RSA

21st century enlightenment

Appendix: technical report for 7 portraits of economic security and modern work in the UK segmentation

January 2018

FUTURE WORK CENTRE

Technical report for segmentation

Objective

RSA asked Populus to create a segmentation based on a survey of attitudes to work conducted in May 2017 The objective was to generate a segmentation about the public's experience of work which identified groups in the workforce with respect to their combinations of work experience and economic security.[1]

The survey contained sections asking about:

- Their current job and attitudes towards work
- Their financial position (including income and savings
- The views about the role of employers and the government in the workplace
- The impact of Brexit, immigration, technology, housing costs and income levels on work

Input variables

A segmentation is a way of grouping respondents by way of the similarity of their characteristics and attitudes as measured by survey responses. Regardless of methodology, the final segmentation will therefore almost entirely be a function of the choice of the survey variables used to construct the segmentation. Considerable thought and effort was therefore put into ensuring the most appropriate variables were selected.

The initial pool of variables was selected on the basis of the questions which were best felt to reflect the dimensions upon which the segmentation should differentiate. All these variables were examined in terms of the distribution of their responses. For variables to make effective segmentation variables, they need to have a good distribution and not be too skewed towards one end of the scale. Binary or categorical variables can also prove problematic in this sense as they can be too rigid when trying to measure latent constructs. As a result, variables which were deemed to have too much of a skew were excluded and the remaining ones fed into a factor analysis.

The full list of variables can be found in an explorable data dashboard available on the RSA website.[2]

1. This document is adapted from the technical report provided by Populus.
2. https://www.thersa.org/discover/publications-and-articles/reports/seven-portraits-of-economic-security-and-modern-work-in-the-uk/interactive

Method: Data preperation and factor analysis

The resulting attitudinal variables were fed into a principle components analysis, commonly referred to as a factor analysis. This has two benefits. Firstly, it creates orthogonal dimensions, meaning that the segmentation will not be affected by the questionnaire design or high correlation between certain variables. Secondly, it makes interpretation of the resulting segments easier.

The factor analysis was run using mean substitution for any missing data and using Varimax rotation to create orthogonal factors.

Following a number of model runs, a 12 factor solution was selected This solution had 66% efficiency (i.e. 66% of the variance from the 44 variables was explicable in the 12 factors.).

The factors were given the following names:

- Factor 1: Fulfilment
- Factor 2: Job security
- Factor 3: Respect and fairness
- Factor 4: Autonomy
- Factor 5: Work relations
- Factor 6: Financial security
- Factor 7: Wellbeing
- Factor 8: Career progression
- Factor 9: Career change
- Factor 10: Certainty about working hours
- Factor 11: Commute and caring responsibilities
- Factor 12: Financial support

Following subsequent analysis of the factors, it was decided to further decided to split factor 6 into two separate factors by running a PCA on just the questions that were contained within that factor.
The new factors were labelled as follows:

- Factor 6a: Pay
- Factor 6b: Financial stability

The third factor on pensions was disregarded.

There was also a wish for the model to encompass the measures of savings and income, which were captured in the questionnaire to obtain a harder measure of financial circumstances. Incorporating hard measures such as this poses a number of challenges, which were overcome in the following ways.

The first problem – especially with potentially sensitive questions about personal finances – is the sizable proportion of respondents who will refuse to divulge this information. In this instance, the non-response was relatively low, so only 8% of respondents refused to give their income and 9% refused to give information about their savings. In both instances, respondents who refused to give information had their income bracket imputed, based on the mean value for each respective question.

The second problem is the scale. The income and savings questions are neither on a similar scale of the attitudinal questions (13 point for income and 10 point for savings) and neither lend themselves easily to factor analysis.

The solution therefore is to 'normalise' both variables. In this case, we do this by taking the overall mean from each observation and dividing by the overall standard error. This has the benefit of generating variables with the same mean (0) and standard deviation (1) as the factors, so puts them on an even playing field in terms of influencing the segmentation.

The third issue to be tackled is implicit weighting of the variables. Applying factor analysis to the attitudinal questions means that each of the factors / dimensions identified are given equal weighting in the cluster analysis. When we use variables outside of the factor analysis, especially ones which are correlated (which these are, r = 0.39), then this can have an undue influence on the segmentation outputs. The solution to this is manual and explicit weighting of the input variables, achieved through manually altering their standard deviation. To obtain the correct level of explicit weighting requires a degree of trial and error with observation on how much influence each of the variables is having on the segmentation. We also needed to consider in this instance the non-zero correlations, in this case between income, savings and the additional financial factors. In this instance, the weight attributed to the savings and income variables was 0.8 (80%).

