Heterogeneity in the UK Labour Market: Using Machine Learning To Test Macroeconomic Models

Chris Martin* Magdalyn Okolo†

October 2, 2022

Abstract

We use machine learning techniques to investigate the main sources of heterogeneity in the UK labour market. A recent theoretical literature argues that accounting for differences in productivity between workers can help resolve some long-standing issues in the analysis of labour markets. This literature assumes that heterogeneity only reflects differences in productivity and not other types of heterogeneity, such as differences in age, sex, ethnicity, location, or the type of contract a worker has. We test this assumption. Applying a clustering algorithm to individual-level data, we find that cluster membership is mainly driven by productivity. This finding provides empirical support for the recent theoretical literature. Our results also imply that differences in productivity are systematic rather than purely random; that there is a strong relationship between education and productivity and that UK labour markets are not fully segmented.

Keywords: labour market heterogeneity; machine learning in Economics; UK labour market.

JEL Classification: E23, E32, J23, J30, J64

*corresponding author; cim21@bath.ac.uk; Department of Economics, University of Bath, Bath BA2 7AY UK. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

†m.u.a.okolo@bath.ac.uk; Department of Economics, University of Bath, Bath BA27AY, UK.

1 Introduction

The labour market is highly diverse. Until recently, macroeconomic models of the labour market did not take account of this diversity. This is changing, with a growing number of studies now arguing that accounting for differences between workers can help resolve some long-standing issues in the analysis of labour markets. These issues include the large volatility of unemployment relative to wages (Adjemian et al., 2021); the pattern of recovery from deep and shallow recessions (Hall and Kudlyak, 2021); the negative relationship between the duration of unemployment and the job finding rate of unemployed workers, and the impact of job loss of subsequent earnings (Gregory et al., 2021); and the slope of the Phillips Curve and the equilibrium rate of unemployment (Abriti and Consolo, 2022). These models all assume there are differences in productivity between otherwise identical workers[1]. Differences in productivity between workers are well documented. But other dimensions of heterogeneity in the labour market have also been highlighted, including the distinction between males and females (eg Hall and Kudlyak, 2020), between young versus older workers (eg Brotherhood et al., 2021), between different ethnicities (eg Duzhak, 2021), between temporary versus permanent jobs (eg Booth et al., 2002), and between fixed and zero hours contracts (Dolado, Lalé and Turon, 2021)[2]. Current theoretical models assume that heterogeneity in the labour market only reflects differences in productivity and does not reflect these other types of heterogeneity. In this paper, we test this assumption.

We also present evidence that helps resolve three important model design issues within the literature. The first is whether productivity is entirely random. Some models including Pries (2004), Gertler et al. (2020) and Faccini and Melosi (2021), assume that is, with productivity varying stochastically between "good matches" and "bad matches'. The second is whether productivity is related to education. Dolado, Motyovski and Pappa (2021), Adjemian et al. (2021), Abriti and Consolo (2022)) and Martin and Okolo (2022) assume that it is, whereas Gregory et al. (2020), Gregory et al. (2021) and Baley et al. (2022) assume it is not. The third is whether the labour market is segmented, so only more highly educated workers can do high productivity jobs; Dolado, Motyovski and Pappa (2021), Gregory et al. (2021), Adjemian et al. (2021) and Abriti and Consolo (2022)) assume that it is, while Grigsby (2021) and Martin and Okolo (2022) assume that more lowly educated workers can do high productivity jobs, and vice versa.

We obtain our results using data from the UK Labour Force Survey. This is a rich source of individual-level data on the dimensions of heterogeneity that have been highlighted in the literature, including sex, age, ethnicity, whether a job is permanent or temporary and whether the worker has a zero-hours contract. However, these data do not include a measure of productivity, and so we use proxies for this. The theoretical literature suggests two proxies[3]. The first is occupation, since the literature typically assumes that workers

[1]These models assume the surplus from job matches is smaller for lower productivity workers. A smaller surplus leads to a higher volatility of unemployment relative to wages for these workers, Ljungqvist and Sargent (2017), Adjemian et al. (2021) and Abriti and Consolo (2022), so lower productivity workers account for a large share of the overall volatility of unemployment. The small surplus implies that job matches with lower productivity workers break down more readily, so the flow from employment to unemployment contains a larger share of less productive workers following a shallow recession than following a deep recession. Since more productive workers find new jobs more quickly, this composition effect also explains the relatively rapid recovery of employment following a deep recession (Hall and Kudlyak, 2021). Since it is harder for less productive workers to find good job matches, the impact of unemployment on earnings is more severe for these workers (Gregory et al., 2021). And the negative relationship between the duration of unemployment and the job finding rate of unemployed workers arises because the share of less productive workers in the unemployed increases over time during a recovery (Gregory et al., 2021). Abriti and Consolo (2022) further argue that worker heterogeneity also affects the slope of the Phillips Curve (since the elasticity of marginal cost to the output gap differs between workers of differing productivities) and the equilibrium rate of unemployment, since mismatch between workers of differing productivities affects the Beveridge Curve.

[2]There is also a large literature that distinguishes between formal and informal sectors (Zenou, 2008; Ulyssea, 2010), mainly in the context of developing economies.

[3]Wages are also a possible proxy for productivity; we do not use wages in our analysis because they are endogenous in theoretical models we are evaluating, and also because there are well-documented issues with the quality of wage data derived from individual-level surveys such as the LFS.

in higher skill occupations are more productive than workers in lower skill occupations[4]. Our second proxy is education. Most of the literature assumes that skills are related to education, so we use measures of educational attainment as proxies for productivity. Some models in the theoretical literature relate productivity to job tenure, because workers gain productivity on the job (eg Baley et al., 2022). In this paper, we do not use tenure as a proxy for productivity, but we do include measures of job tenure in our data set.

