Segmented Labour Markets in the UK: A Machine Learning Approach
Labour Force and Annual Population Surveys User Conference 2023

Chris Martin and Magdalyn Okolo

University of Bath

January 24, 2023

Summary

- This paper uses 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 two proxies for productivity: occupation and education.
- This supports a recent theoretical literature which argues that accounting for differences in productivity can help resolve long-standing issues in the analysis of labour markets.
- Our results also imply:
  - differences in productivity are systematic rather than purely random
  - increased education is associated with increased productivity
  - labour markets are not segmented.

The Theoretical Literature (1)

A growing literature argues that accounting for heterogeneity enables us to resolve some long-standing issues

1. the large volatility of unemployment relative to wages
2. the pattern of recovery from deep and shallow recessions
3. the impact of job loss of subsequent earnings
4. the negative relationship between the duration of unemployment and the job finding rate of unemployed workers
5. the slope of the Phillips Curve
6. the equilibrium rate of unemployment

The key assumption in the literature is that that the most important aspect of labour market heterogeneity reflects differences in productivity between workers.

⏩ References ⏩ More Detail

The Theoretical Literature (2)

- There are three main unresolved model design issues in the literature
  1. Is productivity random?
  2. Is the labour market segmented, so only highly skilled workers can do high productivity jobs?
  3. Do differences in ability or skill reflect differences in education?
- In this paper, we present evidence on the relationships between productivity, occupation and education that is relevant to these debates.

⏩ References

Diverse Sources of Heterogeneity

- Differences in productivity across workers are well documented.
- But other dimensions of heterogeneity in the labour market have also been highlighted, including the distinctions between
  - males versus females
  - young versus older workers
  - different ethnicities
  - temporary versus permanent jobs
  - fixed hours versus zero hours contracts
- Current theoretical models assume that differences between job matches reflects differences in productivity, rather than these other types of heterogeneity

Research Questions and Methodology

- In this paper, we address two main research questions
  1. are differences in productivity the main source of heterogeneity in the UK labour market?
  2. can we discriminate between different models in the theoretical literature?
- We address these research questions using a clustering approach.
- Clustering is a form of Unsupervised Machine Learning, used to partition data into one of a pre-determined number of sub-samples or clusters
- Clustering aims to allocate each data point into a cluster in order that data points within a cluster are more similar to points in that cluster than to data points in other clusters
- Clustering is a useful tool in addressing our research questions since it enables us to analyse which of the many dimensions of heterogeneity in our data is the most important.
- A regression-based approach is not well suited to doing this.

Data

- We use data from the 2019Q4 UK Labour Force Survey; the last survey before the onset of the Covid-19 pandemic.
- These data include measures of
  1. sex
  2. age
  3. ethnicity
  4. whether a job is permanent or temporary
  5. whether the worker has a zero-hours contract
  6. tenure in the current job
  7. whether the individual is employed in the public sector
  8. and whether they are searching for an alternative job.

Proxies

- There are no measures of productivity in our data
- So we use proxies for productivity
- The theoretical literature suggests two proxies
    1. Occupation.
    2. Education
- In this paper, we will use both

Occupation and Education

- The UK ONS defines three broad occupational categories

  1. high skill: managerial, technical and scientific occupations
  2. medium skill: administrative and secretarial roles, skilled trade occupations and roles in caring, leisure and other services
  3. low skill: sales and customer service workers, process and machine operatives and elementary occupations

- For education, we use indicators of whether the individual has

  1. at least a degree
  2. has A-levels or higher qualifications: A-levels are broadly similar to US SATs or APs
  3. has GCSE or higher qualifications: GCSEs are GCSEs are roughly equivalent to the US High School Diploma

Data

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

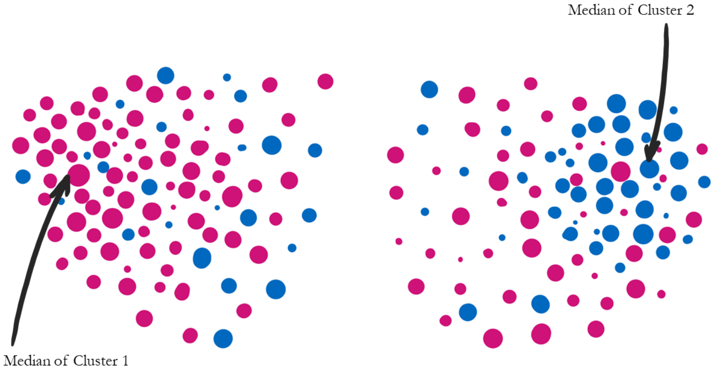
25,000 observations

Clustering (1)

- 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
- Clusters are defined by a central point or centroid
- The clustering algorithm allocates data points to clusters in order to minimise the distance between data points and centroids
- We
  1. use the median value of a cluster as the centroid
  2. use the Hamming measure of distance
  3. use the Clustering Large Applications' (CLARA) variant of the popular Partitioning Around Medoids (PAM) algorithm

⏩ More Details

Clustering (2)



