Clusters in UK Self-Employment

Jack Blundell

November 2019
Abstract
UK Self-employment has soared in recent years. With existing labour market policy designed to cater for conventional employee relationships, policymakers in this field are increasingly seeking to better understand these workers’ characteristics in order to ensure that new labour market regulations are designed appropriately, and are targeted towards the groups that require social assistance. In this paper I apply a machine learning method to ask whether there exist distinct ‘clusters’ of workers within self-employment, corresponding to groups with similar observable characteristics. My analysis first uncovers a two-group typology, with a distinct divide between a low-educated male group and a high-educated female group. While groups differ on characteristics, drawing on new survey data I find that both are similarly satisfied with self-employment. I also uncover a six-group typology. This detailed clustering reveals a sub-group of low-educated young men who are dissatisfied with self-employment and are most likely to report self-employment as their only employment option, many of whom can be broadly classified as ‘gig economy’ workers.

This work was funded by the Turing-HSBC-ONS Economic Data Science Award. The author thanks Guy Michaels, Giulia Giupponi, Nikhil Datta and the Labour Markets Directorate at BEIS for invaluable comments.

Jack Blundell is a PhD student at Stanford University and a Research Assistant at the Centre for Economic Performance, London School of Economics.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means without the prior permission in writing of the publisher nor be issued to the public or circulated in any form other than that in which it is published.

Requests for permission to reproduce any article or part of the Working Paper should be sent to the editor at the above address.

© J. Blundell, submitted 2020
1 Introduction

Self-employment in the UK has risen dramatically over the last two decades. From 2000 to 2017, self-employment swelled from 12.0% to 15.1% of the labour force, with the years since the financial crisis seeing particularly rapid growth (ONS, 2018). A myriad of drivers have been proposed for the rise. These range from structural changes such as an aging work-force, tax incentives and technological developments to cyclical shifts, most prominently a lack of better work options post-Great Recession. Coincident with the rise in self-employment has been a change in the nature of self-employed workers (D’Arcy and Gardiner, 2014). Much of the new "gig economy" workforce qualifies as self-employed, and the self-employed of today represent a diverse community poorly characterised by a single homogeneous group. The seminal Taylor Review of Modern Working Practices (Taylor, 2017) notes that “The experiences and vulnerabilities of this group ranges from billionaire entrepreneurs to taxi drivers working 90 hours a week simply to pay their bills”. Self-employed workers can be old or young, can have left school at 16 or have post-graduate degrees, and can be found in a variety of sectors, from construction to banking and finance.

With self-employment on the rise, there has been increased discussion of the lack of social protection for self-employed workers. Many of the typical benefits enjoyed by employees are not available to self-employed workers, both in the UK and across the EU (Spasova and Wilkens, 2018). If we are to design effective policies directed towards the self-employed, as a first step we must understand who they are, and whether they would benefit from such social protection. With such a variety of self-employed workers, it is likely that the need for further social protection varies across types of workers. In light of this, the goal of this paper is to develop a typology of self-employed workers in the UK. Applying a machine learning clustering algorithm, I ask whether self-employed workers can be classified into distinct groups based on a number of demographic and work characteristics. Such clusters can serve as a useful descriptive tool and can have practical implications for policymakers. If relatively homogeneous sub-groups can be identified within broader self-employment, it raises the scope for policymakers to target policies towards groups they deign to be particularly receptive, or those who in need of further social protection.

I first apply a ‘Partitioning around Medoids’ (PAM) algorithm to cluster self-employed workers from the Labour Force Survey (LFS) into worker types based on a set of demographic and work characteristics. As will be detailed in Section 3, the PAM algorithm identifies groups of similar individuals and provides diagnostics to determine how well alternative groupings summarise variation in the data. As this is an under-utilised and widely-applicable tool in the analysis of labour markets, an overview is given of the key choices involved and an outline the algorithm itself is provided.

Applying the algorithm to the LFS, it is shown that self-employed workers can be clustered into two groups.1 The two groups differ on gender, education and job characteristics, with one group being predominantly male and less educated and the other female and highly educated. The groups are segregated across occupations and industries, with the former often found in construction and the latter in professional services. To complement the LFS analysis, by drawing on the recent LSE-CEP Survey of Alternative Work Arrangements, I am able to present evidence on differences in the motivations, preferences and constraints at greater detail than would be possible using conventional survey data. Applying the LFS clustering to LSE-CEP survey, overall satisfaction with self-employment is found to be high across both major groups. In each group, more than three quarters of workers surveyed state that they are happier in self-employment than they would be as an employee. Flexibility is highly valued across both groups of self-employed workers.

