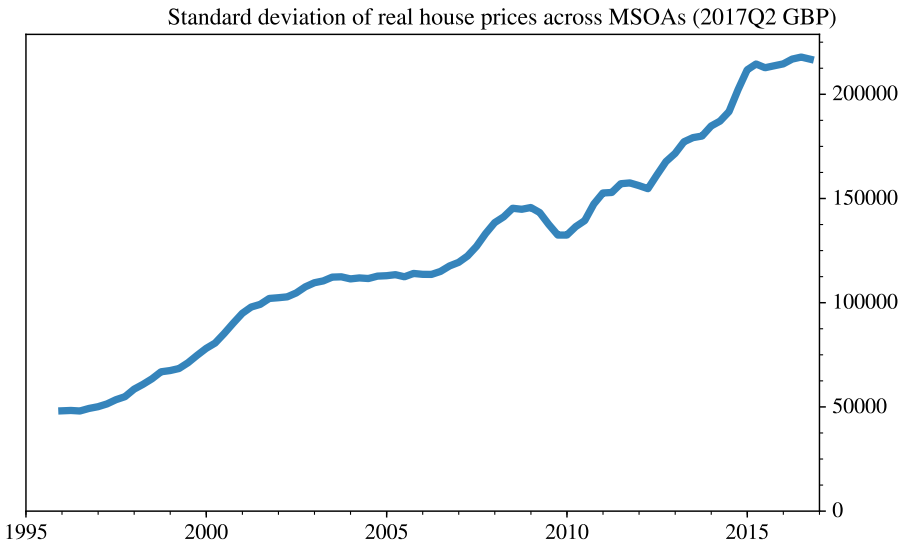


Figure 19: The standard deviation of UK house prices across medium super output areas (MSOAs) in real terms using the GDP deflator. Source: ONS, Author calculations

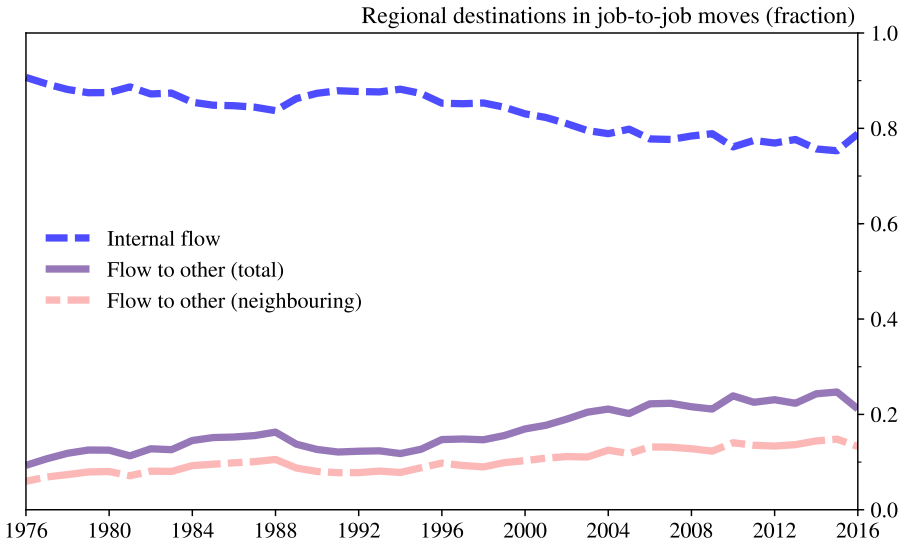


Figure 20: Regional job-to-job flows between 1-character UK NUTS regions have changed very little over two decades despite highly variable macroeconomic conditions. Source: ONS (ASHE), Author calculations

ities, some of which have grown more severe over time. Real house price differences across regions grew by almost a factor of five between 1996 and 2017, as shown in Figure 19 and, as Rabe and Taylor (2012) show, differences in house price levels are important determinants of migration for homeowners – although house price levels themselves are likely to be driven by employment opportunities. In agreement with previous UK studies, such as Battu, Ma and Phimister (2008), data from the *Annual Survey of Hours and Earnings* suggests that of those who do move jobs, very few also move region.[10] Even when people do move region for a job, they mostly move to a neighbouring region. The fraction of job-to-job moves which are within a region, without a region and to neighbouring regions over time are shown in Figure 20 using NUTS regions.

[10]The *Labour Force Survey* is unsuitable for this as it is a survey of establishments rather than individuals, and so does not accurately capture the geographical moves undertaken by individuals.