We use these data to investigate whether productivity, proxied by occupation and education, is the most important type of heterogeneity between workers. A regression-based approach is not well suited to doing this, since there is no obvious measure of heterogeneity to explain. Clustering offers an alternative way to address this question. As discussed by eg Hastie et al. (2009), Kaufman and Rousseeuw (1990) and Malik and Tuckfield (2019), clustering is a form of Unsupervised Machine Learning, used to partition data into a number of sub-samples or clusters. Clustering algorithms seek to form clusters so that data points in the same cluster are as similar as possible and as distinct as possible from data points in other clusters. As a result, cluster membership will reflect the most important sources of difference in a dataset.

So, to test the assumption that heterogeneity between workers only reflects productivity, we can simply examine whether cluster membership is driven by our proxies for productivity, rather than sex, age or some other characteristic. Finding that it is would give empirical support to the recent theoretical literature, whereas the basis of the literature may need to be rethought if it is not. We can address the model design issues in the literature by examining the roles of occupation and education in driving cluster membership. Finding that there are no systematic differences between individuals in our clusters would suggest there are unexplained differences in productivity between individuals, supporting the use of models in which productivity is entirely random. Models in which productivity is related to education imply that cluster membership is driven by occupation and education, whereas models in which productivity is independent of education imply that cluster membership is related to occupation but not education. And segmentation of the labour market implies that highly educated and less educated workers should be in distinct clusters: evidence that this is not the case would support models in which labour markets are not segmented.

There is increasing interest in using machine learning techniques in Economics, as discussed in Athey (2019) and Athey and Imbens (2019). Gregory et al. (2021) use a clustering algorithm to allocate individuals into one of three clusters, using US data on the proportion of time spent as unemployed, the duration of spells of employment and unemployment and the number of job-to-job transitions. They find three clusters; most workers are "α" types, with short and infrequent spells of unemployment. Around 15% of workers are "γ" types, with long and frequent unemployment spells. And around 25% are intermediate "β" types. They find that cluster membership is not related to characteristics such as age, gender or education. These findings are not directly comparable with our results, since workers are not clustered on the basis of occupation, education, gender or other variables and so the question of what is the most important source of heterogeneity is not addressed. Faia et al. (2021) and Grigsby (2021) use a clustering approach to investigate the relationship between occupation, education and productivity; this is not addressed in this paper since we keep occupation and education distinct in order to address issues of model design[5].

This paper is structured as follows. In section 2), we outline our data and the variables we use, discussing our proxies for productivity. In section 3), we discuss clustering and explain our choice of a k-Medoids algorithm; we then outline how this algorithm works and explain how we determine the number of clusters

[4]This assumption is supported by UK evidence in Turrell et al. (2021).

[5]Faia et al. (2021) and Grigsby (2021), use a clustering algorithm to classify jobs into occupational groups and through this to construct measures of the productivity of households in these different groups, using O-Net data on the skill requirements of nearly 1000 occupations.

using the silhouette value (Kaufman and Rousseeuw (1990)), a measure of how well a data point fits within its assigned cluster. Section 4) contains our main results. Silhouette values indicate using two clusters. The algorithm allocates each data point into one of two very different clusters. One cluster contains a large share (83%) of workers in high skill occupations and is dominated by graduates (74%) and workers with A-levels or higher qualifications (87%). The other cluster contains a large share (79%) of workers in medium- or low skill occupations and contains very few workers who are graduates (2%) or have A-levels (7%). Other variables are not associated with differences in cluster membership. These results show that occupation and education have a dominant role in clustering workers into different groups, supporting the assumption that differences in productivity are the most important source of heterogeneity between workers. Considering the model design issues within the literature, the clear differences in worker characteristics between clusters suggests that differences in productivity are systematic rather than purely random; the strong relationship between education and occupation suggests that education is associated with productivity; and the findings that not all members of the high productivity cluster are highly educated and that some members of the low productivity cluster are highly educated suggests that labour markets are not fully segmented.

We validate and establish the robustness of our main results in two ways. First we randomly divide our sample into two sub-samples and cluster each sub-sample separately. We find that in 96% of cases, individuals are allocated to the same cluster with sub-sample clustering as they are with the full sample. Second, we use alternative clustering algorithms. The first is the well-known k-Means algorithm. Although this is better suited to clustering with continuous variables, using this algorithm provides useful insights. In 90% of cases, the k-Means algorithm allocates individuals to the same cluster as with the k-Medians algorithm. The drivers of cluster membership are broadly similar, but with some differences: for example the role of educational attainment is more sharply defined with k-Means. We also explore "soft clustering"; in contrast to "hard clustering", this allows an individual to belong, in a probabilistic sense, to multiple clusters. We use the 'Fuzzy c-Means" algorithm, which is a generalisation of k-Means. The results are similar to those obtained for k-Means clustering but, as we would expect, the differences between clusters are not as clearly defined.

We also present results for three and four clusters. In the case of three clusters, we find a cluster that is dominated by highly-educated workers in high skill occupations, a cluster that is dominated by less well-educated workers in medium skill occupations and a cluster containing mainly workers with low educational attainment in low skill occupations. With four clusters, we find two clusters containing highly educated workers in high skill occupations as well as clusters dominated by workers in medium and low skill occupations. With a larger number of clusters, occupation and education retain their dominant role in determining cluster membership, but the increased granularity reveals interesting additional sources of worker heterogeneity in terms of sex, tenure and sector. Finally, in section 5), we summarise and conclude.

2 Data

We use data from the 2019Q4 UK Labour Force Survey, the last survey before the onset of the Covid-19 pandemic. Our sample comprises 25,000 observations. We restrict our analysis to the study of heterogeneity between employed workers, leaving analysis of differences between a wider sample that includes the employed, the self-employed, the unemployed and the inactive for subsequent work. We use the binary variables[6]

[6]We use only binary variables as clustering algorithms are more effective when applied to binary variables rather than a mixture of binary and more continuous variables. This explained in more detail below, when we outline the steps used in a clustering algorithm.

outlined in Table 1) in our analysis. To measure occupation, we use three binary variables which indicate whether an individual is employed in a "high skill", a "medium skill' or a "low skill' occupation, as defined by the UK Office for National Statistics[7]

For education, we use indicators of whether the individual has at least a degree, has A-levels or higher qualifications and has GCSE or higher qualifications[8]. We also use binary variables that indicate whether the individual is a graduate, female, non-white or young (aged less than 30). We include indicators of whether the respondent is employed in London or in the South East. And we use indicators of whether the individual has a temporary contract, has a zero-hour contract, has been been in their current job for less than a year, has been in their current job for more than five years and whether they are searching for a different job. We also include a binary variable that indicates whether the respondent is employed in the public sector.