---

1The terms ‘clusters’, ‘types’ and ‘groups’ are used interchangeably throughout this paper
While the two-group typology is useful, a second subdivision into six smaller groups is presented, primarily determined by occupation / industry differences. I find that while aggregate satisfaction is high among the self-employed, there exists a small group of predominantly low-educated young men outside of construction who report dissatisfaction with self-employment. Many of this group would rather be in traditional employment. This group are also less satisfied with their hours and most likely to report that they are self-employed due to a lack of better options. Given the immense policy interest in potential exploitation and one-sided flexibility in the gig economy, it is notable that this group is predominantly made up of drivers, a group who have been at the forefront of technological disruption in the labour market. This is a group that has been driving the increase in gig work in the US (Abraham, Haltiwanger, Sandusky, and Spletzer, 2019), and with the advent of self-driving cars are likely to be subject to further disruption in the future. Of course, one interpretation of these patterns is that while not the most desirable jobs, for these workers self-employment represents a potentially valuable insurance option.

The remainder of this paper proceeds as follows. Section 2 briefly overviews the most relevant aspects of the voluminous literature on self-employment. In Section 3 I describe our clustering approach, before discussing the two data sets in Section 4. In Section 5 I discuss our results. In Section 6 I test the robustness of our clustering approach, and I conclude in Section 7.

2 Related literature

There exists an extensive academic literature on theoretical and empirical issues surrounding self-employment, both on broader issues and on the UK in particular. Much of the early academic literature on self-employment is motivated by its association with entrepreneurship and innovation, and ultimately by its link to economic growth. A literature summarised in van Praag and Versloot (2007) tends to find that higher rates of entrepreneurship, typically defined as self-employed individuals or owner-managers, are associated with higher productivity and employment growth. The perception of self-employment as a particularly economically beneficial segment of the labor market stimulates numerous papers seeking to determine who becomes self-employed and the motivating factors behind the decision. Early papers such as Evans and Jovanovic (1989) explore the role of capital and risk in determining who becomes self-employed. Other papers have demonstrated particularly high self-employment rates among immigrants and ethnic minorities (Borjas (1986), Clark and Drinkwater (1998)). Also of interest in the literature is the question of what makes a successful self-employed individual (Lazear, 2004).

A more-recent literature has looked within self-employment, motivated by the observation that self-employed workers are an increasingly varied group. Dawson, Henley, and Latreille (2009) for example find substantial heterogeneity in the motivations behind self-employment. One notable difference is between genders, with female self-employed workers tending to be more concerned with “lifestyle factors” and less with financial gain. Related to this, Datta (2019) demonstrates that workers value both security and flexibility, but that there is substantial variation in preferences across individuals. Lenton (2017) uses British survey data to demonstrate substantial differences in the characteristics and motivations between types of self-employed workers. She finds for example that education strongly dictates whether conditional on being self-employed, a worker is a sole trader or a sub-contractor/freelancer. Using Swedish administrative data, Humphries (2018) applies machine learning methods to cluster self-employed workers based on life-cycle employment profiles, finding that self-employed workers fall into a small number of economically distinct groups. The methodology here is related to this previous work, but the research question differs substantially.
Alongside the above academic literature, in recent years there has emerged a large policy literature on self-employment in the UK. Much of this is motivated by a perception that self-employed workers represent a particularly vulnerable group in society. The Taylor Review (Taylor, 2017) establishes self-employment as a central issue facing policymakers, focusing on the increase in variety of self-employment and the risk of one-sided flexibility facing contractors. The Resolution Foundation’s 2014 report (D’Arcy and Gardiner, 2014) provides a thorough analysis of recent trends in UK self-employment, and on the aggregate characteristics of the group. Building on ideas in Hutton (1996), Tomlinson and Corlett (2017) propose the presence of two types of self-employed workers, those in ‘precarious’ and those in ‘privileged’ sectors. Precarious sectors include retail, cleaning and construction, whereas privileged sectors include health, IT and consultancy. Workers in precarious sectors are younger, more likely to be underemployed, are less educated, have less housing wealth and are more likely to be on some form of government assistance. While this and other work in the policy literature have indeed discussed there being different groups in self-employment, this is the first paper (to the author’s knowledge) to formalize this intuition by applying a clustering algorithm.