Cluster analysis

Once the input variables were finalised, cluster solutions were run. The broad method chosen to run the segmentation solutions was k-means. This is a cluster algorithm which optimises a specified number of segments to maximise homogeneity within clusters and maximise heterogeneity between segments. However, k-means by itself suffers from a number of weaknesses; the main one is that it is highly sensitive to the initial cluster centres of the data. That is to say it is very biased towards the initial 'guesses' (a requirement for k-means cluster analysis) and can end up producing solutions which are functions of local maxima, in respect of finding optimal clusters, rather than global maxima.

To get around this, we employ a method of clustering that makes use of multiple starting points, based on multiple criteria, producing multiple initial cluster solutions. Each of those cluster solutions is then analysed and solutions with the highest level of reproducibility, frequently an indication of stability, are selected.[3]

In turn, several different cluster solutions were run with different numbers of clusters and slight variations in respect of the inputs, until a satisfactory solution was arrived at. Around 50 such models were run and examined.

3. Technical details of the clustering method used are given in the following paper: http://www.sawtoothsoftware.com/download/techpap/ccatech.pdf

Additional changes

The assessment of a segmentation solution includes a qualitative evaluation of the outputs and each of the segments in turn. In assessing the solutions presented, in particular an eight segment solution, it was decided that one of the segments did not differentiate as much as desired. However, the running of the seven segment solution did not result in a solution which was more intuitive. In other words, there was what is sometimes referred to as a 'bucket segment' – a segment from a segmentation solution which doesn't add to the understanding of the overall analysis. There are a number of options for how to handle this:

1. Retain the solution in full but ignore that segment
2. Re-run the solution but without the respondents who fall into that cluster
3. Re-allocation respondents who fall into that cluster and reallocated them to their nearest alternative cluster

Option 3 was selected here as it means we don't lose any respondents and can maximise the information available across the segmentation. Re-allocating respondents to their nearest alternative cluster involves calculating the cluster centres for all the alternative clusters and calculating the Euclidean distance from each of the respondents who were formerly in that cluster to each of the other clusters and assigning it to the cluster which has the shortest distance. This process produced a satisfactory seven cluster solution, which was selected to be the final solution.

The cluster solution gave segments with the following weighted sizes:

1. Steady staters: 14.0%
2. Flexi-workers: 11.7%
3. Strivers: 15.5%
4. Erratically precarious: 16.1%
5. Chronically precarious: 15.4%
6. Idealists: 14.3%
7. High flyers: 13.0%