Table 1: **Characteristics of Jobs and Workers Used in Clustering Analysis**

Characteristic	**Definition**	Whole-Sample Average
High skill	1-digit SOC code between 1 and 3	0.54
Medium skill	1-digit SOC code between 4 and 6	0.27
Low skill	1-digit SOC code between 7 and 9	0.19
Graduate	Respondent is a graduate	0.40
A-level or higher	Respondent has A-Levels or higher qualifications	0.50
GCSE or higher	Respondent has GCSEs or higher qualifications	0.73
Female	Respondent is female	0.41
Non-White	Respondent is non-white	0.12
Young	Respondent is aged ≤ 30	0.28
London	Respondent is employed in London	0.17
South East	Respondent is employed in the South East of England	0.13
Temp	Employed on a temporary contract	0.03
Zero Hour Contract	Employed on a zero hour contract	0.01
Short Tenure	Respondent has been in current job for $\leq$ one year	0.16
Long Tenure	Respondent has been in current job for $\geq$ five years	0.49
Searching for New Job	Respondent is currently searching for a different job	0.06
Public Sector	Employed in the Public Sector	0.25

Source: UK Labour Force Survey, 2019Q4
25,000 observations

3 Methodology

3.1 Clustering

This subsection draws on Hastie et al. (2009). Related texts include Kaufman and Rousseeuw (1990), Rogers and Girolani (2012) and Malik and Tuckfield (2019). Suppose we have data on P characteristics. Then data point i can be represented as the $(P \times 1)$ array of x_i, where x_i^p denotes the pth value of x_i for $p = 1, .., P$. We can represent the dissimilarity or distance between data points i and j as $D(x_i, x_j)$. Then if there are N data points, the total dissimilarity in our data is the constant $T = \sum_{i=1}^{N} \sum_{j=1}^{N} D(x_i, x_j)$. Suppose

[7]The ONS classifies occupations with 1-digit Standard Occupational Classification codes between 1-3 as high skill, occupations with SOC codes between 4-6 as medium skill; and occupations with SOC codes between 7-9 as low skill. High skill occupations comprise managerial, professional and technical roles, medium skill occupations include administrative and secretarial roles, skilled trade occupations and roles in caring, leisure and other services; low skill occupations comprise sales and customer service workers, process and machine operatives and elementary occupations.

[8]GCSEs are roughly equivalent to the US High School Diploma; A-Levels are broadly similar to the US SATs or APs.

there are K clusters. Clustering gives a mapping $k_i = C(i)$ that assigns each data point i, $i = 1, .., N$ into a unique cluster[9] k, $k = 1, .., K$. Then, as shown in, eg Hastie et al. (2009), $T = W + B$, where $W = \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(j)=k} D(x_i, x_j)$ is the *within cluster distance* and $B = \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(j)\neq k} D(x_i, x_j)$ is the *between cluster distance*. A clustering algorithm aims to maximise the similarity of data points within a cluster and thus seeks to minimise W. In doing so, it maximises B. In other words, choosing cluster membership in order to minimise the difference between data points in the same cluster leads to the maximisation of the total distance between clusters. So, in essence, clustering aims to allocate each data point into one of a pre-determined number of clusters in order that data points within a cluster are more similar to points in that cluster than to data points in other clusters.

Different approaches to clustering are distinguished by the measure of distance and the algorithm that is used. Since we are using binary data, we use the Hamming Distance, so $D(x_i, x_j)$ is the Hamming Distance between data points i and j. This is the proportion of cases in which $x_i^p \neq x_j^p$ for $p = 1, ..P$[10]. Cluster membership is determined through an algorithm that is based around the central point or centroid of a cluster. Since our data is binary, we use the median point or medoid of a cluster[11]. Since there are $\frac{1}{K!} \sum_{k=1}^{K} \binom{K}{k} k^N$ ways to assign N data points into K clusters (eg Hastie et al. (2009)), it is not feasible for a clustering algorithm to search over all possible assignments when using large datasets, such as ours. Therefore, clustering algorithms use a sampling approach to large datasets. We use the "Clustering Large Applications" (CLARA) variant of the Partitioning Around Medoids (PAM) algorithm, proposed by Kaufman and Rousseeuw (1990) to address the challenges posed by large datasets through sampling.

The CLARA algorithm for the case where there are K medoids proceeds as follows: (i) an initial random sample is taken from the full dataset; (ii) K initial candidate medoids are randomly selected from within the subsample; (iii) each data point in the subsample is assigned to the closest cluster, as measured by the Hamming Distance between that data point and the candidate medoid; (iv) the distance between each data point and every other data point in the cluster is calculated; the data point with the smallest distance within a cluster becomes the new candidate medoid for that cluster[12]; (v) steps (iii)-(iv) are repeated until there is no change in cluster membership within the subsample; (vi) cluster membership for the full sample is then determined by allocating each data point to the cluster containing the closest candidate medoid; (vii) a new subsample is taken from the full sample and steps (iii)-(vi) are repeated; (viii) this is repeated until cluster membership of the full sample does not change. The initial subset of data in step (i) is selected using the "k-means++" sampling approach of Arthur and Vassilvitskii (2007). The CLARA algorithm converges to a local minimum of W; the algorithm is run 90 times with different starting values, with the best result being selected, in order to give assurance that a global solution has been found.

3.2 Evaluating the Fit of Datapoints Within Clusters

We evaluate the fit of data points within their assigned clusters using the Silhouette value; this is a measure of how similar a data point is to other points in its cluster, compared to data points in other clusters.

[9] This describes hard clustering; we will also explore soft clustering, which allows a data point to belong to multiple clusters.

[10] For example, the Hamming Distance between (1,0,0,0) and (0,1,0,0) is 0.5 and Hamming distance between (1,0,0,0) and (1,1,0,0) is 0.25. The Hamming Distance is widely used in Information Theory and Cryptography.