3 Methodology: Identifying clusters

In this section an overview of the PAM algorithm is provided, focusing on the key choices faced by the researcher. The interested reader can refer to Hennig and Liao (2013) for further technical details. The general approach of clustering methods in machine learning is first to build a distance measure, expressing the dissimilarity of all observations from one another based on characteristics, then to employ an algorithm to assign observations to groups such that the distance between observations within groups is minimized. This results in a set of groups containing similar observations, based on the characteristics chosen. These methods have been widely used outside of Economics, for example in Pew (2018) where a typology of religious beliefs is constructed, and are common in market segmentation exercises in industry. The growth of such methods within Economics is growing, for example in Bonhomme and Manresa (2015) where clustering methods are used to group fixed effects in panel data.

While it is attractive to see clustering methods as ‘letting the data speak’, in practice identifying clusters involves many choices by the researcher and some degree of subjectivity. Nonetheless, it offers a more transparent way of assigning individuals to groups based on observable characteristics.

3.1 Choosing a distance measure

The current application is unusual as a clustering problem, as the data contains combination of categorical data (e.g. occupation), ordinal data (e.g. education) and continuous data (e.g. age). While clustering is more commonly applied to continuous data, this mixed-data-type environment is the dominant one in social science settings, particularly those using survey data. In this environment, the conventional strategy of using Euclidian distance to express the dissimilarity between two observations along a particular dimension is inappropriate, so instead a Gower distance measure is used. Distance is given by the average of ‘partial dissimilarities’ across observations. The partial dissimilarity between an observation $i$ and another observation $j$ is given by the following formula:

$$d(i, j) = \frac{1}{p} \sum_{f=1}^{p} d(i, j)^f$$  \hspace{1cm} (1)
Here, \( p \) is the number of characteristics (or ‘features’), and \( f \) indexes an individual characteristic. If a characteristic is continuous, the contribution of that characteristic to the partial dissimilarity is:

\[
d(i, j)^f = \frac{|x_i^f - x_j^f|}{\max_n(x^f) - \min_n(x^f)}
\]  

The numerator of the formula above is the absolute value of the difference across individuals in that particular characteristic \( x^f \). The denominator is the maximum difference found across all individuals in the dataset.\(^2\) The partial dissimilarity by construction has a range of \([0, 1]\). Unlike in many distance measures, scaling of continuous variables before calculation of the distance measure is not required, as it is subsumed in the calculation of \( d(i, j)^f \). For unordered categorical characteristics, \( d(i, j)^f = 1 \) if the characteristics for \( i \) and \( j \) match, and \( d(i, j)^f = 0 \) otherwise. Ordinal characteristics are ranked according to order, normalized to range \([0, 1]\), then treated as continuous. This enforces a constant distance between adjacent categories, which may or may not be desirable depending on context. Categories may need to be re-formed based on institutional knowledge and research question in light of this implicit scaling. With any clustering, there is always an implicit weighting on different characteristics. Robustness to alternative weightings is demonstrated in Section 6.

### 3.2 The PAM algorithm

With the \( n \times n \) dissimilarity matrix formed from partial dissimilarities \( d(i, j) \) in hand, an algorithm is then chosen which groups observations to minimize distance within groups. For this mixed-type setting, the most appropriate algorithm is Partitioning around Medoids (PAM), also known as K-medoids (Kaufman and Rousseeuw, 1990). This algorithm is valuable in that it is more robust to noise and extreme observations than many alternatives, and provides an ‘example’ individual for each type. The main downside of the method is that it can be time-consuming with large datasets, though this is not a constraint in this application where the data is relatively small.

Intuitively, PAM identifies a central representative observation (medoid) for each of a set number of groups \( k \), then assign groups according to proximity to medoids. The algorithm is fairly simple, and contains four steps, the first two of which are often labelled the ‘build’ phase, and the final two the ‘swap’ phase:

1. Select \( k \) observations to serve as initial medoids. These are observations for which the sum of distances to all non-medoid observations is at its minimum.\(^3\)

2. Assign each observation to a cluster based on its closest medoid according to the distance measure

3. By cluster, search for alternative medoids which can reduce average dissimilarity, based on the dissimilarity matrix. If such an observation is identified, the observation which reduces average dissimilarity the most becomes the new medoid for the cluster.

4. If any medoids have changed, return to 2., otherwise end algorithm

\(^2\)It is worth noting that this particular distance measure can be strongly affected by outliers. This is not an issue in the data used here but ought be given careful consideration in other applications. Many thanks to Nikhil Datta for emphasising this point.

