Appendix: Using Sequence Analysis to Understand Career Progression: an Application to the UK House of Commons

August 14, 2018

Optimal Matching

Optimal matching relies on the Needleman-Wunsch algorithm (Needleman and Wunsch 1970). By taking into consideration the costs set up by the researcher, this algorithm calculates the minimum number of substitutions and indels needed to align two sequences, namely to make them identical. This algorithm was originally introduced to calculate the largest number of amino acids of one protein than can be matched with those of another protein allowing for all possible substitutions and deletions in either sequence (Needleman and Wunsch 1970). In this Appendix we briefly illustrate how this algorithm works. Take as example the sequence number 1 and the sequence number 2 mentioned in the main text and take also the following substitution cost matrix:

$$\begin{array}{c} \\ BB \\ FB \\ GO \end{array} \begin{array}{c} \begin{array}{ccc} BB & FB & GO \end{array} \\ \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix} \end{array}$$

Assume that the cost of substituting the element FB with the element BB is 1 and with the element GO the cost is 2. Also, assume that the substitution cost of an element with itself is equal to 0 and that substituting FB with BB is equally costly to substituting BB with FB, for instance. By considering these substitution costs, we calculate the distance between the sequences $BB-BB-BB-FB-FB-GO$ and $BB-BB-BB-GO-GO-GO$ with the following formula:

$$S(BB,BB)+S(BB,BB)+S(BB,BB)+S(FB,GO)+S(FB,GO)+S(GO,GO)=$$

$$0+0+0+2+2+0=4$$

This is rather straightforward when we compare two sequences of identical length and

consider only substitution costs. In this case, only one alignment is possible. Yet, when we compare sequences of different lengths, different alignments are possible. To solve this issue Needleman and Wunsch (1970) introduce a way to find the optimal alignment between two sequences by using a series of pairwise comparisons between their single elements. For the sake of simplicity, we illustrate how to compare two sub-sequences of the one mentioned above, namely $FB-FB-FB-BB$ and $FB-FB-FB$, characterized by different length. In this vein, four different alignments are possible:

$FB-FB-FB-BB$

1. $FB-FB-FB-XX$
2. $FB-FB-XX-FB$
3. $FB-XX-FB-FB$
4. $XX-FB-FB-FB$

If we compare those possible alignments, we conclude that the first one is the optimal one being the distance equal to the cost of inserting a gap. The others are associated with distance indexes equal to the cost of inserting a gap plus the cost of substituting BB with FB. What follows is the illustration of how the algorithm works, as explained by Needleman and Wunsch (1970).[1] A matrix $S(i,j)$ is created recursively such that:

$$min \begin{cases} S(i-1,j-1)+s(x_i,y_j) \\ S(i-1,j)+g \\ S(i,j-1)+g \end{cases}$$

The element $s(x_i,y_j)$ represents the substitution cost between the elements in the row i and in the column j and g is the cost of inserting a gap. We assume that the cost of

1. Differently from the original discussion we do not use similarities but distances between elements, based on substitution costs. In this vein, some parts of the discussion as well as of the formalization are modified accordingly, such as that the calculations at the basis of the algorithm in this work use the minimum and not maximum function. Indeed, in the original discussion the aim of the algorithm is to maximise the similarity between sequences while in this case it is to minimize their distance.

inserting a gap is 0.5 in this example. For the sake of this example it is logical that the cost of inserting a gap is half of the lowest substitution cost (which equals 1, in this example), in that a substitution is equivalent to a deletion of an element and the insertion of a gap. We start by compiling the matrix from the top-left cell which is associated the value 0 and from the values on the first row and the first column to which the values of insertion of gaps are associated. As clarified below, indeed, if the sequences are aligned the cell takes value 0 whereas along the first row and first column a movement rightwards or downwards refers to the insertion of a gap. For instance, $S(1,5)$ is four steps on the right from $S(0,0)$: this means it will assume the value of the insertion of four gaps. Once we set those cells up, all other cells are calculated through the algorithm.

Take the example of cell $S(2,2)$, its value will be:

$$min \begin{cases} S(1,1) + s(FB, FB) \\ S(1,2) + 0.5 \\ S(2,1) + 0.5 \end{cases}$$

$$min \begin{cases} 0 \\ 1 \\ 1 \end{cases}$$

By applying the same reasoning to all the other cells we obtain the following matrix:

		FB	FB	FB	BB
	$S(1,1)0$	$S(1,2)0.5$	$S(1,3)1$	$S(1,4)1.5$	$S(1,5)2$
FB	$S(2,1)0.5$	$S(2,2)0$	$S(2,3)0.5$	$S(2,4)1$	$S(2,5)1.5$
FB	$S(3,1)1$	$S(3,2)0.5$	$S(3,3)0$	$S(3,4)0.5$	$S(3,5)1$
FB	$S(4,1)1.5$	$S(4,2)1$	$S(4,3)0.5$	$S(4,4)0$	$S(4,5)0.5$

We now show how the algorithm works in an intuitive manner. A movement on the diagonal means that the elements are aligned whereas a movement rightwards or downwards

represents either an insertion of a gap or a substitution. Starting from $S(1,1)$ we move diagonally into $S(2,2)$: the latter contains the minimum value with respect to the other options, namely $S(1,2)$ and $S(2,1)$. Inserting a gap respectively in the column or in the row sequence at this stage is not an optimal move. We do the same until we get to cell $S(3,3)$ where we have two choices: either we proceed diagonally and we add a gap thus moving rightwards or we insert a gap and then we substitute the element BB with the element FB (the dash arrows). As we can see the second strategy is more costly: the distance index is 1.5. Contrariwise, the first strategy represents the optimal one in that it minimizes the distance index, namely 0.5.

