Is the UK's productivity puzzle mostly driven by occupational mismatch? An analysis using big data on job vacancies*

Arthur Turrell† Bradley Speigner‡ David Copple§ Jyldyz Djumalieva¶
James Thurgood‖

Abstract

Uncertainty still remains as to the cause of the UK's dramatic productivity puzzle that began during the Great Financial Crisis. Occupational mismatch has been implicated as driving up to two thirds of it. However, obtaining the high quality time series data for vacancies by job occupation that are required to measure occupational mismatch is a significant challenge. We confront this issue by using a weighted dataset of 15 million job adverts posted online that cover most of the post-crisis period and which enable us to test whether occupational mismatch still stands up as an explanation for the UK productivity puzzle. We find little evidence that it does, mainly because, relative to the data used in similar analysis by Patterson et al. (2016), our vacancy data imply greater heterogeneity in occupational matching frictions, a key determinant of the optimal distribution of labour across job types.

Keywords: vacancies, mismatch, productivity puzzle

JEL Codes: E24, C55, J63

*First version: July, 2018. This version: June, 2021. The views expressed in this work do not represent the views of the Bank of England, ONS, or Natwest. We are grateful to Katharine Abraham, James Barker, David Bholat, Emmet Cassidy, Matthew Corder, Daniel Durling, Rodrigo Guimaraes, Frances Hill, Tomas Key, Graham Logan, Michaela Morris, Michael Osbourne, Kate Reinold, Paul Robinson, Ayşegül Şahin, Ben Sole, Vincent Sterk, anonymous reviewers, and conference participants at the European Economic Association, the American Economic Association, Federal Reserve Board of Governors, Bank of England, ONS, University of Oxford, and other seminars for comments. Special thanks to William Abel and David Bradnum. Published as "Is the UK's productivity puzzle mostly driven by occupational mismatch? An analysis using big data on job vacancies." *Labour Economics* (2021): 102013.

†Office for National Statistics, UK, and Bank of England, UK.
‡Bank of England, Threadneedle St, London, EC2R 8AH, UK.
§Bank of England, Threadneedle St, London, EC2R 8AH, UK.
¶NESTA, 58 Victoria Embankment, London, EC4Y 0DS, UK.
‖Natwest, 175 Glasgow Road, Edinburgh, EH12 1HQ.

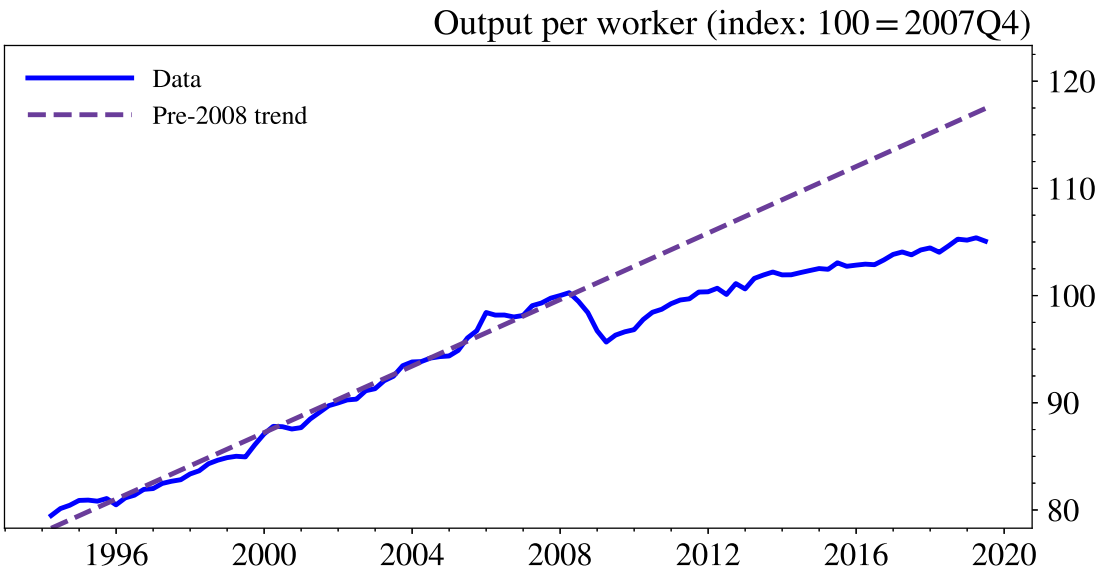

Figure 1: The aggregate output per worker in the UK (seasonally adjusted). Trend lines are fit using data from 1990Q1–2007Q4. Source: ONS.

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 (Barnett et al., 2014b; Bryson and Forth, 2015; Haldane, 2017), as shown in Figure 1. There has been controversy over what has caused this dramatic change in the behaviour of productivity. One explanation of the puzzle, introduced by Patterson et al. (2016), is that occupational mismatch between workers and available jobs is causing a significant fraction, up to two-thirds, of the productivity puzzle.

This paper uses new, big data on job vacancies to show occupational mismatch does not in fact explain a significant proportion of the UK productivity puzzle. Our vacancy data side-steps issues with the administrative vacancy data previously used to look at the contribution of mismatch to the productivity puzzle and we also include all of the UK in our analysis. We use a combination of vacancy stock data and official data to construct counter-factual paths for output, productivity, and employment that show what would have happened in the absence of occupational mismatch.

Using new data, we fundamentally challenge the view that the productivity puzzle is largely driven by occupational mismatch. We find no evidence of mismatch contributing to the productivity puzzle. Our results are driven by significant heterogeneity in how easily workers are matched to jobs across occupations.

The UK productivity puzzle remains stubbornly difficult to solve. A range of mismeasurement or output revision issues have been offered as explanations, 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. (2014a) 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. The UK's high wage flexibility has also been touted as a possible cause Blundell, Crawford and Jin (2014); Pessoa and Van Reenen (2014): 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. Cyclical explanations, including credit constraints (Riley, Rosazza-Bondibene and Young, 2014) and labour hoarding (Martin and Rowthorn, 2012), struggle to explain the long duration of below trend productivity growth.

The analysis of mismatch in the labour market has a long history that begins with the indices of Nickell (1982), Lilien (1982), and Jackman and Roper (1987). The seminal work on how mismatch can negatively affect aggregate labour market outcomes is by Şahin et al. (2014) and looks at how mismatch can add to unemployment. This framework gives counter-factuals for unemployment, productivity, and output in a scenario in which a social planner can optimally assign jobseekers to reduce mismatch. 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. We also adopt this mismatch framework.

The closest paper to ours is Patterson et al. (2016) who use the framework of Şahin et al. (2014) to show that the difference between the counter-factual and realised paths for productivity explain a significant fraction of the UK productivity puzzle between 2007 and 2012; namely up to two thirds of the deviation from the trend-growth in labour productivity. A substantial boost in aggregate output is also shown to result from eliminating occupational mismatch.

We make several key contributions relative to the existing literature. We revisit previous work by Patterson et al. (2016) suggesting that occupational mismatch is a major contributor to the UK's productivity puzzle, finding a different conclusion. We extend the previous analysis to cover the whole of the UK[1] and use a measure of unemployment based on designated national statistics rather than

[1]The job vacancy data used in Patterson et al. (2016) are from JobCentre Plus, available from https://www.nomisweb.co.uk/, and cover Great Britain but not Northern Ireland.

the less accurate Jobseeker Claimant Count. We demonstrate how naturally occurring[2] big data on online job ads can be used for macroeconomic labour market analysis. We create new estimates of matching elasticity and matching efficiency by occupation. Finally, we uncover significant labour market heterogeneity across occupations that was not apparent in similar previous work.

The structure of the paper is as follows: Section 2 describes the online job vacancies data and the other data we use to construct counter-factuals, Section 3 describes the search and matching theory we use and the mismatch theory of Şahin et al. (2014), and Section 4 presents results. Results are split into the estimation of the matching function (Section 4.1), counter-factual simulations (Section 4.3), and accounting for differences with the results of Patterson et al. (2016) (Section 4.4). Section 5 concludes.

2 Data

We use several datasets from the UK's Office for National Statistics (ONS), including the *Labour Force Survey* (LFS) (Office for National Statistics, 2017), the *Vacancy Survey*, and sectoral productivity measures.

