

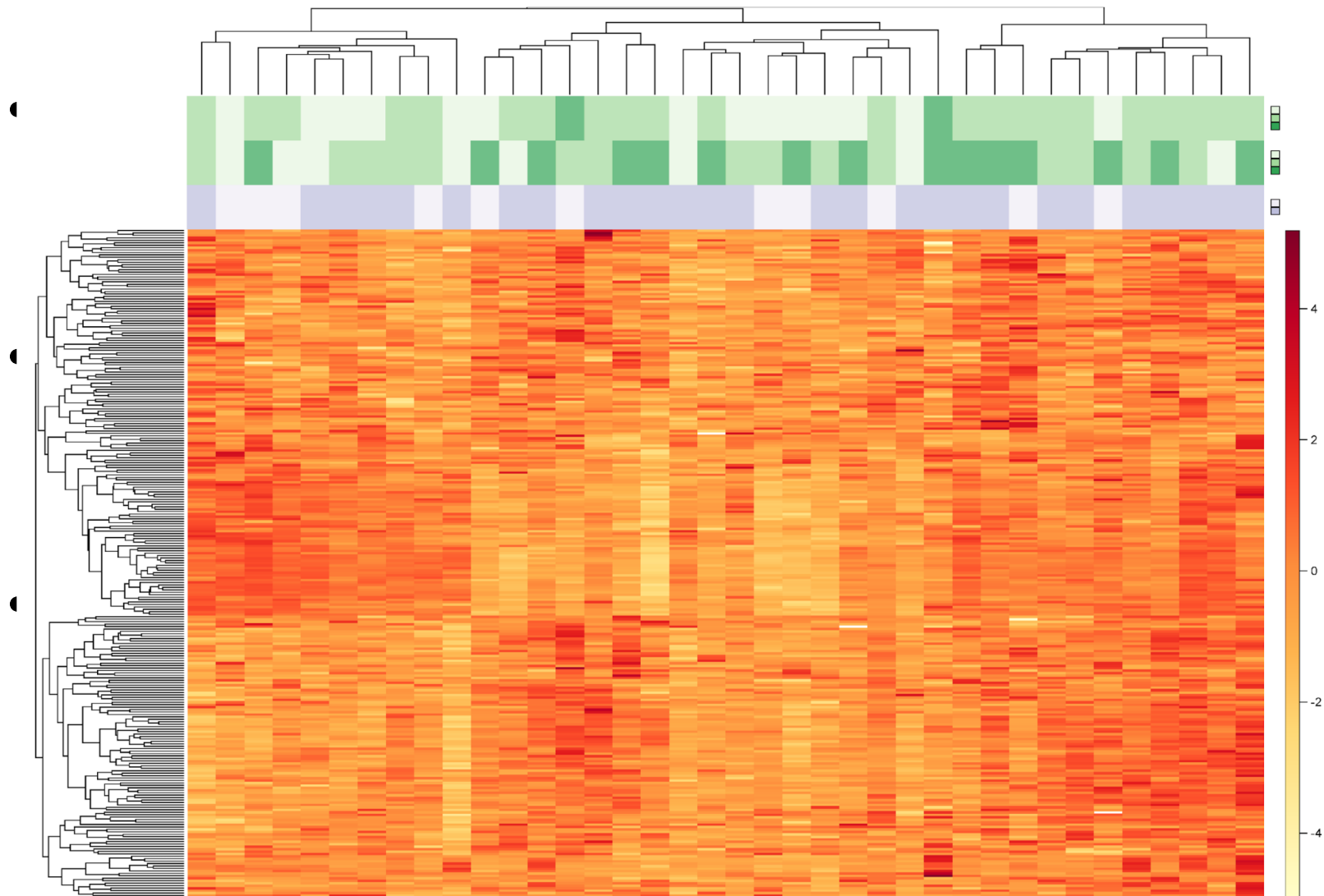
Aligning techniques between R and SAS

A practical example using cluster analysis

Mike Cooper, Steve Prust
Elderbrook Solutions GmbH

- Introduction
- Cluster analysis
- Moving data between SAS and R
- Replicating results
 - Distance matrix
 - Clustering (dendrogram)
 - Data structure alignment
 - Problem resolution
- Results
- Other topics
- Conclusion