Finally, we show how to compare two sequences which differ due to the different length of episodes, namely 'chunks' of sequences (and only the sequence length, as done above). These sequences will be used below as well, to compare sequence analysis and cluster analysis. We take the career paths of observations $id = 676$ and $id = 682$, respectively

$BB - BB - BB - BB - FB - FB - FB - BB - BB - BB - BB - BB - BB - BB$

$BB - BB - FB - FB - FB - FB - FB - FB - BB - BB - BB - BB$

Remember that the cost of substituting the element FB with the element BB (and the other way round) is 1 and that the cost of inserting a gap or deleting an element is 0.5 (hence the indel costs, namely the combined costs of insertion and deletion, equals 1). First, try to replace all the elements FB in the second sequence and to add the necessary gaps to make the two sequences of the same length.

$BB - BB - BB - BB - FB - FB - FB - BB - BB - BB - BB - BB - BB - BB$

$BB - BB - BB\cancel{FB} - BB\cancel{FB} - FB - FB - FB - BB\cancel{FB} - BB - BB - BB - BB - XX - XX$

The cost of this option is 4, namely the cost of substituting the element FB with BB three times, which is 3, plus the cost of adding two gaps, which is 1.

Instead, if we consider using only the insertion of gaps and the deletion of elements, our option will be

$BB - BB - BB - BB - FB - FB - FB - BB - BB - BB - BB - BB - BB - BB$

$BB - BB - XX - XX - FB - FB - FB - \cancel{FB} - \cancel{FB} - \cancel{FB} - BB - BB - BB - BB - XX - XX - XX$

The total cost of this option is 4, namely the cost of five gap insertions and the cost of three element deletions. In this case, the two options are the same in terms of costs. Yet, it should be noted that in this example we set the cost of both the insertion of gaps and deletion of elements to 0.5 (making the indel cost equal to 1), which is half of the cost of substituting the element FB with the element BB. We would find different results, if the substitution costs were higher, which could be the case in this example if the second sequence contained some GO elements (as shown in the matrix above, substituting BB with GO is more costly than substituting BB with FB). In conclusion, we have shown that in case of two sequences which differ due to the length of episodes, the optimal alignment depends on the cost of substitutions and indels.

TraMineR

For the analysis in this paper, we use the TraMineR R package. This package provides a series of functions which allow mining and visualizing sequences, calculating distance matrices, deriving sequence clusters and running basic inferential statistics, such as the discrepancy analysis, which is used in the paper.[2].

In the analysis in the paper, we use the 'TRATE' (transition rates) method to calculate the substitution matrix, where the substitution costs between two elements depend on the transition rates between the two elements, namely the probabilities of transition from one element to the other in the dataset (Studer et al. 2011). This is common practice where there is no theoretical motivation to assign specific substitution costs. As for missing values in sequences, we choose not to treat missing values as separate elements. Treating missing

2. For more information, see http://traminer.unige.ch/index.shtml

values as separate elements would be useful if missing values were present inside the sequences and if the absolute meaning of time points (in this case, the specific years when the elements are measured) was of interest for the analysis. This is not the case in our analysis. Moreover, to compute pairwise distances between sequences we use Optimal Matching, by relying on the substitution matrix. In computing the distances, we leave indel costs at the default setting of 1 (Studer et al. 2011). Also, we cluster sequences based on the distance matrix using Ward (hierarchical) clustering.[3] Finally, as a robustness test, we run a multi-factor analysis of variance using the distance matrix with a specific function provided by the TraMineR package.

Cluster Analysis v. Sequence Analysis

In this section we provide a simple comparison between cluster analysis (used by itself) and sequence analysis (in conjunction with cluster analysis). As it is clear from the paper, one of the main steps in sequence analysis is the clustering of the sequences based on the dissimilarity/distance matrix obtained with optimal matching. In this vein, in order to show the validity of sequence analysis, we compare the results of the clustering using different dissimilarity matrices, obtained respectively with optimal matching (sequence analysis) and the Euclidean distance (the default option in cluster analysis). To be consistent, in both cases we use Ward (hierarchical) clustering.

Optimal matching is explained in detail above. The Euclidean distance is

$$\sqrt{\sum (x_i - y_i)^2}$$

It should be noted that, in order to perform the Euclidean distance on the variables measuring the different stages of the sequences, we assume that these variables are interval

3. Ward's method starts with each point (sequence, in this case) being in its own cluster and keeps merging cluster by minimizing the 'merging costs', which in turn depend on the distance between (the center of the) clusters and the number of points in those clusters (Mirkin 2013).

level (i.e. $BB = 0$, $FB = 1$ and $GO = 2$) and we assign a distinct value to missing data (i.e. they are included in the analysis).[4]

We calculate the Euclidean distance across the variables measuring the different stages of the sequences, we obtain a dissimilarity/distance matrix and then we cluster the sequences using this matrix. To be consistent with what we do in the paper, we choose eight clusters and then we compare the results. The variables resulting from the two clustering procedures are highly correlated, with p value lower than 0.01. Yet, by looking at the data, we find that sequence analysis provides a more fine-grained measure of sequences, which take into consideration the duration of career paths and the order of stages of a career within a path.

Take the career paths of observations $id = 508$ and $id = 281$ (where $BB = 0$, $FB = 1$ and $GO = 2$), respectively

$0-0-0-0-0-0-0-0-0-0-0-1-1-1-1-1-1-1$

$1-1-1-1-1-0-0-0-0-0-0-0-0-0-0-0-0-0$

The hierarchical clustering with the default option of the Euclidean distance groups these two observations into the same cluster ('Cluster 3'), whereas the clustering with the dissimilarity matrix calculated with optimal matching groups them into different clusters (respectively, 'Career 4' and 'Career 6'). The traditional clustering procedures does not take into consideration the order of the elements in a sequence, whereas sequence analysis does. Especially when studying career paths, the order of elements is a rather important piece of information. The first MP above started in the backbench and moved his/her way up to the top, whereas the second MP started in the frontbench and then, for some reasons, he/she moved to the backbench. By relying on this information, it might be concluded that, for instance, the first MP has a brilliant career ahead, whereas the second MP has somehow burnt out and his/her political career is dead.