[11] The most widely used approach is k-Means clustering: this uses a Euclidian measure of distance and defines the centre of a cluster as the mean value of cluster members. This is useful for analysing continuous variables, but is less well suited than k-Medians clustering for the analysis of binary data (eg Hastie et al. (2009)).

[12] This feature influenced our choice to include only binary variables in our data set. If the data set contained a mixture of binary and more continuous variables, the clustering algorithm would focus on the binary variables in this step, since the change in W would be larger if the change in the medoid led to changes in a binary variable compared to a change in a more continuous variable.

Suppose there are N_k data points in cluster k. Then we define $\zeta_i^k = \frac{1}{N_k} \sum_{C(j)=k} D(x_i, x_j)$ to be the average of the Hamming distances between data point i in cluster k and other data points in that cluster. And we define $\zeta_i^{k'} = min_{\{C(j)=k', k' \neq k\}} \frac{1}{N_{k'}} \sum_{C(j)=k'} D(x_i, x_j)$ to be the average of the Hamming distances between data point i in cluster k and data points in the cluster that is closest to k. Then we define the Silhouette measure for data point i in cluster k when there are K clusters as (Kaufman and Rousseeuw (1990)).

$$S_i^k(K) = \frac{(\zeta_i^{k'} - \zeta_i^k)}{\max(\zeta_i^{k'}, \zeta_i^k)} \tag{1}$$

This measure lies between -1 and 1, where a high value indicates a good fit of data point i within cluster k and a negative value indicates that the data point is closer to points in another cluster. The Silhouette measure for each data point can be conveniently illustrated using a Silhouette plot.

3.3 The Number of Clusters

As discussed in, eg Hastie et al. (2009) and Rogers and Girolani (2012), there is no natural criterion that can be used to determine the number of clusters to be used; a pragmatic approach is therefore advised. We assess the number of clusters in our data using two statistics: the average value of Silhouette measures across all data points; and the percentage of cases for which the silhouette statistic is negative. We define the average Silhouette value for cluster k as

$$S^k(K) = \frac{1}{N_k} \sum_{i=1}^{N_k} S_i^k(K) \tag{2}$$

where N_k is the number of data points in cluster k. The overall average Silhouette statistic is then

$$S(K) = \frac{1}{K} \sum_{i=1}^{K} S^k(K) \tag{3}$$

The proportion of cases for which the silhouette statistic is negative is given by

$$\sigma(K) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{N_k} I\{S_i^k(K) < 0\} \tag{4}$$

where $I\{S_i^k(K) < 0\} = 1$ if $S_i^k(K) < 0$ and $I\{S_i^k(K) < 0\} = 0$ if $S_i^k(K) \geq 0$.

4 Results

4.1 The Number of Clusters

Figure 1) shows values of $S(K)$ and $\sigma(K)$ for different values of K. Both measures favour two clusters. In what follows, we therefore concentrate on the case where $K = 2$, but we also present results for $K = 3$ and $K = 4$. Figure 2) plots silhouette values for each data point for $K = 2$, $K = 3$ and $K = 4$.

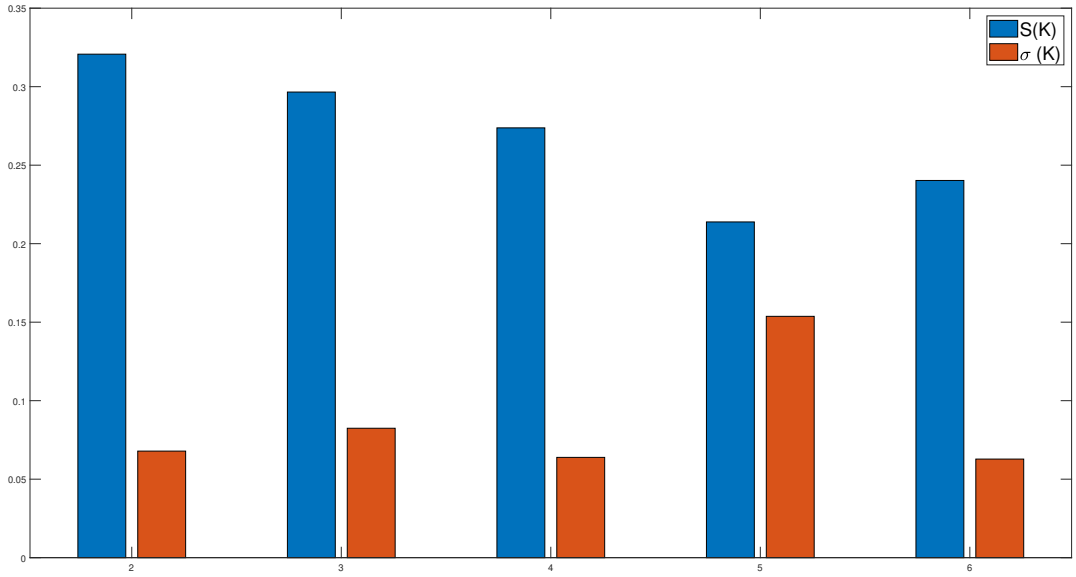

Figure 1: **Average and Negative Silhouette Statistics**
This figure plots S(K) and $\sigma(K)$ for $K = 2, ...6$.

4.2 Two Clusters

For $K = 2$, we label the clusters as clusters 2A and 2B. Table 2) contains our main results, documenting the average values of the variables in Table 1) for each cluster. There are marked differences between clusters 2A and 2B. For all variables, the difference between the mean value reported for cluster 2A and the mean value reported for cluster 2B is statistically significant at the 0.005% level[13]. To focus on the main drivers of differences between clusters, we highlight cases where the average value of a variable in the cluster differs from the average value for the whole sample by more than 40% in red.

The algorithm allocates data points into one of two very different clusters. 83% of members of cluster 2A work in high skill occupations, compared to only 9% and 8% respectively from medium and low skill occupations. By contrast, 79% of members of cluster 2B work in medium or low skill occupations, compared to only 21% in high skill occupations. 74% of workers in cluster 2A are graduates, 87% have A-levels and 98% have GCSEs. By contrast, only 2% of workers in cluster 2B are graduates, only 7% have A-levels and only 42% have GCSEs. There are no other major differences in cluster membership[14].