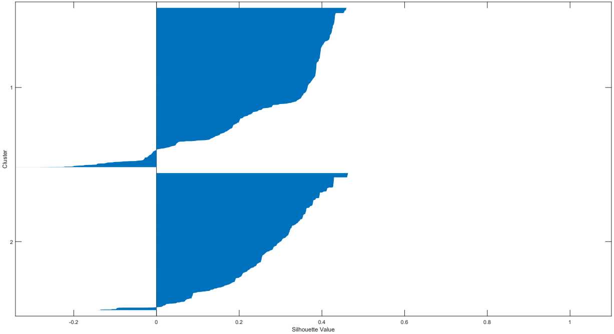
The Silhouette (1)

- We measure how well a data point sits within its cluster using the Silhouette
- This is a measure of how similar a data point is to other points in its cluster, compared to data points in other clusters.
- This measure lies between -1 and 1, where a high value indicates a good fit of the data point within it's allocated cluster and a negative value indicates that the data point is closer to points in another cluster.

⏩ More Details on the Silhouette

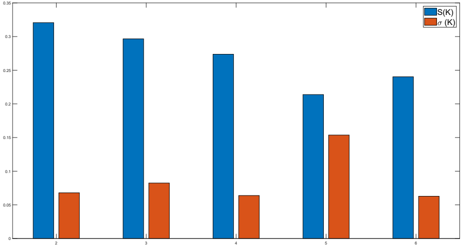
The Silhouette (2)

- This is the plot of Silhouette values, in the case of two clusters



How Many Clusters?

- We assess the number of clusters using two statistics:
  1. The average Silhouette value across all observations: $S(K)$.
  2. The proportion of cases for which the silhouette statistic is negative: $\sigma(K)$.
- Both suggest using two clusters
- We focus on this case, but also have results for three and four clusters



More Details on Silhouette Statistics

Results (1)

- The algorithm allocates data points into one of two very different clusters; we label these clusters 2A and 2B
- 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

Results (2)

- These results show that differences in productivity are the most important source of heterogeneity between workers.
- This supports the recent macroeconomic literature
- Considering our model design issues:
  - the clear differences in the characteristics of workers in different clusters implies that differences in productivity are systematic rather than purely random.
  - the strong relationship between education and cluster membership implies that higher educational attainment is associated with higher productivity.
  - the fact that not all workers in the high productivity cluster have higher educational qualifications suggests that UK labour markets are not fully segmented.

Results (3)

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.

Model Validation

- To demonstrate the credibility of these results, we next perform a validation exercise.
- We randomly divide our sample into two equally sized sub-samples
- We repeat our clustering exercise on these sub-samples
- In 96.3% of cases, data points are allocated to the same cluster whether the sub-sample or full sample data is used
- If we split the data so that the first sub-sample contains, respectively, 90%, 75%, 25% and 10% of the full dataset, we find the same cluster in respectivey 98.5%, 98.1%, 100% and 98.7% of cases.

▸▸ More Details on Validation

Alternative Approaches to Clustering

- To demonstrate the robustess of our results, we also clustered our data using
  - a k-Means algorithm
  - a soft clustering algorithm
- We find
  - 90% of data points are assigned by k-Means to the same cluster as with k-Medians
  - 91% of data points are assigned by soft clustering to the same cluster as with k-Medians.

⏩ More Details on k-Means ⏩ More Details on Soft Clustering

What Have We Learnt?

- The primary driver of cluster membership is occupation and education
- If these are good proxies for productivity, then differences in productivity are the main source of heterogeneity in the UK labour market
- This supports the recent theoretical literature which argues that accounting for differences in productivity can help resolve long-standing issues in the analysis of labour markets.
- Our results do not support models in which
    1. productivity is random
    2. productivity is unrelated to education
    3. labour markets are segmented
- Our results tend to support models in which
    1. productivity is systematic and increased by education
    2. labour markets are not fully segmented
    3. so any worker can do any job; but more highly educated workers are more productive

Going Forward

- This work can be extended in several directions.
    - 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.
    - the public/private sector distinction seems to be important; 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 this would be interesting.
- 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.

Thanks!

Technical Appendices

The Theoretical Literature (1)

A growing literature argues that accounting for heterogeneity enables us to resolve some long-standing issues

1. the large volatility of unemployment relative to wages (Adjemian et al. (2021))
2. the pattern of recovery from deep and shallow recessions (Hall and Kudlyak (2021)).)
3. the impact of job loss of subsequent earnings (Gregory, Menzio and Wiczer (2021))
4. the negative relationship between the duration of unemployment and the job finding rate of unemployed workers (Gregory, Menzio and Wiczer (2021))
5. the slope of the Phillips Curve (Abriti and Consolo (2022))
6. the equilibrium rate of unemployment (Abriti and Consolo (2022))

The key assumption in the literature is that that the most important aspect of labour market heterogeneity reflects differences in productivity between workers.