Also, take the career paths of observations $id = 676$ and $id = 682$ (where $BB = 0$, $FB = 1$

4. We assign a numerical value to missing values only in the cluster analysis used by itself. In the sequence analysis in conjunction with cluster analysis used in the main text, we drop missing values, as explained above.

and $GO = 2$; these sequences are reported below without missing values for illustrative purposes), respectively

$0-0-0-0-1-1-1-0-0-0-0-0-0-0$

$0-0-1-1-1-1-1-1-0-0-0-0$

The hierarchical clustering with the default option of the Euclidean distance groups these two observations into the same cluster ('Cluster 6'), whereas the clustering with the dissimilarity matrix calculated with optimal matching groups them into different clusters (respectively, 'Career 6' and 'Career 3'). The traditional clustering misses an important piece of information, at least for the study of career paths: the duration of episodes (i.e. chunks of the sequence). Both MPs started in the backbench, made their way up to the top and then got back to the backbench. Yet, the second MP spent more time in the frontbench than the first MP. This is an important distinction in the study of career paths. This might suggest, for instance, that as the second MP has more experience in the cabinet, maybe he/she is more likely to get back to power soon.

In conclusion, sequence analysis (by which we mean here the use of optimal matching to create a dissimilarity matrix to be used in cluster analysis) provides a more fine-grained measure of sequences. Differently from cluster analysis used alone, sequence analysis takes into consideration the order of elements in a sequence, namely whether the MP's career is upward or downward, in this case, and the duration of episodes, namely whether the MP has spent more or less time in the frontbench, in this case. These two pieces of information are rather important in the study of political careers.

Descriptive Statistics

Table A1: University Education

University Degree	Freq.	Percent	Cum.
No University	83	11.42	11.42
Undergraduate	407	55.98	67.40
Postgraduate	237	32.60	100.00
Total	727	100.00	

Table A2: Legislature of Entry

Legislature	Freq.	Percent	Cum.
1997-2001	280	38.51	38.51
2001-2005	90	12.38	50.89
2005-2010	123	16.92	67.81
2010-2015	234	32.19	100.00
Total	727	100.00	

Table A3: Gender, State School, Job and Age at Entry

Variable	Obs	Mean	Std. Dev.	Min	Max
Gender	727	.2654746	.4418896	0	1
State School	697	.6140603	.4871661	0	1
Job	727	.2984869	.4579093	0	1
Age at Entry	725	43.27862	8.099836	25	69

Robustness Checks

Table A4: Clustering

Cluster Numbers	R Squared
Eight	0.81368903
Seven	0.80837233
Six	0.76708005
Five	0.68583622
Four	0.62082634