These results show that occupation and education have a dominant role in clustering workers into different groups[15]. These are the characteristics that most strongly make data points in the same cluster similar, reflected in a smaller Hamming Distance, and make data points in one cluster more dissimilar from data points in the other cluster. We can therefore conclude that occupation and education are the most important type of heterogeneity between workers. These findings therefore support the assumption made in the literature

[13]The exception is Searching for New Job, where the difference is significant at 0.2%.
[14]With the exception of working on a fixed-term contract; the number of workers on these contracts is however low.
[15]Our choice to highlight variables where the difference between the cluster mean and the whole sample mean exceeds 40% is arbitrary. If we use a 50% criterion, the only change in Table 2) would be a loss of highlight for the GCSE measure. With a 30% criterion, differences in zero-hours contracts and working in London or in the public sector would be highlighted. The impact of these variables will become more apparent as we increase the number of clusters.

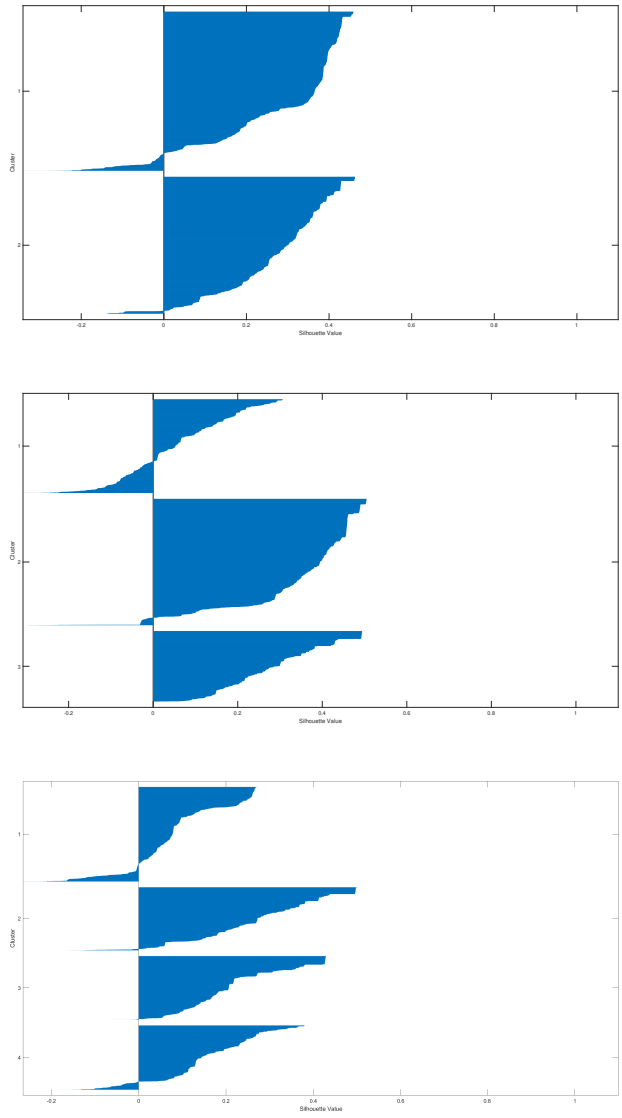

Figure 2: **Silhouette Diagrams for $K = 2$, $K = 3$ and $K = 4$**

on macroeconomic models with heterogeneity, that differences in productivity are the most important source of heterogeneity between workers. Considering the first model design issue within the literature that we are addressing, the clear differences in the characteristics of workers in different clusters implies that differences in productivity are systematic rather than purely random. For the second issue, it is clear that there is a strong relationship between education and cluster membership, implying that higher educational attainment is associated with higher productivity. For the third issue, the fact that not all workers in the high productivity cluster have higher educational qualifications suggests that UK labour markets are not fully segmented. This is a feature of the data, not an artifact of clustering: 19% of workers with a degree are not employed in a high skill occupation, as are 23% of those with A-levels or higher qualifications. Also, 27% of workers with qualifications below GCSEs are employed in high skill occupations. We also note that these results do not

support the approach of distinguishing between workers on the basis of differences in job tenure[16].

Table 2: **Summary Statistics:** $K = 2$**.**

Characteristic	All	2A	2B
Sample Share	1.00	0.47	0.53
High skill	0.54	0.83	0.21
Medium skill	0.27	0.09	0.46
Low skill	0.19	0.08	0.33
Graduate	0.40	0.74	0.02
A-Level or higher	0.50	0.87	0.07
GCSE or higher	0.73	0.98	0.42
Female	0.41	0.50	0.30
Non-White	0.12	0.14	0.09
Young	0.28	0.29	0.26
London	0.17	0.23	0.11
South East	0.13	0.13	0.12
Temp	0.03	0.04	0.03
Zero Hour Contract	0.01	0.01	0.01
Short Tenure	0.16	0.17	0.13
Long Tenure	0.49	0.42	0.57
Searching for New Job	0.06	0.06	0.05
Public Sector	0.25	0.32	0.16

Source: UK Labour Force Survey, 2019 Q4
25,000 observations
red indicates $\geq 40\%$ difference from the sample average.

4.3 Model Validation and Alternative Approaches to Clustering

To demonstrate the credibility of these results, we next perform a validation exercise. The clustering algorithm gives a mapping $k_i = C(i)$, so $k_i = 1$ if data point i is assigned to cluster $2A$ and $k_i = 2$ if data point i is assigned to cluster $2B$. We randomly divide our sample into two equally sized sub-samples, so data point i is randomly allocated to sub-sample s, where $s \in \{1, 2\}$. We repeat our clustering exercise on these sub-samples, giving the mappings $k_i^s = C^s(i)$, for $s \in \{1, 2\}$; $k_i^s = 1$ if data point i is assigned to cluster $2A$ and $k_i^s = 2$ if data point i is assigned to cluster $2B$. We compute the proportion of cases in which a data point is allocated to the same cluster in the cases where the full sample is used and where sub-samples are used, using the statistic $\omega = \frac{\sum_{i=1}^{N} I\{k_i = k_i^s\}}{N}$, where $I\{k_i = k_i^s\} = 1$ if $k_i = k_i^s$ and $k_i = 0$ otherwise. We find $\omega = 0.963$, showing that almost all data points are allocated to the same cluster whether the sub-sample or full sample data is used[17].