Our measures of the number of people transitioning from unemployment to employment (ie employment flows), or vice versa, come from the 2-quarter longitudinal LFS, while the counts of those currently unemployed and employed come from the cross-sectional LFS.[3] Both are measured within market segments, for instance sector or occupation, as necessary. The LFS is the source of official UK national statistics on employment.[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. There is no occupational productivity measure so we construct one from (with occupation labelled here by μ)

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

[2]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

[3]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. To get the measure of unemployment by SOC code, we ascribe job seekers to occupations based on their previous job.

[4]It is, however, a survey of addresses, which contributes a potential downward bias in the rate of job-finding: if someone moves job *and* moves address at the same time, they are lost from the survey. However, the ONS take measures to reduce the extent of this bias with their weighting of the data and, although this is a downward bias, the average rate of flow into employment according to the LFS is higher (at the 1-digit occupational level) over the period we study than for other measures that have occasionally been used such as JobCentre Plus vacancy outflows–suggesting that this downward bias cannot be the source of any significant difference with previous work.

so that the value-added by occupation μ at time t 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 $15,242,000$ individual jobs posted at daily frequency from January 2008 to December 2016 online at Reed.co.uk. Online vacancy data, such as those obtained from Reed, can add value to economic analysis of the labour market because of their timeliness, granularity, extra fields relative to other sources, and by virtue of them being a direct measurement of economic activity (as opposed to an indirect one, such as a survey). Here we make use of the granularity and extra fields to perform analysis that would not be possible with the only other data source that is available for the same period we study (the ONS' Vacancy Survey). Specifically, the Reed vacancies data allows us to look at the demand for labour through the lens of occupations and to a high degree of granularity, down to 3-digit Standard Occupational Classification (SOC) codes.

We employ the methodology of Turrell et al. (2019) for using job title and job description text to classify online vacancies according to UK Standard Occupation Classification (SOC) codes.[5] We choose to label the vacancies with SOC codes so that we can combine them with labour market data on employment and unemployment that also uses SOC codes. This allows us to estimate labour market matching functions and run counter-factual simulations using consistent definitions for sub-markets (here, occupations).

To apply occupational labels to the job vacancies following Turrell et al. (2019), we make use of text fields in the raw data that capture job title, job description, and job sector. We also use official documentation of the SOC taxonomy, including a list of all known possible job titles and a short official description for each SOC code. We aggregate this text at the 3-digit SOC code. We then solve a matching problem; we wish to find the SOC code of the official text that is 'closest' to the text from the job ad. We do this in three steps following standard text cleaning. First, if there are exact matches between job ad titles and official jobs, we accept those matches. Second, we use term frequency-inverse document frequency (tf-idf) vectors to represent the SOC code strings with a matrix with dimension $T \times D$ where t is a term from the text associated with a SOC code and d is the number of SOC codes. Our terms are comprised of all 1–3-term long phrases excluding words that are not informative, known as 'stop words'. We express each job vacancy as a vector in the vector space defined by the official term-frequencies, and then take the five SOC codes with the largest cosine similarity to each job ad.

[5] See http://github.com/aeturrell/occupationcoder for computer code.

Finally, to choose between the top five SOC codes from the tf-idf step, we use the Levenshtein distance (Levenshtein, 1966). In a validation exercise conducted with the UK's Office for National Statistics, this method agreed with their automated assignment procedure on 91% of records that they were able to give labels to, and a much smaller sample had a 76% accuracy versus human coders (which is nevertheless high relative to pre-existing algorithms).

The most similar vacancy data to the Reed data we use are the UK Jobcentre Plus (JCP) 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. They are known to suffer from severe bias including exhibiting a disproportionate share of manual and so-called low-skilled jobs (Burgess and Profit, 2001), over-representation of some sectors as noted in Patterson et al. (2016), and large variations between regions, sectors, and over time due to the policies of individual Jobcentre Plus offices (Machin, 2003). They 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).

Problems with JCP data included that a significant percentage of the entire vacancy stock was not always updated when filled or withdrawn by employers. This had the effect of biasing the stock upwards by numbers as high as the multiple tens of thousands out of vacancies in the few hundreds of thousands. The number of ways for firms to communicate to JCP offices increased at that time, leading to structural breaks in the series, and the reliance on firms to notify JCP offices when vacancies were filled or withdrawn made the outflow series, and therefore the stock, vulnerable to bias.

Previous work using job vacancies to examine mismatch in the labour market used JobCentre Plus data without any correction for its biases. Here, we bring a different measure of vacancies–from Reed–to the same question, and we use an adjusted version of it that attempts to correct for some of the likely biases of online vacancy data.

In the US, the most similar datasets to the vacancy survey and the Reed measure of vacancies are the Job Openings and Labor Turnover Survey and the Conference Board Help Wanted Online series respectively.

It is important to note that online job vacancy data cannot be expected to be perfectly represen-

tative over the time period we study here. The unweighted Reed series accounts for around 40% of UK vacancies annually (see Fig. 2). The unweighted stock will have, like the JobCentre Plus data, many biases relative to the best available aggregate measure, the ONS Vacancy Survey. Because of these biases, we reweight the Reed data using the methodology of Turrell et al. (2019), which corrects for some–but not all–of the biases.

There are three major sources of bias we may worry about in the unweighted data: aggregate stock bias, online representativeness bias, and occupational bias. The first is just whether, on aggregate, there are as many vacancies captured by Reed as there are in reality. The second is the bias that arises because some job vacancies are less likely to be advertised online. Finally, the third bias arises because the distribution of occupations, the dimension most relevant to our analysis of mismatch, is different in Reed data versus reality.

We reweight the Reed data using the monthly sectoral (via the Standard Industrial Classification) disaggregation of the Vacancy Survey and the fact that the Reed monthly stock of vacancies also has a sectoral breakdown. Their ratios are used as weights. Reweighting can almost completely eliminate any aggregate vacancy stock bias. It will reduce the online representativeness bias and differential occupational representativeness bias only to the extent that sectoral differences are correlated with these other compositional differences. In the reweighting, 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. Throughout the rest of the paper, we use the weighted data.

The aggregate time series of the Vacancy Survey, raw Reed stock of vacancies, reweighted Reed vacancy stock, and JobCentre Plus Statistics are shown in Figure 2. The Reed data is better correlated to the Vacancy Survey than the JobCentre Plus data; the correlations between the weighted and unweighted Reed vacancy series, and the Vacancy Survey, are 0.97 and 0.99 respectively, compared to 0.75 for the (seasonally adjusted) JobCentre Plus data. The trends in job vacancies posted online match the trends in job vacancies more broadly. The JCP series has a poorer correlation with the Vacancy Survey than even the raw Reed data. Finally, despite its known biases, JCP was used with no reweighting in the mismatch study that found occupational mismatch *did* cause the productivity puzzle–which justifies a re-analysis of the same question with an independent data source.