Figure A1: State Distribution Plot - Seven Clusters

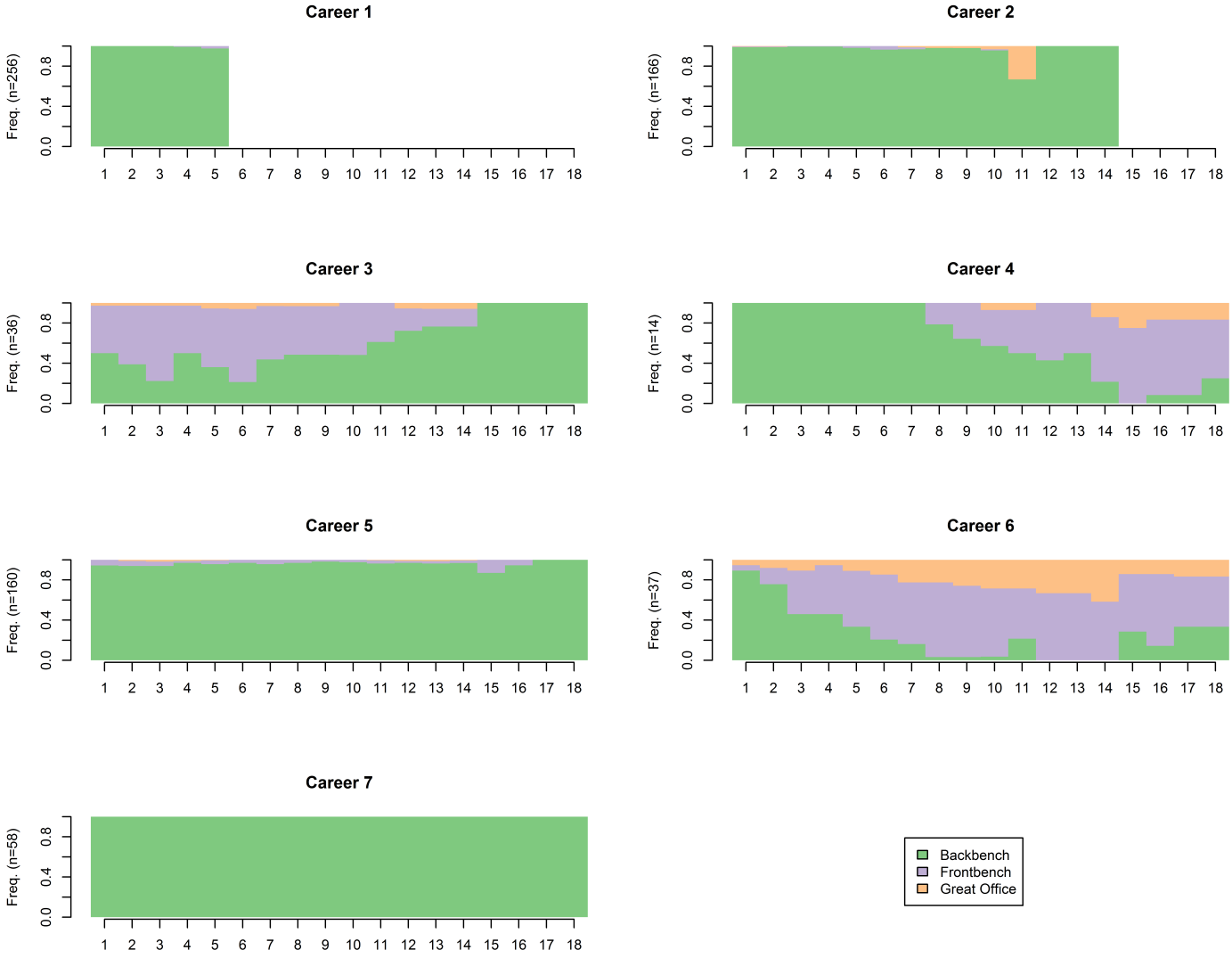

Figure A2: State Distribution Plot - Six Clusters

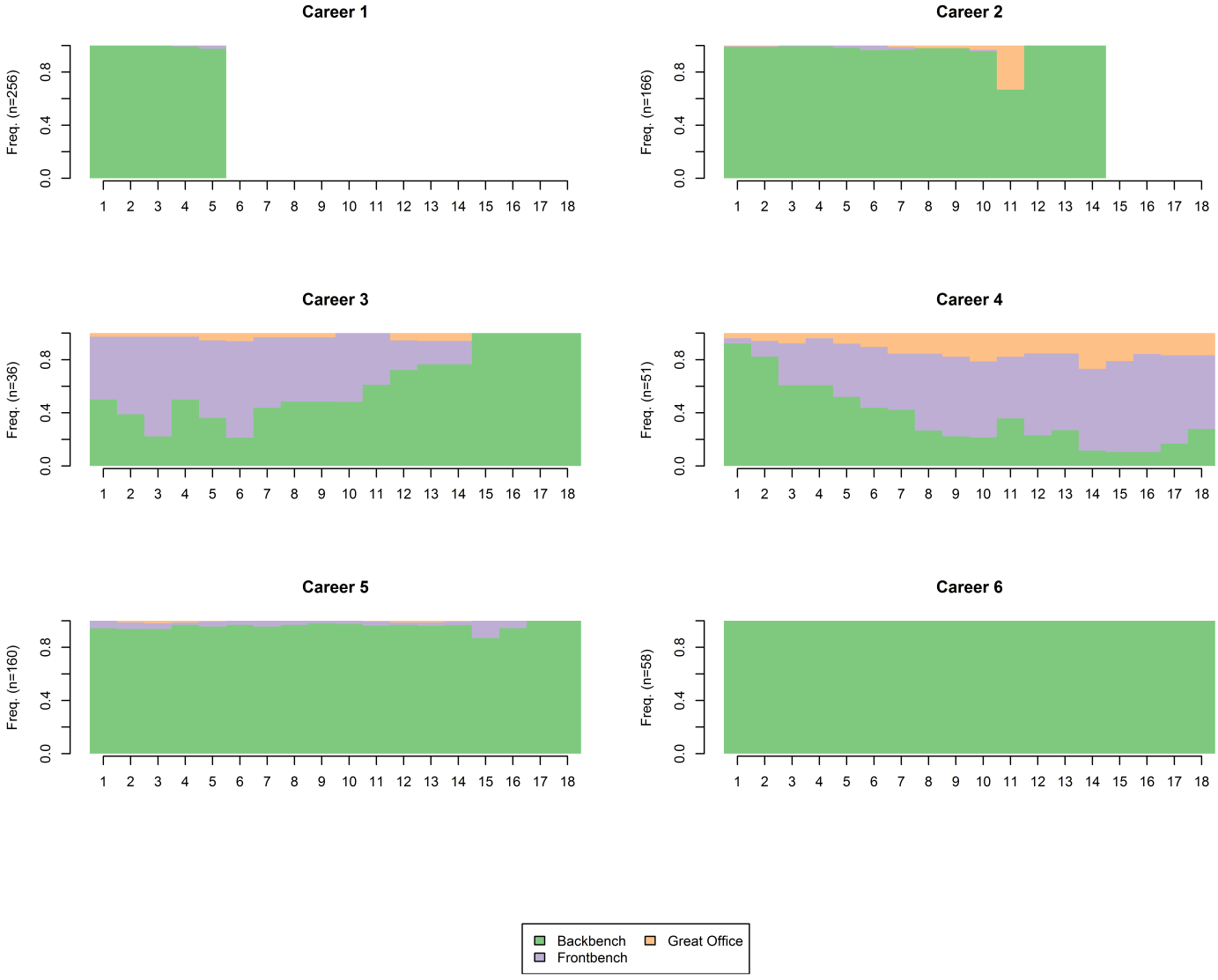

Figure A3: State Distribution Plot - Five Clusters

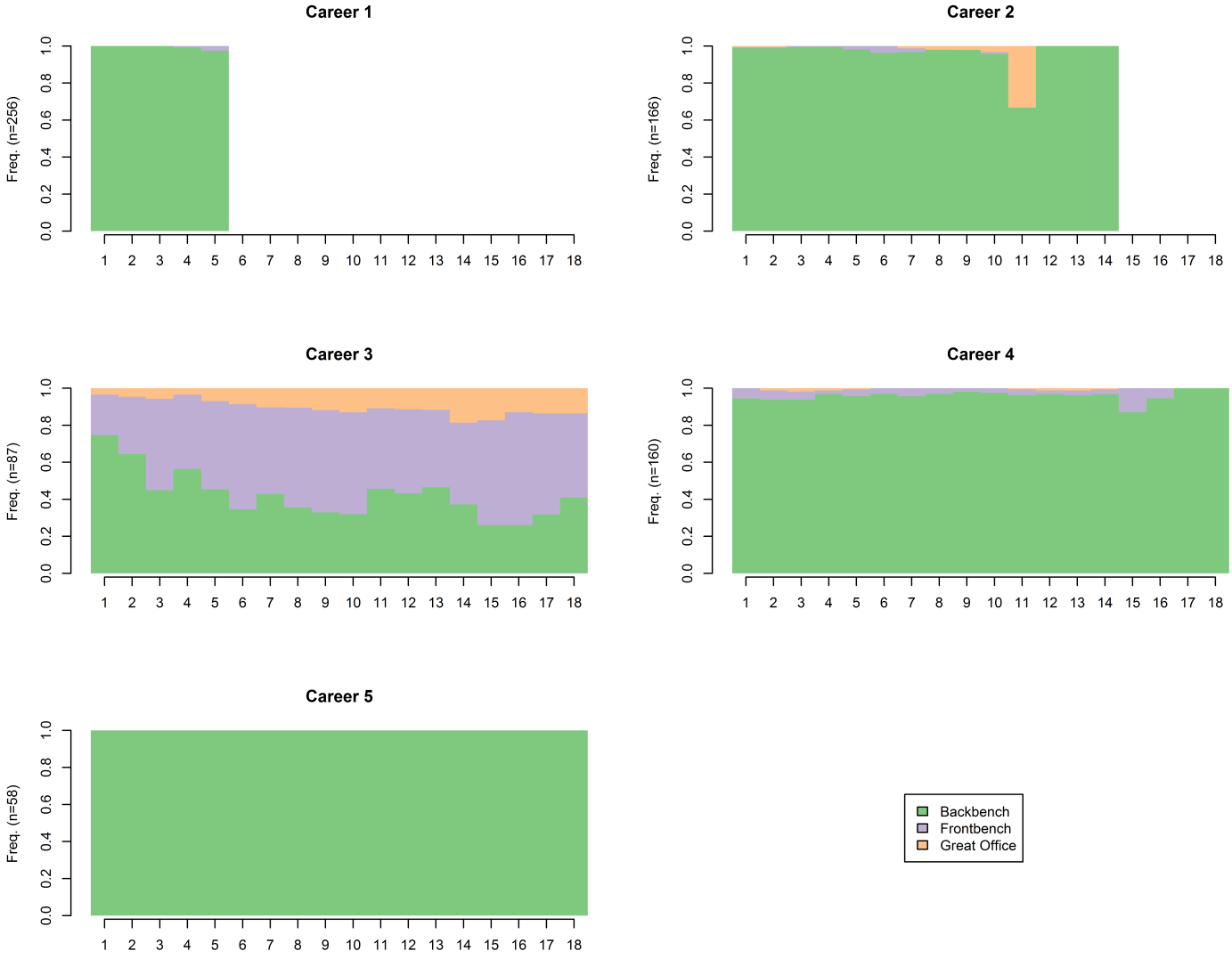

Figure A4: State Distribution Plot - Four Clusters

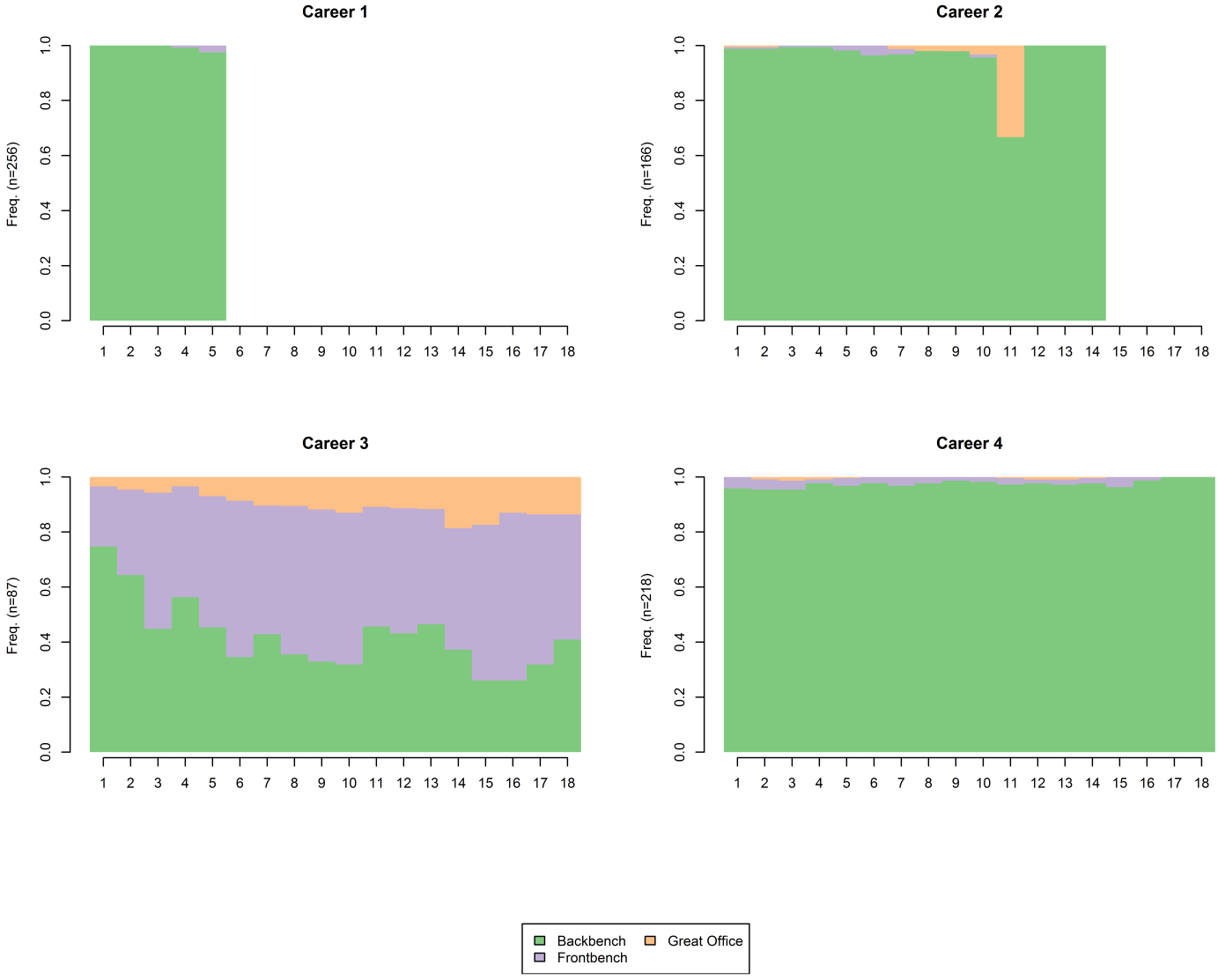

Table A5: Multinomial Analysis - Six Clusters

	Career 1	Career 2	Career 3	Career 4	Career 5	Career 6	Career 1	Career 2	Career 3	Career 4	Career 5	Career 6	Career 1	Career 2	Career 3	Career 4	Career 5	Career 6
Gender		-0.225	-0.420	-0.0326	-0.486**	-0.535		-0.523**	0.0434	-0.118	-0.941***	-1.012***		-0.367	0.432	0.216	-0.225	-0.612
		(0.239)	(0.440)	(0.360)	(0.245)	(0.361)		(0.254)	(0.498)	(0.406)	(0.267)	(0.380)		(0.544)	(0.682)	(0.579)	(0.514)	(0.598)
State School		0.108	-0.0941	-0.0884	0.467**	-0.0459		-0.287	-0.393	-0.477	-0.139	-0.508		-0.721	-0.821	-0.791	-0.179	-0.549
		(0.215)	(0.383)	(0.329)	(0.221)	(0.308)		(0.237)	(0.461)	(0.356)	(0.241)	(0.331)		(0.475)	(0.603)	(0.512)	(0.447)	(0.523)
Job		0.0206	0.198	-0.153	0.0356	0.0907		-0.117	0.234	-0.277	-0.174	-0.126		0.734	0.915	0.416	0.552	0.738