4.3.1 A k-Means Algorithm

If we use the Euclidian Measure of distance, so $D(x_i, x_j) = (x_i - x_j)^2$, then the within cluster distance becomes (eg Hastie et al. (2009)) $W = \sum_{k=1}^{K} N_k \sum_{C(i)=k} (x_i - \bar{x}^k)^2$, where $\bar{x}^k$ is the mean value of x_i for all data points in cluster k. This criteria can be minimised using the k-Means algorithm. The k-Means

[16] Gregory et al. (2021) distingush between workers on the basis of job tenure and unemployment history; with our data, we cannot assess the impact of unemployment histories.

[17] If we split the data so that the first sub-sample contains, respectively, 90%, 75%, 25% and 10% of the full dataset, we find $\omega = 0.985$, $\omega = 0.981$, $\omega = 1.000$ and $\omega = 0.987$, respectively.

algorithm proceeds as follows: (i) K initial candidate centroids are randomly selected; (ii) Euclidian Distances are calculated between each data point and these candidate centroids; (iii) each data point is assigned to the cluster with the closest centroid; (iv) new centroids are calculated as the average of the observations in each cluster; (v) steps (ii)-(iv) are repeated until cluster membership does not change. The algorithm converges to a local minimum of W; the algorithm is run 90 times with different starting values, with the best result being selected, in order to give assurance that a global solution has been found.

Although k-Means is more suited to continuous variables and is more sensitive to data outliers (Malik and Tuckfield (2019)) than k-Medians, application of this algorithm to our data provides useful insights. Our results are shown in Table 3). In 90% of cases, the k-Means algorithm allocates data points to the same cluster as with the k-Medians algorithm. Reflecting this, the drivers of cluster membership are broadly similar to those in Table 4), but with some differences: the role of educational attainment is more sharply defined, as are the differences between public and private sectors and the effect of living in London.

4.3.2 Soft Clustering

Soft clustering is a generalisation of k-Means clustering where each data point can belong to multiple clusters (eg Hastie et al. (2009)). We use the Fuzzy c-means (FCM) algorithm (Bezdec (1981)), in the case where $K = 2$. The weight of data point i in cluster k is $F(i,k)$, where $\sum_{k=1}^{2} F(i,k) = 1$. The centroid of cluster k is the weighted mean of data points within the cluster, given by $\tilde{x}^k = \frac{\sum_{i=1}^{N} F(i,k)x_i}{\sum_{i=1}^{N} F(i,k)}$. With a hard clustering algorithm such as k-Medians or k-Means, each data point belongs to a unique cluster and influences only the centroid of that cluster. As a result, the boundaries between clusters are sharply defined. With soft-clustering, data points can belong to both clusters and so can influence both centroids; this makes the boundaries between clusters "fuzzy" and less distinct. The parameter m, where $m > 1$, determines the extent of overlap between clusters. If $m \to 1$, then the algorithm gives the same results as k-Means clustering. As m increases, the degree of overlap between clusters increases. As $m \to \infty$, there is complete fuzziness, and all clusters are identical. We use the conventional value of $m = 2$. In this case, the FCM algorithm minimises the criterion $W' = \sum_{i=1}^{N} \sum_{k=1}^{K} F(i,k)^2 |x_i - \tilde{x}^k|^2$.

The algorithm proceeds as follows: (i) initial values for $F(i,k)$ are randomly selected; (ii) the centroids of each cluster are calculated; (iii) values for $F(i,k)$ are updated using the rule $F(i,k) = \frac{1}{\sum_{k'=1}^{2} \left\{ \frac{|x_i - \tilde{x}^k|}{|x_i - \tilde{x}^{k'}|} \right\}^2}$ (iv) steps (ii)-(iii) are repeated until convergence. For each data point, the updating rule allocates greater weight to a cluster that gives a smaller distance from the data point to the centroid of that cluster. In 96% of cases, the k-Means and FCM algorithms allocate a data point to the same cluster; 91% of observations are assigned by FCM to the same cluster as with k-Medians. Table 2) contains our results on cluster membership for Fuzzy clustering. The results are similar to those obtained for k-Means clustering but, as we would expect, the differences between clusters are not as clearly defined as with k-Means.

4.4 More Clusters

Tables 4) and 5) document our results for $K = 3$ and $K = 4$. With $K = 3$, cluster 3A is dominated by workers employed in high skill occupations, cluster 3B has a large share of workers in medium skill occupations and cluster 3C is dominated by workers in low skill occupations. With $K = 4$, there are two high skill clusters, 4A and 4B. Clusters 4C and 4D have large shares of workers in medium and low skill occupations respectively. This is summarised in Fig 3).