- Initial discussions and understanding of the problem
 - Analysis complexity
 - New concepts
 - Paucity of previous work to draw upon
- What were the options?
 - R only
 - SAS only
 - Use both R and SAS
- Why decide on the alignment approach we are presenting?
 - Stronger understanding of all aspects of the analysis
 - Improved efficacy in discussions during development
 - Effective validation option



```
data protein;
  length country $ 14;
  input country &$ redmeat whitemeat eggs milk fish cereal starch nuts fruitveg;
  datalines;
Albania      10.1  1.4  0.5   8.9  0.2  42.3  0.6  5.5  1.7
Austria      8.9 14.0  4.3  19.9  2.1  28.0  3.6  1.3  4.3
Belgium     13.5  9.3  4.1  17.5  4.5  26.6  5.7  2.1  4.0
Bulgaria      7.8  6.0  1.6   8.3  1.2  56.7  1.1  3.7  4.2
Czechoslovakia 9.7 11.4  2.8  12.5  2.0  34.3  5.0  1.1  4.0
Denmark     10.6 10.8  3.7  25.0  9.9  21.9  4.8  0.7  2.4
E Germany    8.4 11.6  3.7  11.1  5.4  24.6  6.5  0.8  3.6
Finland      9.5  4.9  2.7  33.7  5.8  26.3  5.1  1.0  1.4
France     18.0  9.9  3.3  19.5  5.7  28.1  4.8  2.4  6.5
Greece     10.2  3.0  2.8  17.6  5.9  41.7  2.2  7.8  6.5
Hungary      5.3 12.4  2.9   9.7  0.3  40.1  4.0  5.4  4.2
Ireland     13.9 10.0  4.7  25.8  2.2  24.0  6.2  1.6  2.9
Italy       9.0  5.1  2.9  13.7  3.4  36.8  2.1  4.3  6.7
Netherlands  9.5 13.6  3.6  23.4  2.5  22.4  4.2  1.8  3.7
Norway      9.4  4.7  2.7  23.3  9.7  23.0  4.6  1.6  2.7
Poland      6.9 10.2  2.7  19.3  3.0  36.1  5.9  2.0  6.6
Portugal     6.2  3.7  1.1   4.9 14.2  27.0  5.9  4.7  7.9
Romania      6.2  6.3  1.5  11.1  1.0  49.6  3.1  5.3  2.8
Spain       7.1  3.4  3.1   8.6  7.0  29.2  5.7  5.9  7.2
Sweden      9.9  7.8  3.5   4.7  7.5  19.5  3.7  1.4  2.0
Switzerland 13.1 10.1  3.1  23.8  2.3  25.6  2.8  2.4  4.9
Uk          17.4  5.7  4.7  20.6  4.3  24.3  4.7  3.4  3.3
Ussr        9.3  4.6  2.1  16.6  3.0  43.6  6.4  3.4  2.9
W Germany   11.4 12.5  4.1  18.8  3.4  18.6  5.2  1.5  3.8
Yugoslavia   4.4  5.0  1.2   9.5  0.6  55.9  3.0  5.7  3.2
;
run;
```

(presentation data from
https://documentation.sas.com/doc/en/statcdc/14.2/statug/statug_code_distangs.htm)

- . Using SAS environment
- . PROC IML to access R
- . SAS datasets (and other information such as macro variables) can be transferred to R environment
- . R environment can send information back to SAS

```
options set=R_HOME='<location>';
ods listing close;

proc iml;

  run ExportDataSetToR("mysasdata", "mysasdata_in_r");
  submit / R;

    library(tibble)
    library(ama)

    mysasdata_in_r_adj = mysasdata_in_r %>%
      column_to_rownames('COUNTRY')

    d.country = Dist(mysasdata_in_r_adj, method = 'euclidean')
    dcountry_matrix <- as.matrix(d.country)

  endsubmit;

  run ImportDataSetFromR("dist_from_r", "dcountry_matrix");
quit;
ods listing ;
```