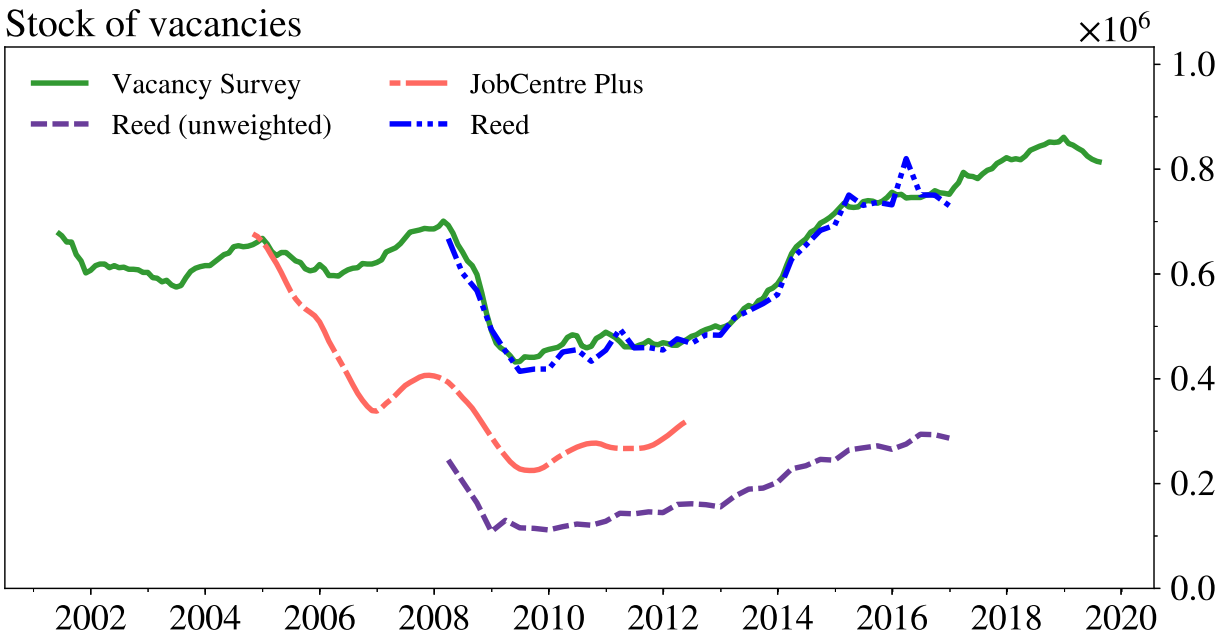

Figure 2: The aggregate stock of vacancies from three data sources. Source: Reed, ONS, National Online Manpower Information System (NOMIS).

3 Theoretical Framework and Evidence of Mismatch

3.1 Theoretical Framework

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).

Mismatch occurs when barriers that prevent worker mobility between occupations lead to misallocation of labour and longer unemployment durations than would otherwise be experienced. By slowing the rate at which the unemployed find jobs, mismatch has a direct negative effect on aggregate employment. A central planner that is able to optimally allocate job seekers across the labour market so as to increase the job finding rate will be able to achieve a higher level of aggregate employment, and therefore output, than the decentralised equilibrium. Mismatch also affects aggregate productivity but in a theoretically ambiguous way. In particular, it is not obvious that the social planner should necessarily

reallocate job seekers from low to high productivity sectors if expected unemployment duration in the high productivity areas of the labour market is much greater than for other less productive sectors. That would result in workers sitting idle for longer periods, with output commensurately lower.

In their assessment of how important mismatch of this type is quantitatively, Patterson et al. (2016) find that it has a large negative effect on employment, output, and productivity. In their model, the social planner eliminates mismatch by optimally reallocating unemployed workers across different occupations in the economy in each period, raising aggregate employment in the process and gradually achieving a more productive distribution of the workforce. In this paper, we follow the same conceptual set-up. But our focus is on how our new data sources affect estimates of the structural parameters of the model, and how that in turn influences the balance of trade-offs facing the social planner in an otherwise similar counterfactual experiment to Patterson et al. (2016).

A basic version of the Diamond-Mortensen-Pissarides (Diamond, 1982; Mortensen and Pissarides, 1994) search-and-match theory of the labour market reveals more details of how mismatch can lower output. We assume a matching function M that 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. Matches and new hires from unemployment are equivalent.

At the disaggregated level, hires are given by h_i for occupation i. Define output per worker in occupation i by z_i, which we take to be exogenous (see Section 2). Hires based upon the theoretical matching function and an occupation-specific matching efficiency are given by

$$h_{it} = \phi_i M(U_{i,t-1}, V_{i,t-1}) = \phi_i U_{i,t-1}^{1-\alpha} V_{i,t-1}^{\alpha}. \tag{2}$$

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 (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). This theory assumes that a social planner can re-allocate jobseekers between occupations subject only to the within-market matching frictions present in each sub-market. Therefore, as stressed by Şahin et al. (2014), the resulting estimate of mismatch unemployment is an upper bound. We maintain this assumption both for simplicity and also for comparability to Patterson et al. (2016). However, we also follow Patterson et al. (2016) in assuming that new hires are temporarily less productive than already-matched workers, which goes some way to capturing match-specific skills loss during reallocation. Other contributions to the literature consider much richer measures of mismatch that account for multiple skill dimensions and human capital accumulation (Guvenen et al., 2020; Lise and Postel-Vinay, 2020). Extending the model along such lines would be more empirically plausible than assuming that workers can move freely across heterogeneous jobs; however, it would also complicate the analysis and make it harder to make direct comparisons with previous work, so we do not pursue such an extension here.

Given I market segments, the Şahin et al. (2014) 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 occupational 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 occupation 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 occupation 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. To ensure consistency with Patterson et al. (2016) 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 A, and the allocation to each occupation 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 $e_{it}^* = (1 - \xi_{t-1})e_{i,t-1}^* + h_{it}^*$. 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. In simulations, we use heterogeneous job destruction rates from the LFS.

3.2 Evidence of Mismatch

Previous work has documented an increase in mismatch beginning in 2008 in both the UK and US as a result of the Great Financial Crisis. In the UK, this coincided with an increase in unemployment that saw the rate climb from 5.2% in quarter 1 of 2008 to a peak of 8.4% in 2011. The rise in mismatch is well-documented. Patterson et al. (2016) show several mismatch indices that begin to climb in the last quarter of 2007. Smith (2012) documents that half the rise in UK unemployment over this period was due to mismatch. Likewise, Şahin et al. (2014) document that the US economy experienced an increase in mismatch around the same period that accounted for a significant part of unemployment.

Our data begin at the end of 2008Q1, but similarly show an increase in mismatch at the beginning of the period we study. To show this, we use the simple but interpretable Jackman-Roper index of labour force mismatch (Jackman and Roper, 1987), which is given by

$$I = \frac{1}{2}\frac{U}{L}\sum_j \left| \frac{U_j}{U} - \frac{V_j}{V} \right|,$$

where j indicates a sub-market, in this case occupation, and L is the total size of the labour market.

Figure 3 shows the Jackman-Roper index for our data, which increased sharply between 2008 and 2009.

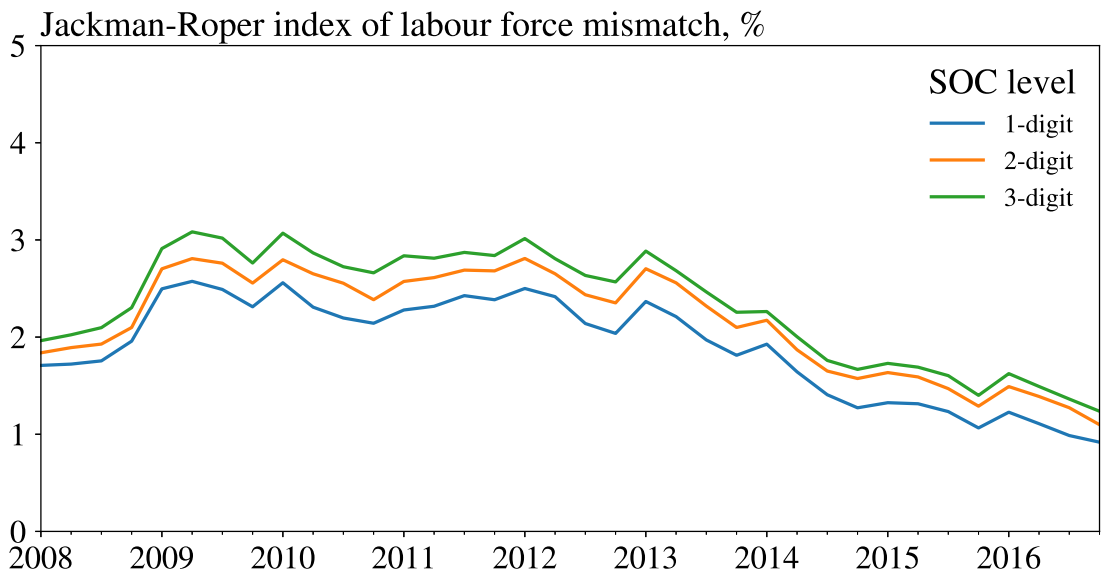

Figure 3: Jackman-Roper index of mismatch of vacancies and unemployed relative to size of the labour force (Jackman and Roper, 1987).

4 Results and Discussion

4.1 Matching Function Estimation

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 Section 3. 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.