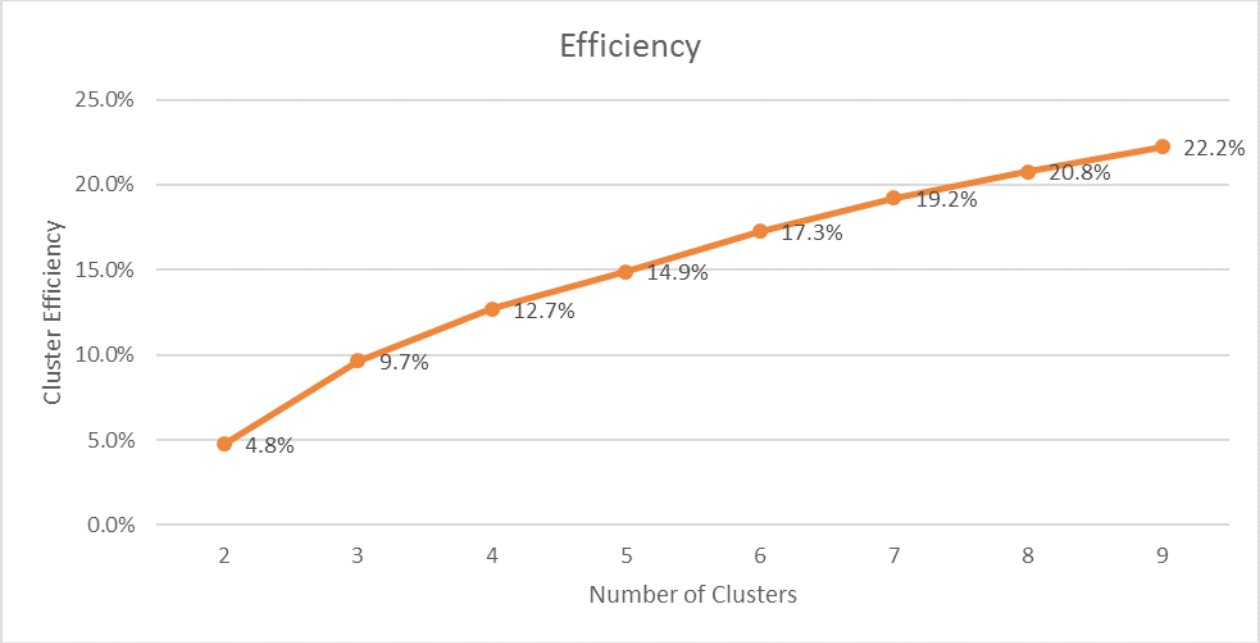
Diagnostics: Efficency

The method of clustering, as well as the algorithms used by the software, ensure that the cluster solutions used are already stable and more likely to be based on global rather than local maxima. However, there are still a number of other metrics we consider when assessing the statistical quality of the cluster analysis. One of these measures is statistical efficiency. This is calculated by way of an ANOVA analysis and enumerates the proportion of the overall sum of squares from the input variables that are explained by dividing the data into the respective number of segments. So a 'one-segment' solution would have 0% efficiency and a segment solution with the same number of segments as respondents would have 100% efficiency.

The efficiency is also a function of the number of input measures. It is much easier to get higher efficiency with fewer input variables, so it is

not the absolute measure of efficiency that we consider, rather the relative efficiency as we increase the number of segments.

In this case, from the raw segmentation solutions (prior to reallocation of the segments), we ran 2 through 9 cluster solutions from the stipulated set of inputs. Their relative efficiencies are given in the chart below:



We can observe a steady increase in the efficiency as the number of clusters increases. Things we are looking out for here include any unusual discontinuities, kinks in the curve or any flattening out as the number of segments increases. The latter can be an indication that the optimal number of segments has been reached and could indicate that segment solutions above this are unstable. No such patterns are observed here, meaning we can revert to using our understanding of the profiles of the segments in order to help choose between them.

Reproductibility

A key measure of the usability of a segmentation is how straightforwardly possible it is to generate a model which predicts the segment membership on the basis of the input variables. The most common way of undertaking this is through discriminant analysis, a technique which derives linear combinations of the model inputs in order to predict a categorical output metric (such as segment membership).

We measure the effectiveness of the model chiefly through its ability to correctly allocate respondents to their 'correct' segment and by comparing this against the chance that it would be correct at random.
In this instance we have a seven segment solution (after reallocation) so, if we were to create an entirely random allocation model, we would expect only 14% of respondents to be allocated to the correct segment.

If we use the factors as inputs (these are the constructed scores from the PCA analysis), we end up with an overall classification rate of 90.7%, with the individual segments varying from 82% to 95% in accuracy.
More realistically, we would want to ascertain how the model performs when we use actual survey data, not just constructed variables like factors.

When we feed the raw survey variables into the model and try to predict the segment membership, we actually achieve a higher allocation accuracy of **91.7%**, with segments varying from 87% - 95%. Even when using a stepwise approach, which reduces the number of inputs from 46 to 30, the overall accuracy remains high at 87.0%. If the number of question inputs is reduced to 10, the allocation accuracy is 72.5%, with individual segment accuracy ranging from 62% - 80%.

The exact tables are at the end of this document.

These figures are useful in their own right in respect of building allocation models. **However, the high numbers achieved tell us that the segmentation models are sufficiently robust and reliable for us to be able to predict the segments with good levels of accuracy.**

Outputs

Profiling of the candidate segmentation solutions are key to understanding the segments and to being able to decide on a final solution. During the process, profiles were created for all candidate solutions and these were examined and poured over in detail.

Segmentation outputs generally contain the following:

- Segment sizes
- Weighted and unweighted counts
- Profiles against input variables
- Where the input variables are factors, profiles against the raw survey data
- Profiles against other survey variables

Segment profiling allows us to understand how each variable, whether an input or an output variable, correspond with each segment. High scores indicate a high association with that segment and vice versa. These can either be used to help name the segments or can be used to help understand or identify them. The factor scores all have a mean of zero, by definition. Typically scores below -0.5 or above +0.5 indicate a noteworthy relationship with that factor.

Full profiles of the segments can be found in the explorable data dashboard available on the RSA website.