		(0.227)	(0.369)	(0.348)	(0.232)	(0.328)		(0.231)	(0.487)	(0.371)	(0.255)	(0.331)		(0.553)	(0.706)	(0.598)	(0.513)	(0.596)
Age at Entry		0.00505	0.0182	-0.0835***	0.0190	-0.00591		0.00265	0.00290	-0.0822***	0.0157	-0.00973		-0.0420	-0.0478	-0.137***	-0.0377	-0.0859**
		(0.0138)	(0.0256)	(0.0232)	(0.0128)	(0.0163)		(0.0148)	(0.0256)	(0.0224)	(0.0144)	(0.0177)		(0.0382)	(0.0422)	(0.0419)	(0.0359)	(0.0382)
University Undergraduate		0.133	0.712	0.555	0.513	-0.0472		0.166	0.949	0.578	0.549	-0.0326		-0.458	0.404	0.0717	-0.0438	-0.582
		(0.316)	(0.633)	(0.648)	(0.353)	(0.475)		(0.328)	(0.739)	(0.646)	(0.366)	(0.487)		(0.758)	(1.001)	(0.894)	(0.773)	(0.830)
University Postgraduate		0.0121	0.634	0.743	0.991***	0.586		-0.0727	0.875	0.659	0.859**	0.488		-1.076	0.0444	-0.241	-0.250	-0.674
		(0.351)	(0.671)	(0.671)	(0.374)	(0.481)		(0.370)	(0.809)	(0.680)	(0.391)	(0.497)		(0.813)	(1.063)	(0.953)	(0.803)	(0.868)
Party Conservative								-0.607	5.120***	1.266**	-1.642***	-1.498*		-2.054**	4.305***	0.255	-2.241***	-1.887**
								(0.450)	(1.050)	(0.495)	(0.564)	(0.802)		(0.817)	(1.150)	(0.780)	(0.741)	(0.914)