One advantage of our approach of assigning conventional SOC codes to online job adverts following Turrell et al. (2019) 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, previous work by Patterson et al. (2016) use the average of vacancy off-flows (from JobCentre Plus data) and jobseeker allowance claimant off-flows (from the National Online Manpower Information Service, or NOMIS) as the dependent variable. These are proxies for labour market transitions. The longitudinal LFS is a weighted and representative sample of employment flows at the UK national level (Jenkins and Chandler, 2010).

We adopt a matching regression specification that assumes the segmented labour market discussed in Section 3, 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. This is a standard specification in the literature (Petrongolo and Pissarides, 2001). 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 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.

In Table 1, we report matching function estimates for data disaggregated to the 1-, 2- and 3-digit SOC 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 0.47, which is in the range described as 'plausible' by Petrongolo and Pissarides (2001).[6] Our estimates are extremely close to those reported by Patterson et al. (2016),

[6]Note that the range they give is the elasticity on unemployment, our $1 - \alpha$.

Table 1: Matching function parameter estimates. Source: Reed, ONS.

	1-digit SOC		2-digit SOC		3-digit SOC		Aggregate	
	IV	OLS	IV	OLS	IV	OLS	IV	OLS
Elasticity parameter (α)	0.490***	0.477***	0.451***	0.477***	0.422***	0.526***	0.522***	0.538***
	(0.038)	(0.022)	(0.038)	(0.024)	(0.023)	(0.017)	(0.026)	(0.027)
Observations	315	324	867	892	2639	2729	35	36
Cross-sections	9	9	25	25	90	90	1	1

who find OLS estimates of 0.559 and 0.463 for the aggregate and 2-digit SOC level respectively.

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 1 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.[7]

The matching function parameters are estimated precisely and the overall fit is especially good at the aggregate and 1-digit SOC code levels. The adjusted R^2's are 0.94, 0.82, 0.64, and 0.52 for OLS estimates at aggregate, and 1-, 2-, and 3-digit SOC levels respectively. Based on previous work, such as Yashiv (2000), adding information on jobseekers or parameters that capture the macroeconomic context in other ways would likely be needed to improve the overall fit more.

Using the aggregated data, we plot the fitted Beveridge curve in Figure 4 against the vacancy-unemployment points at quarterly frequency. We calibrate the job destruction rate in the Beveridge curve to give the best fit to the data while using the aggregate matching efficiency and aggregate elasticity from Table 1. Arrows indicate movements over time, and a shift toward higher unemployment during the Great Recession is evident, as is the unprecedentedly high tightness in the last quarter of 2016.

Figure 5 plots Beveridge curves and quarterly u-v points for each 1-digit SOC code. These curves use the matching efficiency estimates obtained from the regressions in Table 1. The sub-market level Beveridge curves show that a single, aggregate Beveridge curve hides a great deal of important variation in u-v space across SOC codes. There are significant differences between the apparent curves as

[7] This is consistent with Barnichon and Figura (2015), who also find that an instrumental variables approach differs little to their OLS estimate of the matching function elasticity.

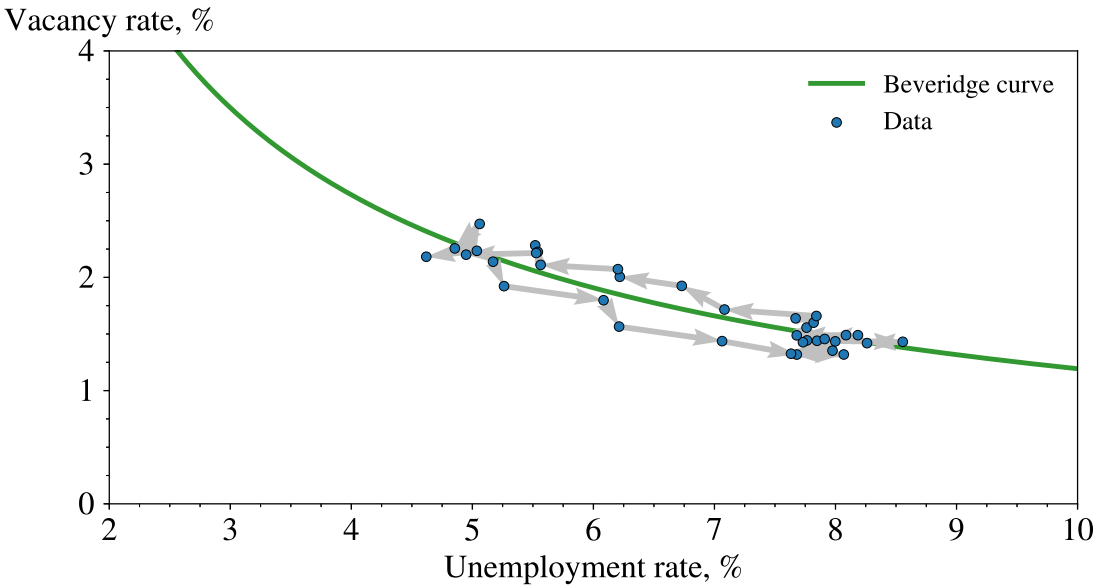

Figure 4: Beveridge curve (line) using estimates in Table 1 versus aggregate u-v data at quarterly frequency. Source: Reed, ONS.

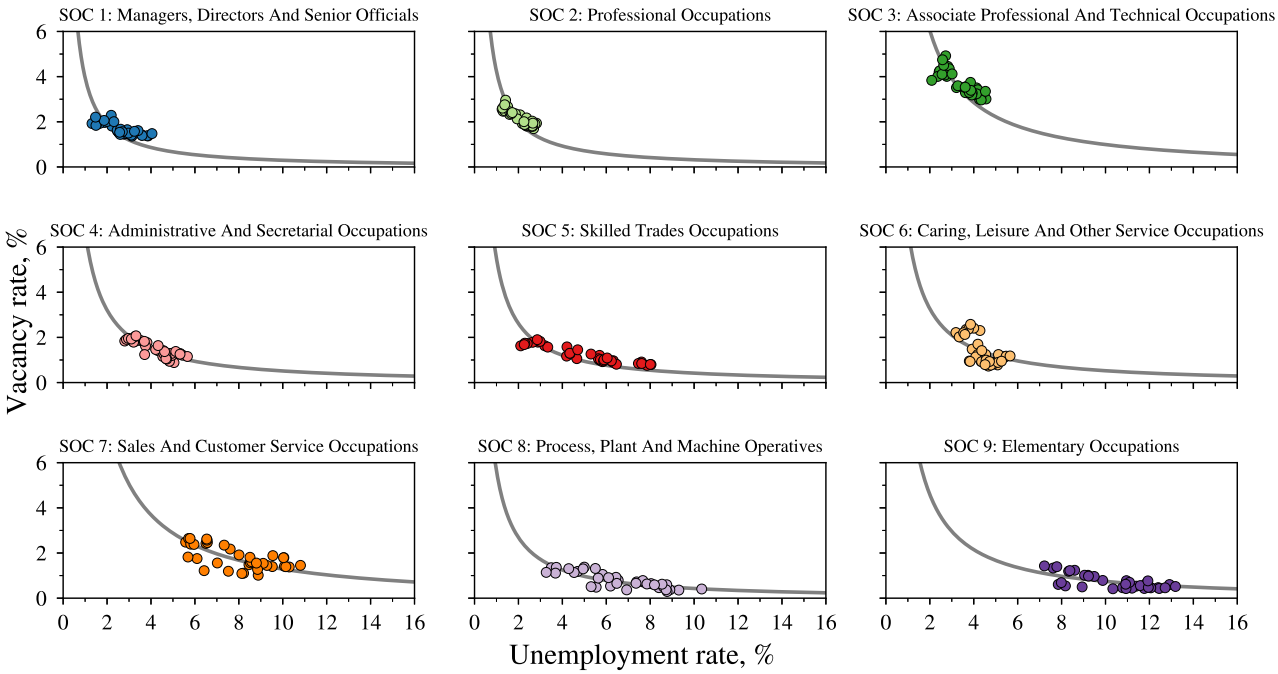

Figure 5: Beveridge curves (lines) using estimates of the parameters in equation (4) and data (points) in u-v space for each 1-digit SOC code at quarterly frequency. Source: Reed, ONS.

separated by SOC ranking, with the curve for associate professional and technical occupations shifted up relative to other occupations. There are also differences in spread; generally, the lower the SOC number of 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. 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. This means that compositional changes in jobs will have an effect on the aggregate rate of outflow, in turn affecting the amount of slack in the economy.