\(^3\)Except for the occasional case of ties, the version of PAM used here is deterministic. It does not rely on a random choice of initial medoids.
3.3 Choosing the number of clusters

The strategy outlined above provides groupings and ‘typical’ individuals (medoids) within groups for a fixed number of groups \( k \). As outlined in Hennig and Liao (2013), for most applications, choosing \( k \) cannot be entirely data driven. In general, while data-driven tools exist which can be used to aid the choice of clusters, the choice is somewhat subjective. The appropriate number of clusters depends partly on the data but also on the question of interest. Researchers must make qualitative judgements on some ‘acceptable’ level of distance between and similarity within clusters. An additional concern is interpretability. Typically researchers are looking for a small number of clusters that can be inspected and labelled according to characteristics. In this setting, it would not be particularly valuable to policymakers to cluster observations into an unmanageable large number of groups, as the purpose of the exercise here is to reduce complexity by classifying diverse observations by several core types which can be compared to one another. It is up to the researcher to decide on a reasonable maximum possible number of clusters to fit and it will depend on the context.

While a strong degree of subjectivity is involved in this stage of the analysis, a variety of tools exist which can help guide and justify decisions over number of clusters. One such tool is the silhouette. The silhouette of each observation measures how similar it is to other observations within its cluster, relative to its similarity to observations in the closest alternative cluster. This can be averaged across observations to form an average silhouette width measure. The measure is bounded between -1 and 1, with higher values reflecting tighter clusters. The number of groups \( k \) which maximizes the average silhouette provides the tightest clusters, according to the distance measure provided. More details on silhouettes are given in Rousseeuw (1987).

4 Data

The primary dataset used in this paper is the largest household study in the UK, the Labour Force Survey (LFS). The January-March version from 2018 is used throughout.\(^4\) This is the principal source of data on UK employment and is both large in size and considered to be high quality. The LFS contains 6,676 self-employed individuals.\(^5\) The characteristics selected to enter into our clustering algorithm are important. They will dictate both cluster assignment and the interpretation of the clusters. The characteristics to be included should reflect the motivation for performing the clustering. In this setting, it is asserted that policymakers are interested in groupings based on core characteristics which they can observe, and therefore plausibly condition on when designing policy.

A naive approach would be to ‘let the data decide’ and include as wide a set of characteristics as possible. However, many variables in the LFS might be considered proxies for the same underlying characteristics. For example, in this setting including both an indicator for part-time work and an hours worked variable would lead to a double weight being put on work hours in the algorithm. It is also poor practice to include categorical variables in which one group strongly dominates, which applies to several potential variables including ethnicity. Taking these concerns into account, six predictors are selected. These are age, sex, education, industry, occupation and part-time/full-time. These variables are widely


\(^5\)In this paper, LFS weights are not used.
available in multiple datasets which would allow any typology found here to be utilised elsewhere. A disadvantage of the chosen clustering method is that it is not possible to integrate observations missing one or more characteristics. Fortunately, this applies to very few individuals in our LFS data. After these 370 individuals are dropped, 6,306 self-employed individuals remain available for analysis.

While cluster assignment is based on a narrow set of demographics and work characteristics, once clustering has been performed a broader set of characteristics can be used to compare groups. Some of these are drawn from the LFS, and data from the LSE-CEP Survey of Alternative Work Arrangements is also used. This survey is complementary to the LFS, giving more detail on the preferences, motivations and constraints of non-standard workers. The survey was run in February 2018 on a representative sample of 20,000 UK workers. Of these, 2,240 workers classified as self-employed. These workers are assigned to their closest clusters based on Gower distance across five of the six variables used for LFS clustering.\(^6\)

5 Results

5.1 Choosing the number of clusters

The first step of analysis is to explore how well the k-medoid algorithm partitions the LFS data, according to age, sex, hours, occupation, industry and part-time / full-time status. Figure 1 shows the silhouette plot.

![Figure 1: Silhouette width by cluster size](image)

Notes: Average silhouette width by number of clusters. Higher values indicates tighter clustering. Source: LFS

The highest average silhouette width is found at two clusters, at 0.293. The best fit, according to this measure, is given by two clusters. In the following section the characteristics of these two groups are explored, as well as whether they align with previous groupings proposed in the literature.

Interestingly, while three, four and five clusters deliver a poor fit to the data, from six clusters upwards the fit improves substantially, reaching a local maximum at nine clusters. While the fit is not as strong

\(^6\)Due to inconsistent coding, it is not possible to match on occupation when mapping clusters to the LSE-CEP survey.
as in the two-cluster case, this suggests that there may be value in inspecting a finer clustering. One approach would be to use the nine-cluster grouping, which delivers the second-best fit overall. However, we might worry that this is too many clusters. Instead, I follow a different approach based on the marginal gain to adding an additional cluster in terms of improved fit. Inspecting the gradient in Figure 1, we see that the steepest improvement in fit is from five to six clusters. The gradient becomes shallower for higher numbers of clusters, meaning that the additional value in terms of improved fit is lower. Arguably, the six-cluster grouping then represents a reasonable trade-off between statistical fit and having practical descriptive value. As is often the case when using clustering methods, there are of course aspects of this choice which are unavoidably subjective and depend on assertions on what a reasonable degree of complexity is. Nonetheless, as we feel the value of a further grouping may be high we also explore a six-cluster in addition to the two-cluster grouping. In Section 6, more evidence is provided on the extent to which each variable drives the clusters identified here.