⏩ Go Back

The Theoretical Literature (2)

How? Because the surplus from job matches with less productive workers is smaller

1. the large volatility of unemployment relative to wages: a small surplus generates a higher volatility of unemployment relative to wages for less productive workers
2. the pattern of recovery from deep and shallow recessions a small surplus implies matches break down more readily, so the recovery from small recessions is slower because there is a larger share of less educated workers among the unemployed
3. the impact of job loss of subsequent earnings it is harder for less productive workers to find good job matches, so the impact of unemployment on earnings is more severe for these workers
4. the negative relationship between the duration of unemployment and the job finding rate of unemployed workers through a composition effect whereby the share of less productive workers in the unemployed increases over time during a recovery

The Theoretical Literature (2)

1. the slope of the Phillips Curve since the elasticity of marginal cost to the output gap differs between workers of differing productivities
2. the equilibrium rate of unemployment since mismatch between workers of differing productivities affects the Beveridge Curve

⏩ Go Back

The Theoretical Literature (4)

- There are three main unresolved model design issues in the literature
  1. Is productivity random? Pries (2004), Gertler et al. (2020) and Faccini and Melosi (2021) assume that productivity differs randomly between "good matches" and "bad matches'.
  2. Is the labour market segmented, so only highly skilled 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 it is, while Grigsby (2021) and Martin and Okolo (2022) assume it is not.
  3. Do differences in ability or skill reflect differences in education? Dolado, Motyovski and Pappa (2021), Adjemian et al. (2021), Abriti and Consolo (2022)) and Martin and Okolo (2022) assume that they do, whereas Gregory et al.(2020) and Gregory et al. (2021)) assume they do not.
- In this paper, we present evidence on the relationships between productivity, occupation and education that is relevant to these debates.

⏩ Go Back

The Silhouette Statistic

- We define ι_i^k to be the average of the Hamming distances between data point i in cluster k and all other points in cluster k.
- And we define $\zeta_i^{k'}$ to be the average of the Hamming distances between data point i in cluster k and all points in cluster k', where k' is the cluster that is closest to cluster k.
- Then the Silhouette measure for data point i in cluster k when there are K clusters is

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

⏩ Go Back

The Centroid

- The most popular type of clustering algorithm is the k-means algorithm
- Here, the centroid is the mean of the data points in the cluster
- This is useful for continuous data
- But our data is categorical
- k-means is problematic with categorical data, as the mean may not be a member of its own cluster
- So we use a k-medoid algorithm: the centroid is the median of the cluster

Distance

- The most popular measure of distance is Euclidian: the root of the sum of squared distances between data points and the centroid
- Again, this is useful for continuous data, but not for categorical data
- We use the Hamming Distance
- This is widely used in Information Theory and Cryptography
- In the case of strings of binary numbers (categorical data), the Hamming Distance is the proportion of cases in which values of two strings at the same point differ:
  - the Hamming Distance between (1,0,0,0) and (0,1,0,0) is 0.5
  - the Hamming Distance between (1,0,0,0) and (1,1,0,0) is 0.25

Algorithm

- We use the Clustering Large Applications' (CLARA) variant of the popular Partitioning Around Medoids (PAM) algorithm
- This has been proposed to address the challenges posed by large datasets through sampling.
- The CLARA algorithm proceeds as follows:
    1. an initial subset of the data is selected and the PAM algorithm is applied to this subset
    2. each data point in the full sample is then allocated to a cluster by selecting the closest medoid, as measured by the Hamming distance
    3. the algorithm repeats these steps until the medoids do not change.
- The initial subset of data is selected using the "k-means++" approach, which involves random sampling of the data.
- The algorithm is repeated 90 times, with the best result being selected, in order to give assurance that a global solution has been found

⏩ Go Back

Silhouette Statistics

- Define the average Silhouette value for cluster k as

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

- Then

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

- And

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

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

Go Back

Model Validation (1)

- To demonstrate the credibility of these results, we next perform a validation exercise.
- The clustering algorithm gives a mapping $k_i = C(i)$
  - $k_i = 1$ if data point i is assigned to cluster $2A$
  - $k_i = 2$ 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 $2A$
  - $k_i^s = 2$ if data point i is assigned to cluster $2B$.

Model Validation (2)

- 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
- 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.

⏩ Go Back

Detail on k-Means

- Although k-Means is more suited to continuous variables and is more sensitive to data outliers than k-Medians, application of this algorithm to our data provides useful insights.
- The k-Means algorithm proceeds as follows:
  1. 2 initial candidate centroids are randomly selected
  2. Euclidian Distances are calculated between each data point and these candidate centroids
  3. each data point is assigned to the cluster with the closest centroid
  4. new centroids are calculated as the average of the observations in each cluster
  5. steps (ii)-(iv) are repeated until cluster membership does not change.

⏩ Go Back

Detail on Soft Clustering (1)

- Soft clustering is a generalisation of k-Means clustering where each data point can belong to multiple cluster.
- We use the Fuzzy c-means (FCM) algorithm; the degree of fuzziness is governed by the parameter m.
  - 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 $m = 2$.

Detail on Soft Clustering (2)

- The algorithm proceeds as follows:

  1. values for the weight of each data point in each cluster are randomly selected
  2. the centroids of each cluster are calculated
  3. values for the weights are updated
  4. 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.

Go Back