4.2 Productivity and matching efficiency

As noted in Section 3, 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 (2013), both within sectors and across sectors (Barnett et al., 2014b; Broadbent, 2012; Haldane, 2017).

Figure 6 shows that matching efficiency estimates, found according to the regressions in Section 4.1, 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 lower numbered SOC codes 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 is consistent with Şahin et al. (2014) who find that matching efficiency is lowest for management, professional and related occupations and that the average number of years of education of workers in lower numbered 1-digit SOC codes is higher, implying a higher level of specialisation.

Figure 6 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 rises with increasing SOC code number, as opposed to productivity, which broadly falls.

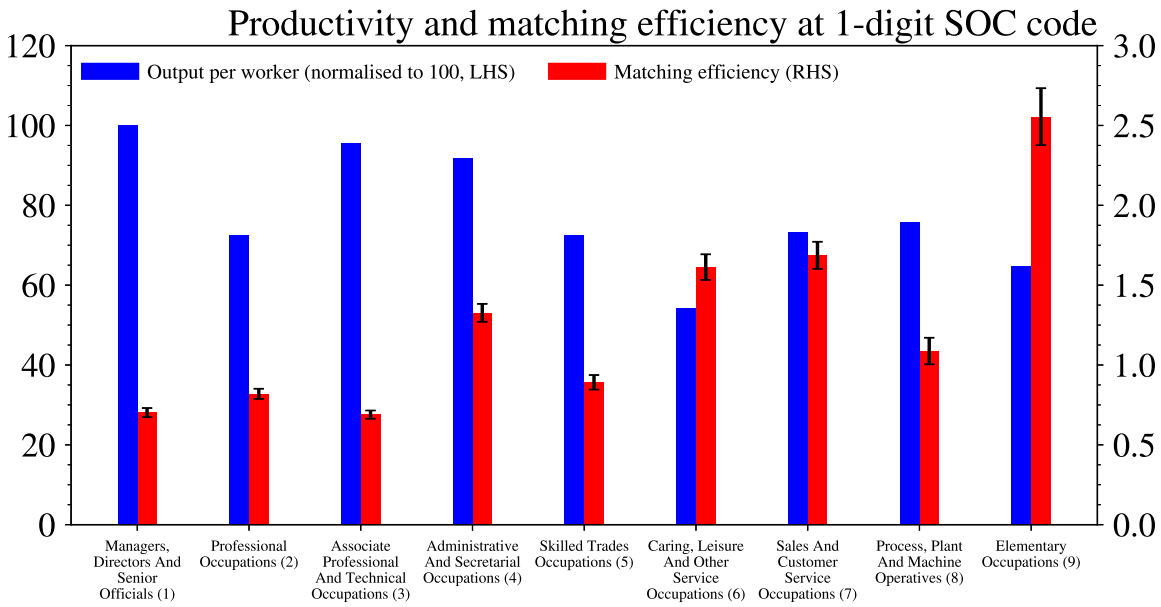

Figure 6: Estimates of productivity (left-hand y-axis) and of the matching efficiency (right-hand y-axis) by 1-digit SOC code. Standard errors are shown for the estimates of the matching efficiency. Source: ONS, Reed.

4.3 Counter-factual Simulations

We run simulations of counter-factual paths for employment, output, and productivity using the matching theory described in Section 3 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 into, employment.

Both our model and matching function estimation abstract from labour force participation flows, which means our counterfactuals do not straightforwardly accord with the true paths taken by employment, output and output per worker. In light of this, we begin counter-factual paths for output, productivity, and employment 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 the relevant counter-factual variable.

Figure 7 shows the social planner's optimal path for output using three different simulations for three different levels of SOC code. Counter-factual paths for employment and output would be around 4 and 0 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

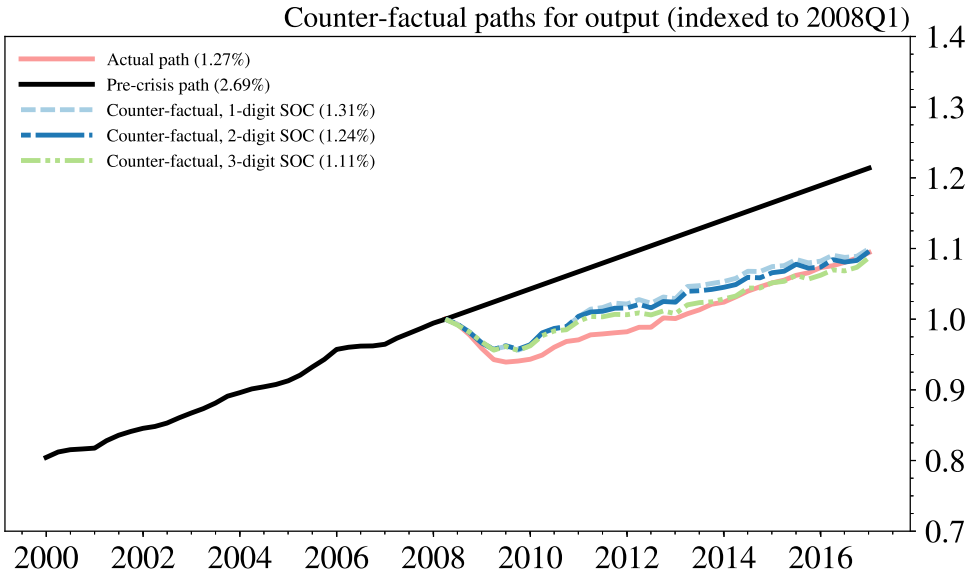

Figure 7: Realised and counter-factual paths for output. Simulations are run at different SOC code levels. Legend growth rates refer to the mean year-on-year growth rates measured quarterly. The shaded line is the pre-crisis trend extrapolated to the post-crisis period.

perhaps implausibly large reduction in the unemployment rate, with the almost complete elimination of structural unemployment. Taking this into account, the estimates of output and productivity growth should be seen as upper bounds, and the conclusion is that even with the most optimistic of scenarios for employment, the extent to which output can be raised is small.

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. The counter-factual path is substantially smaller than the one found in Patterson et al. (2016), which sees output per worker make up more than half the difference to the pre-crisis productivity trend.

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 (8) of Appendix A). As Figure 6 shows, though low SOC code jobs tend to have higher productivity, they also 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 occupation confirm this: at the end of the counter-factual simulation by 1-digit SOC code, although employment overall is higher, the SOC category that includes managers (1) accounts for proportionally fewer employees than those in so-called elementary occupations (9) in

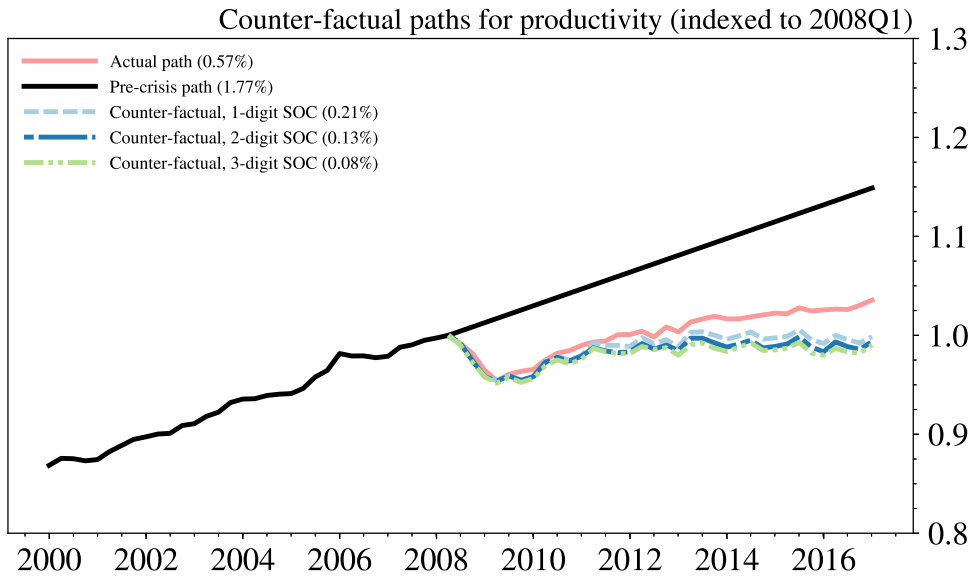

Figure 8: Realised and counter-factual paths for productivity. Simulations are run at different SOC code levels. Legend growth rates refer to the mean year-on-year growth rates measured quarterly. The shaded line is the pre-crisis trend extrapolated to the post-crisis period.