Party LibDem								-1.440***	0.785	-1.254***	-2.109***	-1.772***		-1.251**	1.741	-0.589	-0.741	0.134
								(0.253)	(1.126)	(0.427)	(0.266)	(0.345)		(0.556)	(1.175)	(0.621)	(0.498)	(0.576)
By-elections								0.0385	0.162	0.615	-0.289	-13.29***		-0.123	-0.164	0.448	-0.148	-16.60***
								(0.582)	(1.294)	(0.640)	(0.584)	(0.472)		(1.248)	(1.697)	(1.382)	(1.206)	(1.191)
Legislature 2001-2005														-0.545	-0.571	-0.785	-0.195	-18.65***
														(0.580)	(0.634)	(0.639)	(0.512)	(0.532)
Legislature 2005-2010														3.729***	1.481	2.083*	-15.84***	-16.74***
														(1.068)	(1.255)	(1.158)	(1.038)	(1.057)
Legislature 2010-2015														-20.04***	-19.00***	-5.839***	-20.84***	-21.92***
														(0.478)	(0.613)	(1.094)	(0.447)	(0.592)
Constant		-0.685	-3.180**	1.405	-2.034***	-1.188		0.541	-5.059***	2.032	-0.287	0.450		4.717**	-0.217	6.981***	4.351**	6.399***
		(0.735)	(1.262)	(1.190)	(0.698)	(0.897)		(0.793)	(1.745)	(1.315)	(0.798)	(0.964)		(1.968)	(2.387)	(2.180)	(1.825)	(1.967)
Observations	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A6: Multinomial Analysis - Seven Clusters