Figure 2 shows the relationship between the two and six cluster cases in the LFS data. This figure demonstrates where each of the individuals in the two-cluster grouping (left) fall in the six-cluster grouping (right). The graph illustrates the strong relationship between the two groupings. Clear from the figure for example that the two-cluster group labelled ‘MaLE’ is made up predominantly of six-cluster groups labelled ‘Construction workers’, ‘Low-educated young men’ and ‘Managers’. Group name ‘FeDe’ stands for ‘Female degree holders’ whereas ‘MaLE’ represents ‘Male and low educated’. The groups have been labelled based on LFS characteristics and will be discussed in more detail in the following section. The group labels are not perfect, as seen by the fact that some of the MaLE group are found among the female service worker group in the six-cluster typology.

5.2 Cluster characteristics

5.2.1 Two-cluster typology

In the two-cluster case, the distribution of the characteristics used in the clustering exercise divided by cluster is given in Figure 3. The most salient difference between the two groups is gender. Over 90% of the MaLE group are male and more than three quarters of FeDe are female. This is consistent with the strong gender differences in self-employment found in Dawson et al. (2009). FeDe are higher educated, with more than two thirds holding post-secondary degrees, whereas the majority of the MaLE group have only a secondary education. The FeDe group are also more likely to work part-time. There exist occupation and industry differences. Almost half of MaLE work in skilled trades, whereas FeDe is disproportionately found in professional and technical occupations. In terms of industry, MaLE tends to be found in Construction whereas FeDe are mostly in either the Banking and Finance or Public Administration, Education and Health sectors. FeDe is also on average a little older than MaLE. Comparing our groups to those found in Tomlinson and Corlett (2017), MaLE roughly corresponds to the ‘precarious’ group and FeDe to the ‘privileged’ group in terms of work characteristics, particularly education, occupation and industry. The full set of differences are summarized in Table 1 and shown graphically in Figure 3.

Table 1 also contains four variables from the LSE-CEP Survey of Alternative Work Arrangements. These can first be used to shed light on whether self-employed workers are content with their level of

---

7Performing inference on clustered data is challenging as groups have been formed based on existing characteristics. Given this, all comparisons here are purely based on magnitudes of differences between groups and do not imply statistical significance.

8
working hours. This is particularly important given the recent labour market experience of the UK. While as of 2019 employment is at record levels, much of the growth in employment since the crisis of 2008/2009 is accounted for by self-employment. One explanation for the sluggish growth of wages in recent years is that the high employment rates mask significant under-employment, and that in fact there is a large pool of reserve workers seeking further employment and pushing wages down. Clear from the table and Figure 4 is that on average, satisfaction with hours of work is relatively low across both groups, with only half of respondents satisfied with their hours. Out of those who are dissatisfied, the majority of workers state that they would rather more hours. Under-employment appears rife in self-employment, and this is true for both groups. While under-employment is prevalent, in general, workers in both groups are satisfied with self-employment relative to becoming a conventional employee. More than three-quarters are content in self-employment in both groups.