Classification and Reproducibility Tables from Discriminant Analysis

Model: Using factors as inputs

Classification Results[a]

		RSA_2306_V2b_8 RSA_2306_V2_8 (Cluster 7 reallocated) - FINAL SEGMENT	Predicted Group Membership							
			1	2	3	4	5	6	7	Total
Original	Count	1	156	0	0	11	1	0	0	168
		2	1	114	2	14	0	5	3	140
		3	8	3	166	0	3	5	0	184
		4	4	5	4	166	9	1	4	192
		5	0	0	4	7	173	0	0	184
		6	1	2	2	2	2	162	0	170
		7	2	0	1	2	1	4	146	155
	%	1	92.8	.0	.0	6.4	.8	.0	.0	100.0
		2	.7	81.7	1.6	10.1	.0	3.7	2.1	100.0
		3	4.1	1.6	90.2	.0	1.6	2.5	.0	100.0
		4	2.3	2.5	2.0	86.3	4.5	.4	1.9	100.0
		5	.0	.0	2.3	4.1	93.6	.0	.0	100.0
		6	.3	1.3	1.1	1.1	.9	95.2	.0	100.0
		7	1.3	.0	.6	1.0	.4	2.8	93.9	100.0

a. 90.7% of original grouped cases correctly classified.

Model: Using raw survey data as inputs

Classification Results[a]

		RSA_2306_V2b_8 RSA_2306_V2_8 (Cluster 7 reallocated) - FINAL SEGMENT	Predicted Group Membership							
			1	2	3	4	5	6	7	Total
Original	Count	1	155	0	0	5	3	1	4	168
		2	1	121	2	8	2	3	2	140
		3	3	6	167	0	5	2	2	184
		4	7	5	0	176	2	1	1	192
		5	0	1	6	2	174	1	0	184
		6	1	2	0	4	3	157	2	170
		7	1	0	4	1	1	3	145	155
	%	1	92.2	.0	.0	2.9	1.6	.9	2.4	100.0
		2	1.1	86.7	1.5	5.8	1.7	2.2	1.1	100.0
		3	1.8	3.2	90.3	.0	2.5	1.1	1.1	100.0
		4	3.8	2.5	.0	91.3	1.2	.7	.5	100.0
		5	.0	.4	3.5	1.2	94.5	.5	.0	100.0
		6	.8	1.4	.0	2.5	1.9	92.2	1.3	100.0
		7	.9	.0	2.5	.6	.6	1.7	93.7	100.0

a. 91.7% of original grouped cases correctly classified.

Model: Using raw survey data as inputs (stepwise – 30 variables)

Classification Results[a]

		RSA_2306_V2b_8 RSA_2306_V2_8 (Cluster 7 reallocated) - FINAL SEGMENT	Predicted Group Membership							
			1	2	3	4	5	6	7	Total
Original	Count	1	153	3	2	4	3	1	2	168
		2	4	112	4	10	2	6	2	140
		3	9	10	157	1	6	2	0	184
		4	6	7	1	171	4	1	2	192
		5	3	2	13	0	164	0	3	184
		6	2	5	3	3	2	153	2	170
		7	9	4	5	3	4	2	128	155
	%	1	91.2	2.0	1.0	2.3	1.8	.5	1.2	100.0
		2	2.6	80.3	2.8	7.5	1.2	4.1	1.6	100.0
		3	5.1	5.6	85.0	.4	3.1	.8	.0	100.0
		4	3.0	3.7	.5	88.9	1.9	.7	1.2	100.0
		5	1.6	.8	6.8	.0	89.1	.0	1.7	100.0
		6	1.3	2.6	1.6	1.7	1.4	90.0	1.3	100.0
		7	6.1	2.3	3.4	2.2	2.7	1.0	82.3	100.0

a. 87.0% of original grouped cases correctly classified.

Model: Using raw survey data as inputs (stepwise – 10 variables)

Classification Results[a]

		RSA_2306_V2b_8 RSA_2306_V2_8 (Cluster 7 reallocated) - FINAL SEGMENT	Predicted Group Membership							
			1	2	3	4	5	6	7	Total
Original	Count	1	129	7	3	6	9	7	8	168
		2	4	87	10	9	6	9	14	140
		3	6	15	136	1	14	5	9	184
		4	2	10	0	153	13	3	11	192
		5	12	8	12	5	137	6	5	184
		6	11	4	7	15	4	122	8	170
		7	20	4	6	2	12	9	102	155
	%	1	76.8	4.1	1.6	3.4	5.3	4.2	4.7	100.0
		2	2.8	62.3	7.1	6.5	4.3	6.7	10.3	100.0
		3	3.1	8.1	73.8	.4	7.4	2.5	4.7	100.0
		4	1.2	5.3	.0	79.5	6.5	1.7	5.7	100.0
		5	6.4	4.1	6.6	2.6	74.2	3.1	3.0	100.0
		6	6.2	2.3	4.1	8.8	2.4	71.5	4.7	100.0
		7	12.7	2.7	4.1	1.0	7.8	6.1	65.7	100.0

a. 72.5% of original grouped cases correctly classified.