	Career 1	Career 2	Career 3	Career 4	Career 5	Career 6	Career 7	Career 1	Career 2	Career 3	Career 4	Career 5	Career 6	Career 7	Career 1	Career 2	Career 3	Career 4	Career 5	Career 6	Career 7
Gender		-0.225	-0.419	-0.114	-0.485**	-0.0100	-0.535		-0.529**	0.0921	-0.639	-0.950***	0.0884	-1.022***		-0.392	0.469	-0.392	-0.272	0.368	-0.674
		(0.239)	(0.440)	(0.670)	(0.245)	(0.407)	(0.361)		(0.255)	(0.501)	(0.672)	(0.268)	(0.473)	(0.381)		(0.547)	(0.686)	(0.844)	(0.520)	(0.636)	(0.606)
State School		0.107	-0.0968	0.974	0.465**	-0.427	-0.0452		-0.282	-0.437	0.330	-0.130	-0.731*	-0.499		-0.725	-0.878	0.0680	-0.126	-1.054*	-0.484
		(0.215)	(0.383)	(0.682)	(0.221)	(0.370)	(0.308)		(0.237)	(0.462)	(0.683)	(0.240)	(0.411)	(0.330)		(0.476)	(0.605)	(0.794)	(0.449)	(0.547)	(0.525)
Job		0.0207	0.199	-0.416	0.0358	-0.0562	0.0904		-0.121	0.257	-0.556	-0.180	-0.139	-0.134		0.729	0.937	0.255	0.538	0.513	0.711
		(0.228)	(0.369)	(0.625)	(0.232)	(0.392)	(0.328)		(0.231)	(0.486)	(0.621)	(0.255)	(0.424)	(0.332)		(0.557)	(0.709)	(0.826)	(0.517)	(0.651)	(0.604)
Age at Entry		0.00507	0.0185	-0.180***	0.0190	-0.0528**	-0.00591		0.00208	0.00751	-0.178***	0.0146	-0.0497**	-0.0111		-0.0439	-0.0410	-0.262***	-0.0450	-0.0963**	-0.0966**
		(0.0139)	(0.0257)	(0.0491)	(0.0129)	(0.0238)	(0.0164)		(0.0149)	(0.0252)	(0.0458)	(0.0146)	(0.0225)	(0.0179)		(0.0374)	(0.0408)	(0.0632)	(0.0353)	(0.0408)	(0.0381)
University Undergraduate		0.133	0.715	-0.660	0.512	1.428	-0.0482		0.157	1.036	-0.576	0.531	1.539	-0.0542		-0.499	0.458	-0.540	-0.112	0.844	-0.672
		(0.317)	(0.633)	(0.905)	(0.354)	(1.046)	(0.475)		(0.329)	(0.744)	(0.865)	(0.366)	(1.006)	(0.487)		(0.783)	(1.011)	(1.245)	(0.799)	(1.196)	(0.855)
University Postgraduate		0.0118	0.636	-0.412	0.990***	1.587	0.585		-0.0838	0.960	-0.581	0.839**	1.635	0.464		-1.101	0.126	-1.082	-0.321	0.622	-0.767
		(0.352)	(0.671)	(0.968)	(0.374)	(1.066)	(0.481)		(0.370)	(0.814)	(0.969)	(0.391)	(1.034)	(0.496)		(0.839)	(1.077)	(1.292)	(0.831)	(1.261)	(0.892)
Party Conservative									-0.610	5.155***	-0.887	-1.642***	1.989***	-1.520*		-1.987**	4.261***	-1.867	-2.314***	0.799	-2.135**
									(0.451)	(1.050)	(1.093)	(0.564)	(0.556)	(0.801)		(0.800)	(1.144)	(1.174)	(0.736)	(0.832)	(0.915)
Party LibDem									-1.443***	0.804	-2.507***	-2.110***	-0.650	-1.779***		-1.281**	1.654	-0.981	-0.764	-0.289	0.0852
									(0.254)	(1.127)	(0.792)	(0.267)	(0.521)	(0.345)		(0.565)	(1.173)	(1.071)	(0.504)	(0.717)	(0.583)
By-elections									0.0191	0.347	-0.430	-0.317	0.999	-14.97***		-0.0975	-0.0123	-0.987	-0.225	0.727	-17.81***
									(0.585)	(1.318)	(1.047)	(0.584)	(0.728)	(0.473)		(1.245)	(1.727)	(1.782)	(1.197)	(1.482)	(1.189)
Legislature 2001-2005																-0.590	-0.457	-2.271*	-0.261	-0.154	-19.71***
																(0.579)	(0.636)	(1.166)	(0.511)	(0.696)	(0.535)
Legislature 2005-2010																3.669***	1.659	-17.04***	-16.97***	2.765**	-17.87***
																(1.091)	(1.282)	(1.228)	(1.060)	(1.221)	(1.081)
Legislature 2010-2015																-21.07***	-19.85***	-22.42***	-21.91***	-4.907***	-23.00***
																(0.482)	(0.605)	(0.803)	(0.461)	(1.118)	(0.603)
Constant		-0.685	-3.192**	4.097*	-2.035***	-0.813	-1.187		0.576	-5.363***	5.740***	-0.224	-0.937	0.534		4.879**	-0.594	11.77***	4.768***	3.618	6.999***
		(0.740)	(1.268)	(2.171)	(0.702)	(1.434)	(0.904)		(0.805)	(1.750)	(2.133)	(0.812)	(1.509)	(0.980)		(1.979)	(2.379)	(2.815)	(1.842)	(2.332)	(1.999)
Observations	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695	695

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A7: Multinomial Analysis with Oxbridge - Eight Clusters