A key topic of discussion when considering the self-employed is their lack of access to various benefits associated with conventional employment relationships. In the LSE-CEP survey, respondents are asked which benefit they would most like to receive. For both the MaLE and FeDe groups, the dominant response is retirement savings. Despite the ample differences in work and demographic characteristics, both groups have in common a concern over having enough money after retirement. Respondents are also asked their reason for self-employment. Flexibility dominates responses for both groups, and more-so for FeDe. For this group, more than three-quarters of respondents state either flexibility or being able to
Table 1: Two-cluster characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>MaLE (Male and low educated)</th>
<th>FeDe (Female degree holders)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LFS characteristics:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>90% male</td>
<td>23% male</td>
</tr>
<tr>
<td>Education</td>
<td>27% have a degree, 21% no qualifications</td>
<td>68% have a degree, 6% no qualifications</td>
</tr>
<tr>
<td>Occupation</td>
<td>Predominantly skilled trades (39%), followed by Managers, Directors (18%)</td>
<td>Predominantly Professional (31%) followed by Associate Professional and Technical (21%)</td>
</tr>
<tr>
<td>Industry</td>
<td>Predominantly construction (30%) and wide spread over other industries</td>
<td>Predominantly Banking and Finance (39%) followed by Public Administration, Education and Health (26%)</td>
</tr>
<tr>
<td>Age (mean)</td>
<td>47.9</td>
<td>49.8</td>
</tr>
<tr>
<td>Full-time</td>
<td>90%</td>
<td>30%</td>
</tr>
<tr>
<td><strong>LSE-CEP characteristics:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfied with hours</td>
<td>49% satisfied</td>
<td>50% satisfied</td>
</tr>
<tr>
<td>Preferred employment type</td>
<td>77% say self-employment</td>
<td>82% say self-employment</td>
</tr>
<tr>
<td>Preferred benefit</td>
<td>Predominantly retirement savings (44%)</td>
<td>Predominantly retirement savings (43%)</td>
</tr>
<tr>
<td>Reason for self-employment</td>
<td>Predominantly flexibility (34%), then working from home (24%), then better pay (17%)</td>
<td>Predominantly flexibility (43%), then working from home (33%), then better pay (9%)</td>
</tr>
</tbody>
</table>

Notes: LFS and LSE-CEP survey characteristics across two-cluster grouping.
Source: LFS, LSE-CEP Survey of Alternative Work Arrangements

work from home (a specific form of flexibility) as the primary reason for self-employment. Individuals in MaLE also value flexibility, but are relatively more likely to report better pay as a primary motivation.

To summarise our discussion of the two-cluster typology, there exist clear differences in characteristics across the two clusters. FeDe is more female, works fewer hours and is highly educated than MaLE, and is found in different occupations and industries. Satisfaction with self-employment is high for both groups, but there is also evidence of under-employment among both. Both groups are concerned about saving enough for retirement. Flexibility ranks highly for both groups as a motivation for self-employment, particularly for the female-dominated cluster. Consistent with the presence of a ‘precarious’ and a ‘privileged’ group in self-employment, there is a strong divide in work and demographic characteristics. However, in general both groups are relatively satisfied with self-employment, albeit for different reasons. There is no clear indication then that either of the two major groups in self-employment is on aggregate losing out by being self-employed, though this point will be returned to in the following section.
Figure 3: Two cluster LFS characteristics

5.2.2 Six-cluster typology

As emphasized above and in Section 6, the two-cluster typology appears to be robust and delivers intuitive clusters. However, as discussed above there is an argument for exploring a finer clustering.

Source: LFS
Given this, a six-cluster model is fit. The demographic and work characteristics of the resulting groups are described below, and are given in Figure 7 in the appendix.

1. **Female service workers** – This group is predominantly female, part time and not particularly highly educated. They tend to work in the services sector. An inspection of a more-detailed occupational classification shows the dominant occupations to be hairdressing, cleaning and childcare.

2. **London professionals** – Predominantly male, full time and highly educated, this group are geographically focused in London and the South East. They work in professional occupations typically in the banking and finance sector.

3. **Less-educated young men** – Members of this group are the most likely to have below secondary qualifications. They are predominantly male and noticeably younger than other groups. The transport and communications sector is the most common industry and 51% of this group are road transport drivers. This group likely includes some of those occupation most associated with the ‘gig economy’, such as private hire and delivery drivers. This group is also by far the least likely to be white, with more than a quarter from ethnic minorities.
4. **Managers** – This predominantly male group are older than other groups, working as managers and proprietors in distribution, hotels and restaurants.

5. **Older health/education workers** – This group is the most highly educated and work in a wide set of occupations related to health and education. They are older than other groups and the most likely to be part time.

6. **Construction workers** – The largest group of the six are the most homogeneous, dominated by tradesmen working in the construction industry.

The six groups differ (by construction) on the core LFS characteristics. In Figure 8 in the appendix, differences across the LSE-CEP survey characteristics are shown. There are quite substantial differences across groups. Across all groups, a relatively low proportion are satisfied with their hours. This is chiefly due to hours being lower than desired, though the group of managers is more likely to want fewer hours. Notable here are the group of low-educated young men (note that this group is almost entirely a subset of the MaLE group). The group is the least likely to be satisfied with their hours and most likely to want more hours. This suggests that there is substantial under-employment across the self-employed, and that this is particularly an issue among low-educated young men. This group is also the most likely to prefer a conventional employment relationship. 41% of this group would rather be employed than self-employed.