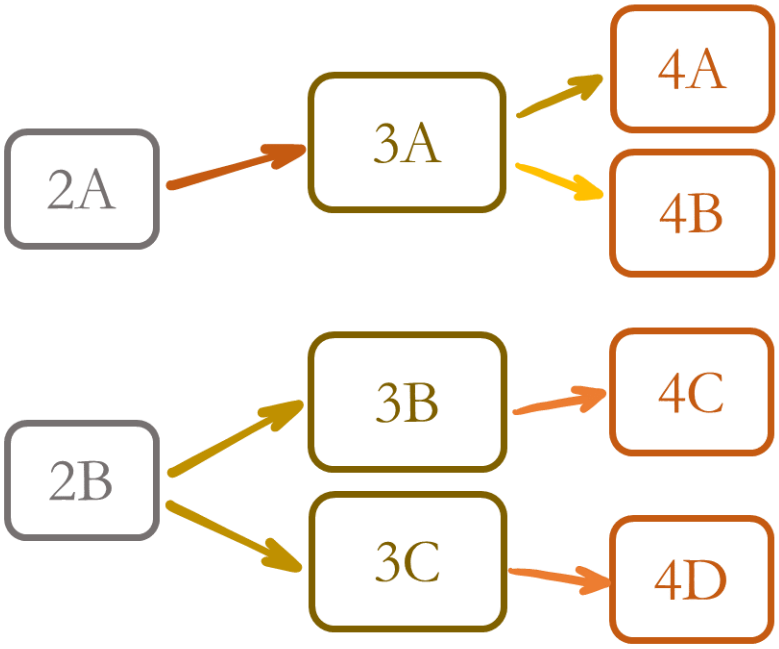

Figure 3: **Similarities between Clusters across $K = 2$, $K = 3$ and $K = 4$**

Occupation and education continue to have a central role in determining cluster membership. We find that 94% of members of cluster 3A work in high skill occupations, as do 86% and 88% of members of clusters 4A and 4B. Clusters 3B and 4C contain 72% and 87% respectively of workers employed in medium skill occupations. And clusters 3C and 4D contain 64% and 57% workers employed in low skill occupations. In terms of education, 82% of cluster 3A have degrees, 93% have A-levels or higher and all have at least GCSEs. The same is true of clusters 4A and 4B with 74% and 80% being graduates, 86% and 91% having at least A-levels and 97% and 100% having at least GCSEs. There is a sharp contrast with low skill clusters, since no member of cluster 3C and only 2% of cluster 4D have a degree. Only 4% of cluster 3C and 5% of cluster 4D have at least A-levels, while only 20% and 29% respectively of these clusters have at least GCSE qualifications. The medium skill clusters are slightly more qualified, but only 16% of members of cluster 3B are graduates, as are only 6% of cluster 4C, while only 27% and 17% respectively of these clusters have at least A-levels.

With three clusters, differences in gender between clusters emerge: cluster 3B contains 60% of female workers, while cluster 3C only contains 20%. This difference is also apparent for $K = 4$. With the increased granularity of $K = 4$, there are differences in tenure, gender and sector between the two high skill clusters. Only 6% of members of cluster 4A have been employed in their current job for less than one year, and 77% have been employed for more than five years, compared to 21% and 28% respectively for cluster 4B. 78% of members of cluster 4A are female and 75% are employed in the public sector, compared to 17% and 11% respectively for cluster 4B.

Table 3: **Summary Statistics: k-Means and Fuzzy k-Means**

Characteristic	All	K-Means:2A	K-Means:2B	Fuzzy:2A	Fuzzy:2B
Sample Share	1.00	0.46	0.54	0.47	0.53
High skill	0.54	0.83	0.30	0.84	0.28
Medium skill	0.26	0.11	0.39	0.11	0.40
Low skill	0.19	0.06	0.31	0.05	0.32
Graduate	0.40	0.89	0.00	0.85	0.00
A-Level or higher	0.50	1.00	0.07	0.97	0.07
GCSE or higher	0.73	1.00	0.49	1.00	0.47
Female	0.41	0.46	0.36	0.46	0.36
Non-White	0.12	0.16	0.09	0.16	0.09
Young	0.28	0.26	0.29	0.26	0.29
London	0.17	0.25	0.11	0.25	0.11
South East	0.13	0.13	0.12	0.13	0.12
Temp	0.03	0.04	0.03	0.04	0.03
Zero Hour Contract	0.01	0.01	0.01	0.01	0.01
Short Tenure	0.16	0.15	0.15	0.15	0.16
Long Tenure	0.49	0.47	0.50	0.48	0.49
Searching for New Job	0.06	0.06	0.06	0.06	0.06
Public Sector	0.25	0.35	0.16	0.37	0.14

Source: UK Labour Force Survey, 2019 Q4
25,000 observations
red indicates $\geq 40\%$ difference from the sample average.

Table 4: **Summary Statistics:** $K = 3$**.**

Characteristic	All	3A	3B	3C
Sample Share	1.00	0.32	0.43	0.25
High skill	0.54	0.94	0.22	0.26
Medium skill	0.27	0.02	0.72	0.10
Low skill	0.19	0.04	0.06	0.64
Graduate	0.40	0.82	0.16	0.00
A-Level or higher	0.50	0.93	0.27	0.04
GCSE or higher	0.73	1.00	0.75	0.20
Female	0.41	0.38	0.60	0.20
Non-White	0.12	0.15	0.10	0.10
Young	0.28	0.23	0.38	0.22
London	0.17	0.24	0.13	0.10
South East	0.13	0.13	0.13	0.11
Temp	0.03	0.03	0.04	0.03
Zero Hour Contract	0.01	0.00	0.02	0.01
Short Tenure	0.16	0.13	0.21	0.13
Long Tenure	0.49	0.55	0.32	0.60
Searching for New Job	0.06	0.05	0.07	0.05
Public Sector	0.25	0.35	0.22	0.11

Source: UK Labour Force Survey, 2019Q4
25,000 observations.
red indicates $\geq 40\%$ difference from the sample average

5 Conclusions

This paper has used machine learning techniques to investigate the main sources of heterogeneity in the UK labour market. Applying a clustering algorithm to individual-level data, we find that cluster membership is mainly driven by occupation. If occupation is a good proxy for productivity, this finding provides support for a recent theoretical literature which argue that accounting for differences in productivity can help resolve long-standing issues in the analysis of labour markets. Different models within this literature are characterised by different assumptions about the relationship between productivity and education. Our results support models in which any worker can do any job, but where more highly educated workers are more productive than the less-well educated.

This work can be extended in several directions. For example, our results suggest that the distinction between working in the public and private sectors is important for the UK labour market. Working in the public sector is associated with longer job tenures, a larger share of female employment and a larger share of graduates. Further investigation of the reasons for this, and in how this affects and reflects wage determination and recruitment practices would be interesting. Also, we have only considered employees in this paper. Widening the scope to include the self-employed, the unemployed and the inactive might well uncover other interesting and important structural differences across the labour market. More widely, this paper illustrates the potential of using machine learning techniques in Economics, especially as a wider range of data sets and types of "big data" become available. Machine learning opens the prospect of addressing a wider range of research questions with a richer set of analytical tools.