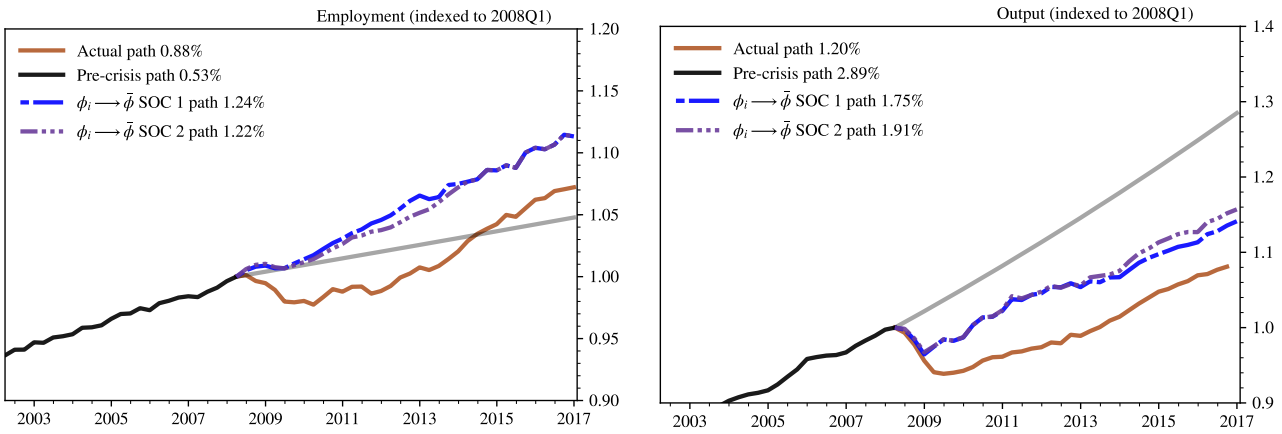


Figure 21: Realised and counter-factual paths for employment and output at the 1- and 2-digit SOC code level assuming all occupations have the same matching efficiency $\bar{\phi}$. Source: ONS, Reed, Author calculations

D Simulations of occupational counter-factuals with equalised matching efficiencies

Simulations imply that lower growth in output per worker is a consequence of optimising in favour of aggregate output, and this optimisation is partly driven by heterogeneity in matching efficiency. To demonstrate this, Figures 21 and 22 show another counter-factual, just at the 1- and 2-digit SOC code levels, in which the matching efficiency is set to be the mean so that $\phi_i \longrightarrow \overline{\phi}$ for all i. The figures show a very modest increase in the level of productivity in this scenario. Output per worker increases if the distribution of matching efficiencies is flat; different to what observed in the data. A homogeneous matching efficiency with the same mean does account for a very small sliver of the productivity puzzle. Output is around 3 percentage points higher; productivity 4 percentage points at most.

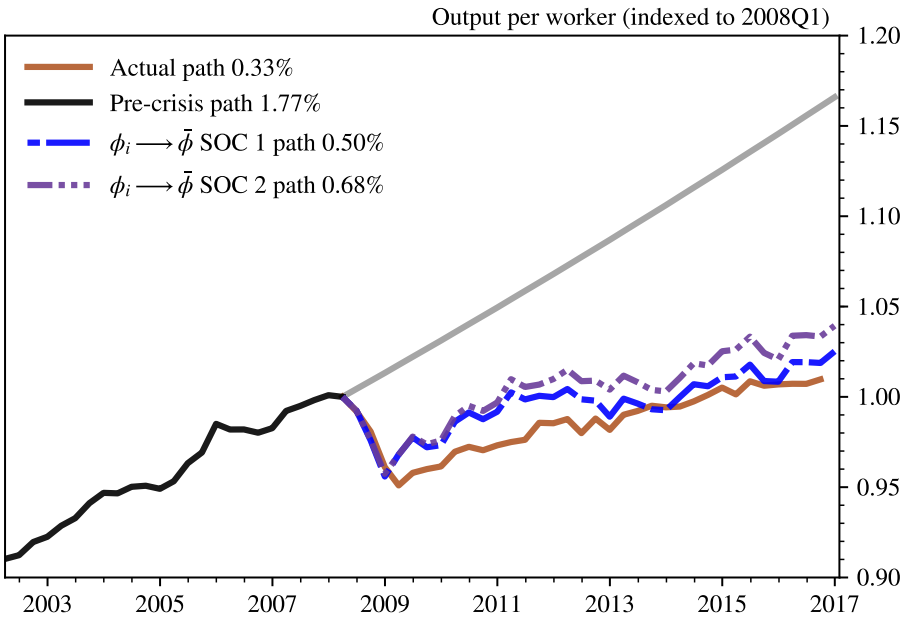


Figure 22: Realised and counter-factual paths for productivity at the 1- and 2-digit SOC code level assuming all occupations have the same matching efficiency. Source: ONS, Reed, Author calculations