For all six groups of self-employed workers, flexibility is the main motivation for self-employment. However, the second most common motivation varies substantially across workers. Construction workers are particularly motivated by better pay, with over a third reporting this as the main reason for self-employment. The two female-dominated groups, female service workers and older health/education workers place a high value on being able to work from home, another form of flexibility. Low-educated younger men are most likely to report that self-employment is their only option, with 24% of this group reporting so. Consistent with the previous evidence, this group of workers seems to be in self-employment as a last resort, and not as an active choice. It could be that for this group, self-employment represents an important source of insurance. In terms of most desired benefit, all types of workers place retirement savings at the top of their list. Again, the response is clearest for the cluster of low-educated young men, who relative to other groups place a lower value on sick leave and more on retirement contributions. This group is then particularly concerned over savings for their future.

### 6 Robustness

In this section the robustness of the clustering approach applied above is explored. Figure 5 shows the results from an exercise in which characteristics are sequentially dropped individually and silhouette widths are calculated based on the remaining five characteristics. In general, across the set of six panels the silhouette width pattern looks similar. The top-left panel shows that dropping gender reduces but does not eliminate the spike in average silhouette width at two clusters. There is evidence then of two clusters in the data even if gender is excluded. Removing either occupation or industry results in an even greater spike at the two-cluster typology. The two-type typology is relatively robust to dropping characteristics, however the six-type grouping is less robust. This grouping relies particularly on occupation and industry differences and is clearest when education is removed from the clustering exercise.
Figure 5: Silhouette width robustness to dropping characteristics

Notes: This plot shows the average silhouette width for different cluster sizes. Each panel shows widths where a single characteristic is dropped, as indicated by labels. Source: LFS

Note that partial dissimilarity between an observation $i$ and another observation $j$ can be generalized as:

$$d(i, j) = \frac{1}{p} \sum_{f=1}^{p} w_f d(i, j)^f$$

(3)

Here, $w_f$ is a weight placed on characteristic $f$. In the main analysis above, implicitly $w_f = 1$ for all characteristics. Ultimately, there is no ‘correct’ choice for weighting. Figure 6 shows the shape of the silhouette plot for 20 randomly generated weight vectors drawn from a $[0,1]$ uniform distribution. Each line represents an a randomly-drawn weight vector. Although there are some exceptions, the dominant pattern is the same as in the main clustering exercise. There exists a strong tendency towards two groupings. The highest silhouette width is found at either 2, or the maximum (8) for almost all weighting vectors. This is then consistent with the previous robustness test, that the two-cluster grouping is relatively robust and the six-cluster grouping less so.

A central issue when clustering using categorical data is the risk of saturation. For example, by construction a dataset with two categorical variables each consisting of two categories can be perfectly divided into 4 distinct clusters, no matter the association between the two variables across individuals. Such a clustering therefore does not reflect any underlying patterns in the data. In this case there exist four unordered categorical variables (occupation, industry, sex and full-time/part-time), one approximately
Notes: This plot shows the average silhouette width for different cluster sizes for 20 randomly generated weight vectors. Each line represents an alternative weight vector. Each series scaled to have mean 0 variance 1.
Source: LFS

continuous variable (age) and one ordered categorical variable (education). There are nearly 1,000 possible combinations of the categorical variables alone in the current setting. Excluding empty cells still leaves over 500 combinations in our data, so saturation is unlikely to be driving the clustering.

7 Discussion/conclusion

The Taylor Review (Taylor, 2017) identified a crucial need for policymakers to better understand the vast heterogeneity among self-employed workers in the UK. In this paper I have contributed to this understanding by applying state-of-the-art methods to survey data to build new typologies of self-employment. I argue that self-employment can broadly be divided into two groups, differing on a wide range of characteristics. Importantly however, and perhaps contrary to some previous work, I find that general satisfaction of self-employment is high in both groups. While motivations for self-employment differ, overall satisfaction is strikingly similar across groups.

While the two-cluster typology suggested a high degree of satisfaction with self-employment, by building finer groups I have identified one which stands out as being relatively dissatisfied with self-employment and typically not working the number of hours they want. Among this group of low-educated young men, a large minority of this group would rather be employed in a conventional relationship and report self-employment as being a last resort. That this group contains many ‘gig economy’ workers ought to be of particular interest to policymakers. Many members of this group are drivers, who may be subject to disruption in the future as self-driving cars come into use.
The current analysis, particularly the evidence from the CEP-LSE Survey of Alternative Work Arrangements, has broader lessons for the UK labour market. I demonstrate evidence of significant under-employment, which suggests that measures of labour market slack need to be augmented to include aspects of self-employment. I show that under-employment is particularly important for a group of low-educated young male workers. In addition, I argue that the clustering methodology applied here is a valuable descriptive tool which can be applied elsewhere in the labour market.