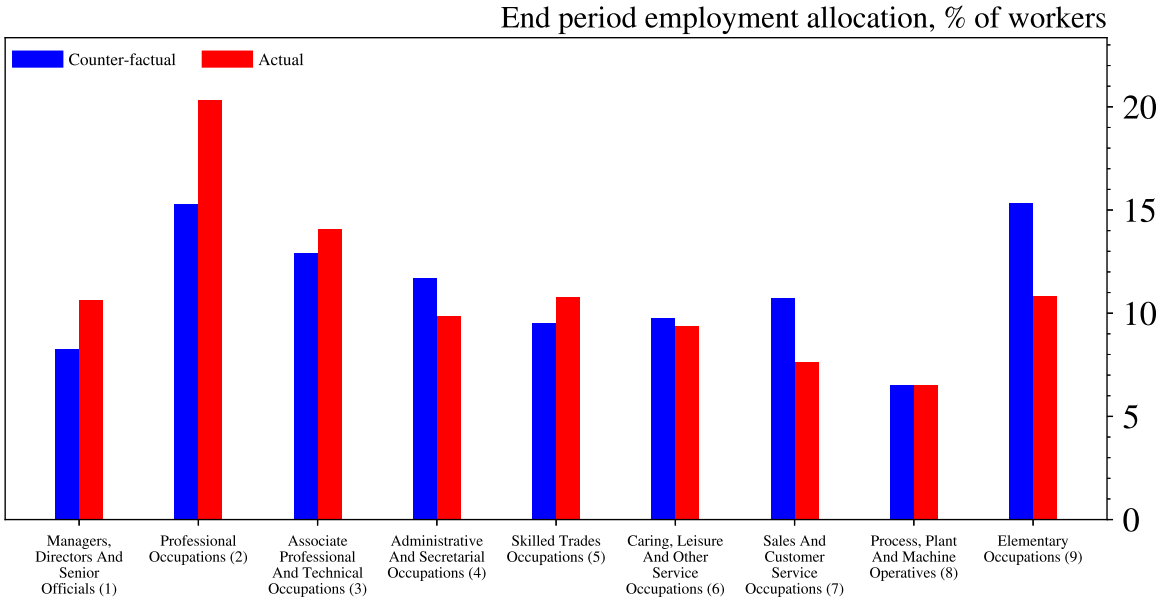

Figure 9: Realised and counter-factual final period employment rates by 1-digit SOC code.

the actual distribution of employment, as shown in Figure 9. Further simulations, in Appendix B, 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.

Unfortunately, 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. Both of these have the potential to increase $\bar{u}$, the steady state of unemployment, given fixed job destruction rates and tightness. Previous results, such as 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. 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 appreciable in 2011, the paths had shown signs of convergence by the end of 2017.

4.4 Accounting for differences with Patterson et al. (2016)

There are many potential sources of difference between the results we find and those of Patterson et al. (2016) that could lie at the root of the two differing narratives they uncover about the UK labour market. In Patterson et al. (2016), the conclusion is that mismatch accounted for more than half of the productivity puzzle by 2012. Using different data and an only partially overlapping time period, our analysis finds that mismatch accounts for none of the productivity puzzle. Here, we dig deeper into the main factors behind the difference in conclusions. We begin with differences in input data; these are summarised in Table 2.

First, the time periods are different; 2006–2012 versus 2008–2017 here. This is important because any benefits of re-allocation can accrue over time with the churn of workers. As our data only begin in 2008, we are necessarily restricted to 2008 onwards, but this is also when productivity first diverged

Variable	Patterson et al. matching est.	Patterson et al. simulations	This paper: matching est. and simulations
Hires	Average of jobseekers' allowance claimant outflows and JCP vacancy outflows (pre-recession)	LFS flows (2006–)	LFS flows (2008–)
Job destruction rates	NA	LFS flows (2006–)	LFS flows (simulations, 2008–)
Vacancies	JCP vacancy data (unweighted; pre-recession)	JCP vacancy data (unweighted; 2006–)	Reed vacancy data (weighted; 2008–)
Unemployment	Jobseekers' allowance claimants (pre-recession)	Jobseekers' allowance claimants	LFS unemployment (2008–)

Table 2: Summary of differences in data used in the two studies.

from trend as shown in Figure 1.

Second, there are small differences in the productivity data. These are likely due to revisions in both the estimates of output and the weighting given to individuals in the Labour Force Survey.

Third, the previous work used different data for estimating the matching function and in simulations of counter-factuals. In contrast, we use the LFS measures of activity in the labour market and the Reed vacancy data for both matching function estimation and simulations, for consistency. Patterson et al. (2016) use some LFS data for the counter-factual simulations but otherwise (including for estimating the matching function) use time series that proxy labour force variables indirectly.

As their measure of unemployment, Patterson et al. (2016) use the jobseekers' allowance claimant counts rather than LFS-unemployment that we use. The former is biased downwards relative to total unemployment as not all unemployed meet the eligibility criteria. The latter is also imperfect, as it ascribes job seekers to their previous occupation.

The measure of hires we use is based on flows into employment from the longitudinal LFS, which is designed to be representative of the UK. In Patterson et al. (2016), the average of the outflow of JCP job vacancies and the outflow of jobseeker's allowance claimants was used. These give quite different measures of hires, as shown in Figure 10–the average is likely to be an under-estimate because not all job vacancies (or types of vacancies) were advertised on JCP, and nor were all those who stopped claiming jobseekers' allowance the same as all who were hired.

The measure of vacancies is also different: JCP versus Reed. The JCP vacancy data do not include Northern Ireland and suffer from biases that caused them to be dropped as national statistics. Figure 11 compares the distributions of these two measures of UK vacancies at the 1-digit SOC code level. We show them alongside the distribution of employment according to the LFS.

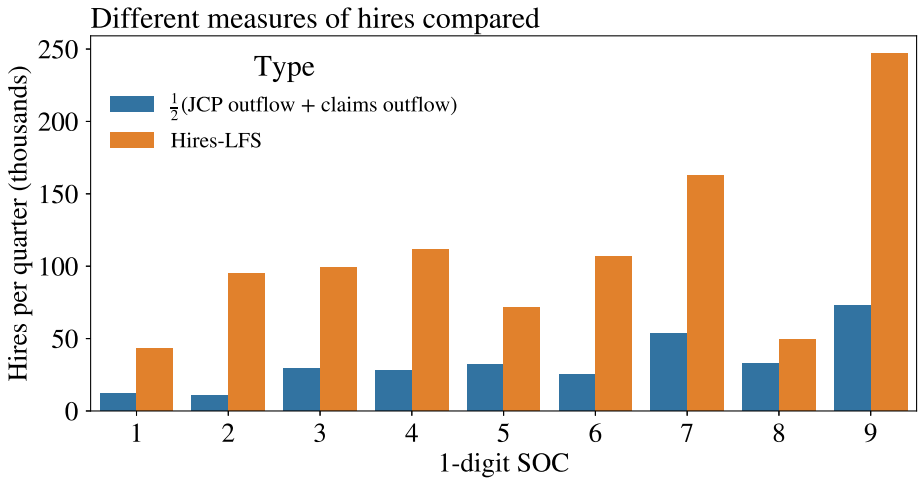

Figure 10: Mean hires per quarter according to different measures of hires, at 1-digit SOC code level.