References

Abriti, M. and Consolo, A. (2022), 'Labour market skills, endogenous productivity and business cycles', ECB Working Paper Series .

Adjemian, S., Karamé, F. and Langot, F. (2021), 'Nonlinearities and Workers' Heterogeneity in Unemployment Dynamics', IZA DP No. 14822 .

Arthur, D. and Vassilvitskii, S. (2007), "k-means++: The advantages of careful seeding", in 'Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms'.

Athey, S. (2019), The impact of machine learning on economics, in A. G. AK Agrawal, J Gans, ed., 'The Economics of Artificial Intelligence: An Agenda', Univ. Chicago Press, Chicago, pp. 685–725.

Athey, S. and Imbens, G. (2019), 'Machine Learning Methods That Economists Should Know About ', Annual Review of Economics **11**, 685–725.

Baley, I., Ljungqvist, L. and Sargent, T. (2022), 'Returns to labor mobility', Unpublished Paper .

Bezdec, J. (1981), Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

Booth, A., Francesconi, M. and Frank, J. (2002), 'Temporary Jobs: Stepping Stones or Dead Ends?', The Economic Journal **112**, 189–213.

Brotherhood, L., Kircher, P., Santos, C. and Tertilt, M. (2021), 'An Economic Model of the Covid-19 Pandemic with Young and Old Agents: Testing and Policies ', LIDAM Discussion Paper CORE 2021 34 .

Dolado, J, ., Lalé, E. and Turon, H. (2021), ' Zero-hours Contracts in a Frictional Labor Market ', CEPR Discussion Paper No. DP16843 .

Dolado, J., Motyovski, G. and Pappa, E. (2021), 'Monetary policy and inequality under labor market frictions and capital-skill complementarity', American Economic Journal: Macroeconomics .

Duzhak, E. (2021), 'How Do Business Cycles Affect Worker Groups Differently? ', FRBSF Economic Letter 2021-25 .

Faccini, R. and Melosi, L. (2021), 'Bad Jobs and Low Inflation', unpublished paper .

Faia, E., Kudlyak, E. and Shabalina, E. (2021), 'Dynamic Labor Reallocation with Heterogeneous Skills and Uninsured Idiosyncratic Risk', IZA DP No. 14794 .

Gertler, M., Huckfeldt, C. and Trigari, A. (2020), 'Unemployment Fluctuations, Match Quality and the Wage Cyclicality of New Hires', The Review of Economic Studies .

Gregory, V., Menzio, G. and Wiczer, D. G. (2020), 'Pandemic Recession: L or V-Shaped?', NBER Working Papers .

Gregory, V., Menzio, G. and Wiczer, D. G. (2021), 'The Alpha Beta Gamma of the Labor Market', NBER Working Papers (28663).

Grigsby, J. (2021), 'Skill heterogeneity and aggregate labor market dynamics', unpublished paper .

Hall, R. and Kudlyak, M. (2020), 'Job-Finding and Job-Losing: A Comprehensive Model of Heterogeneous Individual Labor-Market Dynamics ', Federal reserve Bank of San Francisco Working Paper 2019-05 .

Hall, R. and Kudlyak, M. (2021), 'The Inexorable Recoveries of U.S. Unemployment ', Board of Governors of the Federal Reserve System 2021-20 .

Hastie, T., Tibshirani, R. and Friedman, J. (2009), The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd edition edn, Springer.

Kaufman, L. and Rousseeuw, P. (1990), Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, Inc.

Ljungqvist, L. and Sargent, T. J. (2017), 'The Fundamental Surplus', American Economic Review **107**(9), 2630–65.

Malik, J. and Tuckfield, B. (2019), Applied Unsupervised Learning with R, Packt.

Martin, C. and Okolo, M. (2022), 'Modelling the Differing Impacts of Covid-19 in the UK Labour Market ', Oxford Bulletin of Economics and Statistics .

Pries, M. (2004), 'Persistence of Employment Fluctuations: A Model of Recurring Job Loss', The Review of Economic Studies .

Rogers, O. and Girolani, O. (2012), A FIrst Course in Machine Learning, Taylor and Francis.

Turrell, A., Speigner, B., Copple, D., Djumalieva, J. and Thurgood, J. (2021), 'Is the uk's productivity puzzle mostly driven by occupational mismatch? an analysis using big data on job vacancies', Labour Economics .

Ulyssea, G. (2010), 'Regulation of Entry, Labor Market Institutions and the Informal Sector', Journal of Development Economics **91**, 87–99.

Zenou, Y. (2008), 'Job Search and Mobility in Developing Countries: Theory and Policy Implications', Journal of Development Economics **86**, 336–355.

Table 5: **Summary Statistics:** $K = 4$**.**

Characteristic	All	4A	4B	4C	4D
Sample Share	1.00	0.25	0.21	0.23	0.31
High skill	0.54	0.86	0.88	0.09	0.32
Medium skill	0.26	0.11	0.07	0.87	0.09
Low skill	0.19	0.02	0.06	0.05	0.59
Graduate	0.40	0.74	0.80	0.06	0.02
A-Level or higher	0.50	0.86	0.91	0.17	0.05
GCSE or higher	0.73	0.97	1.00	0.65	0.29
Female	0.41	0.78	0.20	0.62	0.17
Non-White	0.12	0.13	0.16	0.09	0.09
Young	0.28	0.17	0.32	0.39	0.21
London	0.17	0.18	0.28	0.12	0.11
South East	0.13	0.13	0.13	0.11	0.12
Temp	0.03	0.03	0.04	0.03	0.03
Zero Hour Contract	0.01	0.00	0.01	0.02	0.01
Short Tenure	0.16	0.06	0.21	0.21	0.12
Long Tenure	0.49	0.77	0.28	0.32	0.63
Searching for New Job	0.06	0.04	0.07	0.07	0.05
Public Sector	0.25	0.75	0.11	0.17	0.08

Source: UK Labour Force Survey, 2019Q4
25,466 observations.
red indicates $\geq 40\%$ difference from the sample average.