In reference to self-employment, Taylor (2017) states that “Policy interventions have to be tailored to respond to those who require support”. The methodology outlined and typologies developed here will, I hope, be valuable in the design of future policies directed towards particular groups. There are many future avenues for related research. One particularly promising route would be to ask whether the typology of self-employed workers has changed over time, and to quantify the extent to which the growth of self-employment is driven by those of each type.

References


A Six-cluster characteristics

Figure 7: Six cluster LFS characteristics

(a) Sex
(b) Education
(c) Occupation
(d) Industry
(e) Age
(f) FT/PT

Source: LFS
Figure 8: Six cluster LSE-CEP characteristics

(a) More/Fewer hours
(b) Preferred benefit
(c) Preferred employment type
(d) Reason for self-employment

Source: LSE-CEP Survey of Alternative Work Arrangements
<table>
<thead>
<tr>
<th>Number</th>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>Jonathan Wadsworth</td>
<td>Immigration</td>
</tr>
<tr>
<td>46</td>
<td>Nicholas Bloom, Renata Lemos, Raffaella Sadun, Daniela Scur, John Van Reenen</td>
<td>International Data on Measuring Management Practices</td>
</tr>
<tr>
<td>45</td>
<td>Alex Bryson, John Forth</td>
<td>The UK’s Productivity Puzzle</td>
</tr>
<tr>
<td>44</td>
<td>Max Nathan, Anna Rosso</td>
<td>Mapping Information Economy Businesses with Big Data: Findings for the UK</td>
</tr>
<tr>
<td>43</td>
<td>Luis Garicano, Esteban Rossi-Hansberg</td>
<td>Knowledge-based Hierarchies: Using Organizations to Understand the Economy</td>
</tr>
<tr>
<td>42</td>
<td>Chiara Criscuolo, Peter N. Gal, Carlo Menon</td>
<td>Dynemp: A Stata® Routine for Distributed Micro-data Analysis of Business Dynamics</td>
</tr>
<tr>
<td>41</td>
<td>Nicholas Bloom, Renata Lemos, Raffaella Sadun, Daniela Scur, John Van Reenen</td>
<td>The New Empirical Economics of Management</td>
</tr>
<tr>
<td>40</td>
<td>Ghazala Azmat, Barbara Petrongolo</td>
<td>Gender and the Labor Market: What Have We Learned from Field and Lab Experiments?</td>
</tr>
<tr>
<td>39</td>
<td>Christopher J. Boyce, Alex M. Wood, James Banks, Andrew E. Clark, Gordon D.A. Brown</td>
<td>Money, Well-Being, and Loss Aversion: Does an Income Loss Have a Greater Effect on Well-Being than an Equivalent Income Gain?</td>
</tr>
<tr>
<td>38</td>
<td>Nicholas Bloom</td>
<td>Fluctuations in Uncertainty</td>
</tr>
</tbody>
</table>
37 Nicholas Oulton  Medium and Long Run Prospects for UK Growth in the Aftermath of the Financial Crisis

36 Phillipe Aghion Nicholas Bloom John Van Reenen  Incomplete Contracts and the Internal Organisation of Firms

35 Brian Bell John Van Reenen  Bankers and Their Bonuses

34 Brian Bell John Van Reenen  Extreme Wage Inequality: Pay at the Very Top

33 Nicholas Oulton  Has the Growth of Real GDP in the UK been Overstated Because of Mis-Measurement of Banking Output?

32 Mariano Bosch Marco Manacorda  Social Policies and Labor Market Outcomes in Latin America and the Caribbean: A Review of the Existing Evidence

31 Alex Bryson John Forth Minghai Zhou  What Do We Know About China’s CEO’s? Evidence from Across the Whole Economy

30 Nicholas Oulton  Horray for GDP

29 Stephen Machin  Houses and Schools: Valuation of School Quality through the Housing Market – EALE 2010 Presidential Address

28 John Van Reenen  Wage Inequality, Technology and Trade: 21st Century Evidence

27 Barry Anderson Jörg Leib Ralf Martin Marty McGuigan Mirabelle Muûls Laure de Preux Ulrich J. Wagner  Climate Change Policy and Business in Europe Evidence from Interviewing Managers

For more information please contact the Publications Unit
Tel: +44 (0)20 7955 7673;  Email: cep.info@lse.ac.uk
Website http://cep.lse.ac.uk