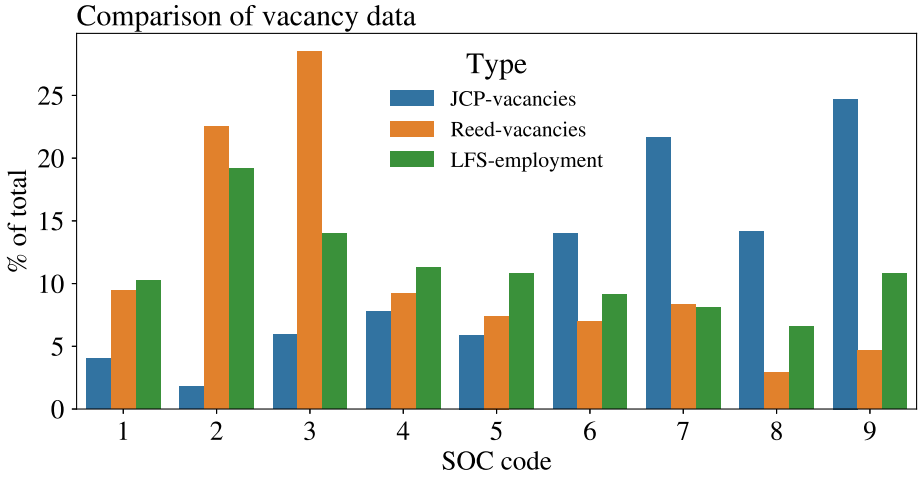

Figure 11: The average distribution of vacancies of JCP and Reed datasets, shown alongside the stocks of employment from the LFS.

We can see from Figure 11 that the JCP data heavily over-represent SOC codes 6, 7, 8, and 9, and heavily under-represents codes 1, 2, and 3, relative to employment stocks. The distribution of Reed vacancies is closer to the distribution of total employment in the LFS, though with an over-representation of SOC code 3. Overall, Reed slightly under-represents lower productivity SOC codes, while JCP chronically under-represents higher productivity SOC codes.

We find that much of the difference in the conclusions that can be drawn from counter-factual simulations between our study and that of Patterson et al. (2016) are determined by differences in estimated match efficiencies (themselves determined by the use of different measures of hires, unemployment, vacancies, and time periods). We show this directly in Figure 12. This is a simulation that uses the distribution of matching efficiencies from Patterson et al. (2016) but uses data from our study otherwise.[8] Figure 12 shows that by using the same distribution of match efficiencies as in Patterson et al. (2016), our simulations can reproduce their result that mismatch accounted for more than half of the productivity puzzle up to 2012. This result persists, and is even stronger, at the 3-digit SOC level.

Importantly, even where we adjust matching efficiencies to create an artificial boost in our counter-factual path for productivity, the elimination-of-mismatch effect begins to wane after 2012. This suggests that other factors are largely driving the long-term productivity puzzle.

We consider another scenario in Appendix B, in which we run counter-factual simulations where all 1- or 2-digit SOC codes have the same matching efficiency (the average across all SOC codes). This is shown in Figure B.1. This constant-match-efficiency experiment produces results that are closer to those of Patterson et al. (2016) for the period 2008–2012 but, again, the contribution of mismatch to the productivity puzzle weakens from 2013.

A combination of data differences are the root cause of the differences in overall results between this study and previous work; and these data differences cause matching efficiencies with a range of other specifications to be more heterogeneous than previously estimated, thereby driving the overall result that mismatch cannot account for the productivity puzzle.

Can we say which overall conclusion is more reliable from this? There are several reasons why we have confidence that these differences favour the conclusion that mismatch does not explain the productivity puzzle, based on our analysis.

First, the measures used in this study are likely to be less biased. Measuring hires directly from labour market flows rather than from data with known incomplete coverage should be more accurate.

[8] We use the distribution of match efficiencies directly from Patterson et al. (2016) replication package but rescale the level so as to be consistent with the Reed and LFS data.

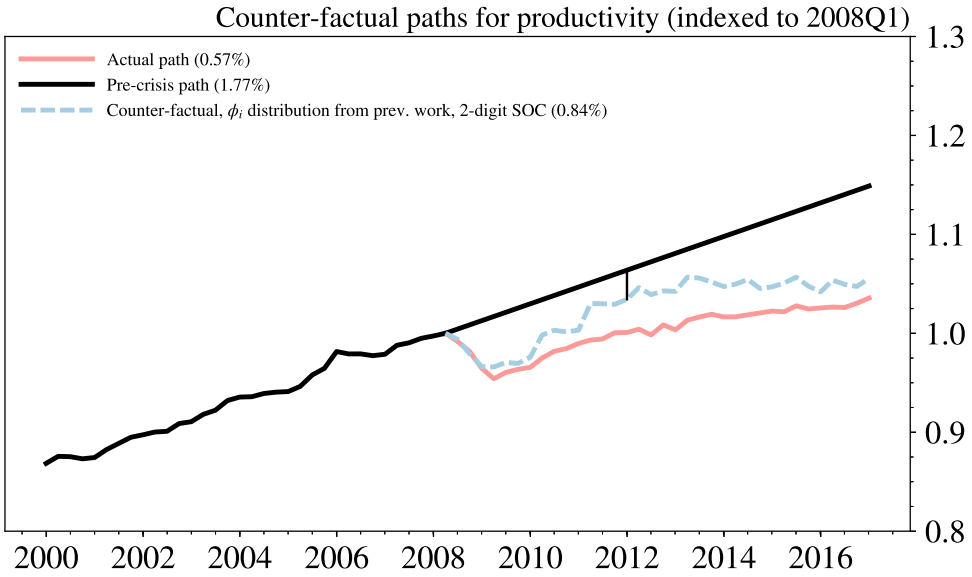

Figure 12: A simulation using the LFS + Reed data but with the distribution of match efficiencies taken from the 2-digit SOC level regression results of Patterson et al. (2016). The match efficiency for SOC code 33 was imputed as an average of the other match efficiencies.

And while the vacancy data we use is not perfect, it has been reweighted and extensively studied for biases, and is therefore more likely to be more reflective of the wider labour market.

Second, even when using the flatter distribution of match efficiencies found in previous work, the higher productivity path for the UK fades away after 2013–and the productivity puzzle is still very much with us. Mismatch is partly a cyclical explanation of the productivity puzzle, so it cannot account for the persistence of the puzzle through to the end of 2019.

Third, the finding that matching efficiencies are fairly homogeneous does not seem to be robust to alternative but plausible specifications, suggesting that match efficiencies are heterogeneous–and this heterogeneity is what causes our simulations to adhere to a lower productivity path. Relatedly, we use matching efficiencies from 2008 onwards, as opposed to from before the crisis.

Fourth, our results for productivity counter-factuals are robust to changes in SOC level. This was less true in the original analysis, where end outcomes between 1- and 2-digit SOC levels differed by as much as four percentage points.

Finally, it is well-known that even on the *same* data, both the choice of specification and the conclusions reached can differ markedly (Huntington-Klein et al., 2021; Silberzahn et al., 2018). Here, we are bringing new data sources, different time periods, revised data, and the benefit of hindsight to our analysis, so it is perhaps not surprising that the well-executed work of Patterson et al. (2016) and

this investigation tell a different story about the drivers of the UK's productivity puzzle.

5 Conclusion

We have used big data on online job vacancies in combination with official statistics to show that, contrary to previous findings using administrative data, occupational mismatch does not account for most of the UK's productivity puzzle. Additionally, we have provided new estimates of the structural parameters of the labour market matching function.

In general, 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.

One limitation of this work, and the literature it follows, is that it treats labour supply as homogeneous. In reality, although it is practical to consider re-allocation across the more granular occupational groupings, it is unlikely that jobseekers could successfully find work in completely a different occupation to that which they have previously worked in. As such, the estimates of what would happen in the absence of mismatch are an upper bound on how different aggregate labour market outcomes could be. New data on labour supply and the behaviour of jobseekers, perhaps from job search websites, could give tighter bounds on what is possible in terms of re-allocation. Another important limitation is that our work does not account for heterogeneity in location, which is likely to be an important factor for at least some pairs of regions in the UK. Finally, our social planner does not take into account several important factors that might cause them to allocate fewer jobseekers to lower productivity jobs, including such considerations as job desirability, long-term human capital accumulation within a job, and the risks of future technological disruption to some types of work. Future work could usefully develop models that take account of these factors.

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 for understanding the aggregate labour market and its potential.