	Career 1	Career 2	Career 3	Career 4	Career 5	Career 6	Career 7	Career 8	Career 1	Career 2	Career 3	Career 4	Career 5	Career 6	Career 7	Career 8	Career 1	Career 2	Career 3	Career 4	Career 5	Career 6	Career 7	Career 8
Gender		-0.363	-0.321	-0.217	-0.0927	-0.451*	0.00128	-0.341		-0.566*	0.322	-0.721	-0.482	-0.951***	0.126	-0.868**		-0.506	0.826	-0.461	-0.154	-0.179	0.512	-0.489
		(0.312)	(0.456)	(0.688)	(0.328)	(0.253)	(0.412)	(0.380)		(0.322)	(0.513)	(0.693)	(0.355)	(0.276)	(0.486)	(0.400)		(0.743)	(0.739)	(0.936)	(0.603)	(0.569)	(0.678)	(0.676)
State School		-0.0115	0.0250	1.674**	0.0355	0.264	-0.141	-0.450		-0.206	-0.227	1.054	-0.624*	-0.330	-0.430	-0.935**		-1.015*	-0.991	0.586	-1.353**	-0.870*	-0.996*	-1.780***
		(0.304)	(0.414)	(0.819)	(0.330)	(0.246)	(0.387)	(0.354)		(0.328)	(0.525)	(0.827)	(0.370)	(0.270)	(0.424)	(0.403)		(0.615)	(0.648)	(0.944)	(0.544)	(0.487)	(0.589)	(0.632)
Job		-0.415	0.157	-0.615	0.332	0.00896	-0.0774	-0.170		-0.510	0.169	-0.639	0.177	-0.150	-0.134	-0.344		-0.0491	0.643	0.271	0.832	0.583	0.189	0.425
		(0.321)	(0.386)	(0.686)	(0.304)	(0.244)	(0.399)	(0.364)		(0.328)	(0.525)	(0.679)	(0.317)	(0.266)	(0.429)	(0.369)		(0.667)	(0.724)	(0.875)	(0.571)	(0.530)	(0.642)	(0.651)
Age at Entry		-0.0182	0.0140	-0.187***	0.0421**	0.0184	-0.0490*	-0.00968		-0.0214	0.00521	-0.175***	0.0433**	0.0172	-0.0452*	-0.0140		-0.0456	-0.0559	-0.301***	-0.0345	-0.0556	-0.104**	-0.119***
		(0.0167)	(0.0268)	(0.0594)	(0.0203)	(0.0142)	(0.0258)	(0.0185)		(0.0180)	(0.0278)	(0.0536)	(0.0213)	(0.0160)	(0.0241)	(0.0202)		(0.0470)	(0.0420)	(0.0696)	(0.0394)	(0.0365)	(0.0435)	(0.0422)
Oxbridge		-0.441	0.399	1.291**	0.212	-0.346	0.704*	-1.430***		-0.379	0.774	1.475**	0.391	-0.203	0.896**	-1.350***		-1.551***	-0.344	0.448	-1.045*	-1.703***	-0.101	-3.331***
		(0.351)	(0.409)	(0.612)	(0.341)	(0.274)	(0.386)	(0.492)		(0.356)	(0.543)	(0.599)	(0.358)	(0.300)	(0.423)	(0.511)		(0.576)	(0.621)	(0.809)	(0.568)	(0.516)	(0.564)	(0.742)
Party Conservative										0.292	5.252***	-0.465	-1.156	-1.699***	2.055***	-2.000*		-1.274	4.349***	-1.618	-2.223**	-2.383***	0.903	-2.832**
										(0.528)	(1.049)	(1.179)	(0.731)	(0.630)	(0.594)	(1.059)		(0.945)	(1.160)	(1.337)	(0.930)	(0.838)	(0.870)	(1.165)
Party LibDem										-0.827**	0.830	-2.159***	-2.142***	-2.055***	-0.649	-1.774***		-0.880	1.824	-0.467	-1.208*	-0.433	-0.140	0.547
										(0.328)	(1.129)	(0.751)	(0.396)	(0.284)	(0.545)	(0.395)		(0.721)	(1.182)	(0.989)	(0.645)	(0.560)	(0.758)	(0.673)
By-elections										-14.28***	0.262	-0.104	0.795	-0.434	0.982	-14.70***		-18.51***	-0.388	-0.171	0.564	-0.150	0.437	-17.83***
										(0.446)	(1.261)	(1.096)	(0.698)	(0.661)	(0.732)	(0.532)		(1.508)	(1.950)	(1.734)	(1.431)	(1.417)	(1.752)	(1.400)
Legislature 2001-2005																		18.06***	-0.201	-2.221**	-2.282**	-0.170	0.216	-20.34***
																		(0.543)	(0.708)	(1.063)	(0.894)	(0.567)	(0.721)	(0.672)
Legislature 2005-2010																		21.88***	1.503	-18.42***	2.183**	-18.60***	2.786**	-19.54***
																		(0.966)	(1.183)	(1.161)	(1.008)	(0.985)	(1.103)	(1.072)
Legislature 2010-2015																		-3.719***	-20.53***	-22.49***	-22.16***	-22.92***	-4.716***	-24.40***
																		(0.519)	(0.661)	(0.834)	(0.632)	(0.613)	(1.139)	(0.863)
Constant		0.0815	-2.524**	2.718	-3.111***	-1.098	0.0753	-0.270		0.877	-4.686***	3.706*	-1.754*	0.462	-0.0862	1.355		-13.60***	0.487	11.48***	4.271**	5.932***	4.558**	8.835***
		(0.814)	(1.215)	(2.207)	(0.966)	(0.671)	(1.184)	(0.838)		(0.908)	(1.668)	(2.122)	(1.001)	(0.800)	(1.185)	(0.959)		(2.297)	(2.311)	(2.791)	(1.982)	(1.831)	(2.176)	(2.134)
Observations	613	613	613	613	613	613	613	613	613	613	613	613	613	613	613	613	613	613	613	613	613	613	613	613

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A8: Discrepancy Analysis

Variable	PseudoF	PseudoR2	p_value
Gender	1.5406378	0.0013683279	0.165
State School	1.0299627	0.0009147683	0.385
Job	0.8401325	0.0007461693	0.535
Age at Entry	3.1711299	0.0028164604	0.005
By-elections	6.0032943	0.0053318663	0.005
Legistature 2001-2005	10.2285072	0.0090845176	0.005
Legistature 2005-2010	60.3632637	0.0536120395	0.005
Legistature 2010-2015	46.5763286	0.2189984284	0.005
University Undergraduate	0.4084218	0.0003627426	0.930
University Postgraduate	0.6823863	0.0006060660	0.655
Party Conservative	69.4499886	0.0616824755	0.005
Party LibDem	3.2946912	0.0029262022	0.005
Total	36.9939491	0.3942771028	0.005

References

Mirkin, B. 2013. *Mathematical Classification and Clustering.* London: Springer.

Needleman, S., and C. Wunsch. 1970. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." *Journal of Molecular Biology* 48 (3): 443–453.

Studer, M., A. Gabadinho, N. S. Muller, and R. G. 2011. *An Introduction to Sequence Analysis Using the TraMineR R Package.* Workshop on Trajectories Paris Institute for Demographic and Life Course Studies, University of Geneva, October.