References

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

Barnett, Alina, Sandra Batten, Adrian Chiu, Jeremy Franklin, Maria Sebastia-Barriel, et al. 2014b. "The UK productivity puzzle." Bank of England Quarterly Bulletin, 54(2): 114–128. 1, 15

Barnichon, Regis, and Andrew Figura. 2015. "Labor market heterogeneity and the aggregate matching function." American Economic Journal: Macroeconomics, 7(4): 222–249. 13

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. 11

Bentley, R. 2005. "Publication of JobCentre Plus vacancy statistics." ONS Reports, Labour Market Trends. 5

Blanchard, Olivier Jean, and Peter A Diamond. 1989. "The aggregate matching function." National Bureau of Economic Research. 13

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. 2

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. 13

Broadbent, Ben. 2012. "Productivity and the allocation of resources." Speech given at Durham Business School, 12. 15

Bryson, Alex, and John Forth. 2015. "The UK's Productivity Puzzle." CEPREMAP CEPREMAP Working Papers (Docweb) 1511. 1

Burgess, Simon, and Stefan Profit. 2001. "Externalities in the Matching of Workers and Firms in Britain." Labour Economics, 8(3): 313–333. 5

Coles, Melvyn G, and Eric Smith. 1996. "Cross-section estimation of the matching function: evidence from England and Wales." Economica, 589–597. 11

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. 9

Diamond, Peter A. 1982. "Wage determination and efficiency in search equilibrium." The Review of Economic Studies, 49(2): 217–227. 8

Field, Simon, and Mark Franklin. 2013. "Micro-data perspectives on the UK productivity conundrum – an update." ONS Reports. 15

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. 19

Guvenen, Fatih, Burhan Kuruscu, Satoshi Tanaka, and David Wiczer. 2020. "Multidimensional skill mismatch." American Economic Journal: Macroeconomics, 12(1): 210–44. 9

Haldane, A. G. 2017. "Productivity Puzzles." Bank of England speech given at the London School of Economics. 1, 15

Hall, Robert E, and Sam Schulhofer-Wohl. 2015. "Measuring job-finding rates and matching efficiency with heterogeneous jobseekers." National Bureau of Economic Research. 19

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. 2

Hunter, J. D. 2007. "Matplotlib: A 2D graphics environment." Computing in Science & Engineering, 9(3): 90–95.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. 2021. "The influence of hidden researcher decisions in applied microeconomics." Economic Inquiry, 1–17. 23

Jackman, Richard, and Stephen Roper. 1987. "Structural unemployment." Oxford bulletin of economics and statistics, 49(1): 9–36. 2, 10, 11

Jenkins, K, and M Chandler. 2010. "Labour market gross flows data from the Labour Force Survey." ONS Reports, Economic and Labour Market Review. 12

Levenshtein, Vladimir I. 1966. "Binary codes capable of correcting deletions, insertions, and reversals." Vol. 10, 707–710. 5

Lilien, David M. 1982. "Sectoral shifts and cyclical unemployment." Journal of political economy, 90(4): 777–793. 2

Lise, Jeremy, and Fabien Postel-Vinay. 2020. "Multidimensional skills, sorting, and human capital accumulation." American Economic Review, 110(8): 2328–76. 9

Machin, Andrew. 2003. "The Vacancy Survey: a new series of National Statistics." ONS Reports, National Statistics feature. 5

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. 5

Martin, Bill, and Robert Rowthorn. 2012. "Is the British economy supply constrained II? A renewed critique of productivity pessimism." ONS Reports. 2

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. 7, 8

Nickell, Stephen. 1982. "The determinants of equilibrium unemployment in Britain." The Economic Journal, 92(367): 555–575. 2

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. 3

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. 0, 1, 2, 3, 5, 8, 9, 10, 12, 17, 19, 20, 22, 23, 28, 31

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. 2

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. 7, 8, 11, 12, 19

Pizzinelli, Carlo, and Bradley Speigner. 2017. "Matching efficiency and labour market heterogeneity in the United Kingdom." Bank of England Staff Working Paper, 667. 9, 19

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. 2

Şahin, Ayşegül, Joseph Song, Giorgio Topa, and Giovanni L Violante. 2014. "Mismatch unemployment." The American Economic Review, 104(11): 3529–3564. 2, 3, 9, 10, 15, 28

Seabold, Skipper, and Josef Perktold. 2010. "statsmodels: Econometric and statistical modeling with python."

Silberzahn, Raphael, Eric L Uhlmann, Daniel P Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, et al. 2018. "Many analysts, one data set: Making transparent how variations in analytic choices affect results." Advances in Methods and Practices in Psychological Science, 1(3): 337–356. 23

Smith, Jennifer C. 2012. "Unemployment and Mismatch in the UK." 2, 5, 10, 28

Turrell, Arthur, Bradley J Speigner, Jyldyz Djumalieva, David Copple, and James Thurgood. 2019. "Transforming Naturally Occurring Text Data Into Economic Statistics: The Case of Online Job Vacancy Postings." National Bureau of Economic Research Working Paper 25837. 4, 6, 12

Wes McKinney. 2010. "Data Structures for Statistical Computing in Python." 56 – 61.

Yashiv, Eran. 2000. "The determinants of equilibrium unemployment." American Economic Review, 90(5): 1297–1322. 13

Is the UK's productivity puzzle mostly driven by occupational mismatch? An analysis using big data on job vacancies

Appendix

Arthur Turrell Bradley Speigner David Copple Jyldyz Djumalieva James Thurgood

A 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 Section 4.1 for plots).

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 n for employed or not employed respectively. The job destruction rate for a segment of the labour market i is given by

$$\xi_{it} = \frac{\sum_p p(e,n)_{it}}{e_i}, \tag{5}$$

while hires into i are given by

$$h_{it} = \sum_p p(n,e)_{it}. \tag{6}$$

Equations (5) and (6) are used, respectively, to define the flow out of, and into, employment.

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). 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:

$$\mathcal{M}_{yt} = \frac{y_t^* - y_t}{y_t^*} = 1 - \sum_{i=1}^I \left(\frac{z_{it}\phi_{it}}{X_t}\right)\left(\frac{v_{it}}{v_t}\right)^{\alpha}\left(\frac{u_{it}}{u_t}\right)^{1-\alpha}, \tag{7}$$

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. Patterson et al. (2016) gives counter-factual hires as $h_{it}^* = h_{it}/(1 - \mathcal{M}_{xt})$ where

$$\mathcal{M}_{xt} = 1 - \sum_i^I \left(\frac{\phi_{it}}{\varphi_t}\right)\left(\frac{v_{it}}{v_t}\right)^{\alpha}\left(\frac{u_{it}}{u_t}\right)^{1-\alpha}, \tag{8}$$

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

$$Y_t^* = \sum_i^I z_{it} e_{it}^* + y_t^*, \tag{9}$$

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.

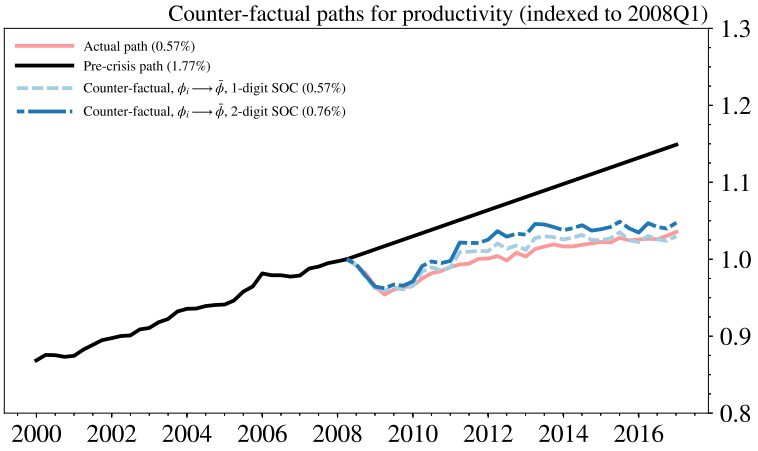

Figure B.1: Realised and counter-factual paths for productivity at the 1- and 2-digit SOC code level assuming all occupations have the same matching efficiency. The shaded line is the pre-crisis trend extrapolated to the post-crisis period.

B 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, Figure B.1 shows 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 some of the productivity puzzle.