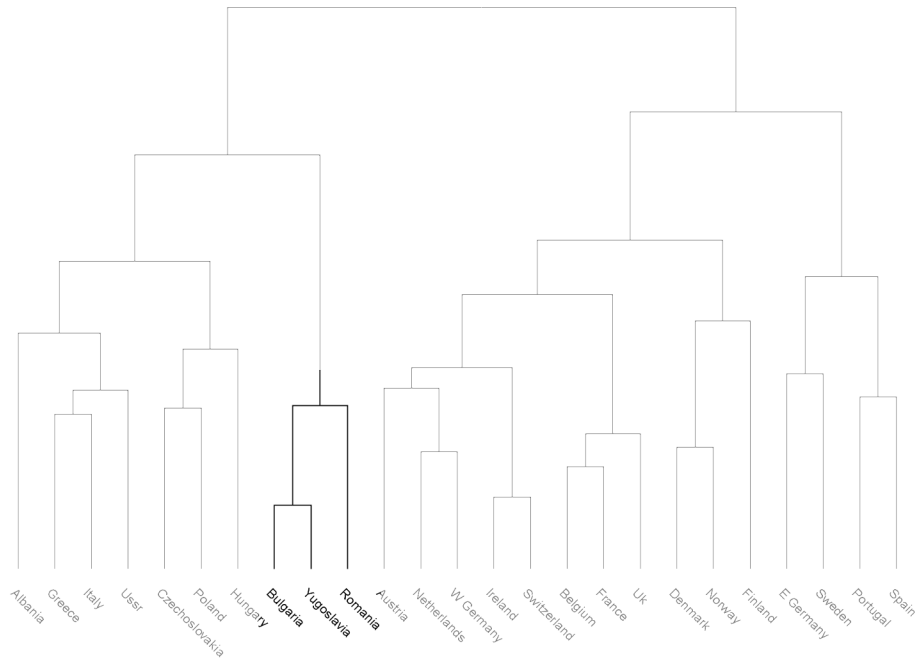
- SAS

- R

```
proc transpose data=protein out=protein_t; var redmeat--fruitveg; id country; run;

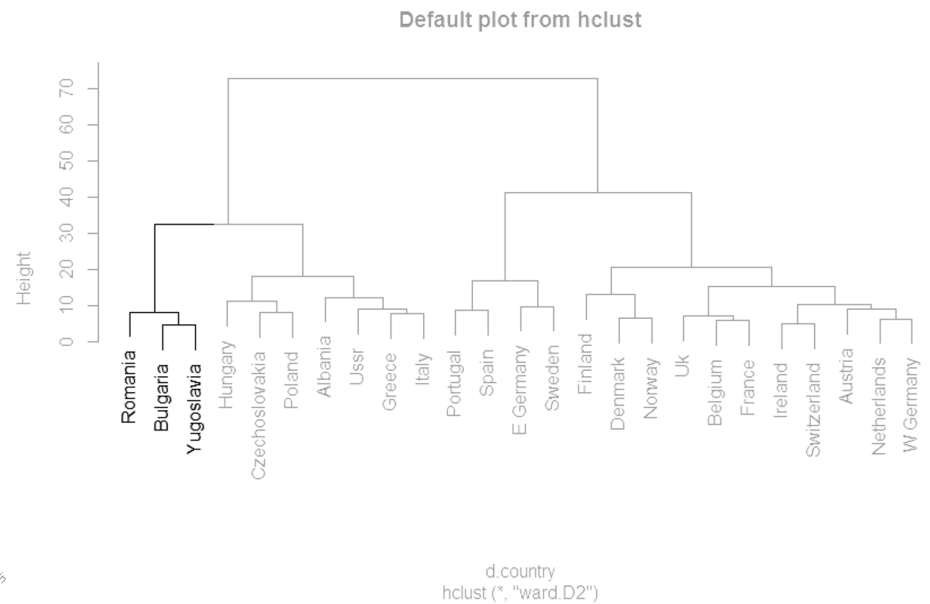
data dist_prog;
  length id $20;
  set protein_t end=last ;
  array res [25,25] _res1-_res625;;
  array val[25] albania austria belgium bulgaria czechoslovakia denmark e_germany finland france greece hungary ireland italy
               netherlands norway poland portugal romania spain sweden switzerland uk ussr w_germany yugoslavia ;
  if _n_ = 1 then do i = 1 to 25; do j=1 to 25; res[i,j] = 0; end; end;
  /*process each row */
  do i = 1 to 25; do j = 1 to 25; if i ne j then do; res[i,j] = res[i,j] + ((val[i]-val[j])**2); end; end; end;
  retain _res;;
  if last then do;
    do i = 1 to 25;
      id = propcase(tranwrd(vname(val[i]),'_',' '));
      do j = 1 to 25;
        if j < i then val[j] = sqrt ( res[i,j] );
        else if i = j then val[j] = 0;
        else val[j] = .;
      end;
      output;
    end;
  end;
  keep id albania austria belgium bulgaria czechoslovakia denmark e_germany finland france greece hungary ireland italy
       netherlands norway poland portugal romania spain sweden switzerland uk ussr w_germany yugoslavia ;
run;
```

- SAS



Dendrogram example - SAS

- R



Replicating results – 3 – data structures

VIEWTABLE: Work.Tree_sas (Ward's Minimum Variance Cluster Analysis)											
	NAME	_PARENT_	_NCL_	_FREQ_	_HEIGHT_	_RMSSTD_	_SPRSQ_	_RSQ_	_PSF_	_PST2_	COUNTRY
1	Bulgaria	CL24	25	1	0	0	0	1	.	.	Bulgaria
2	Yugoslavia	CL24	25	1	0	0	0	1	.	.	Yugoslavia
3	Ireland	CL23	25	1	0	0	0	1	.	.	Ireland
4	Switzerland	CL23	25	1	0	0	0	1	.	.	Switzerland
5	Belgium	CL22	25	1	0	0	0	1	.	.	Belgium
6	France	CL22	25	1	0	0	0	1	.	.	France
7	Netherlands	CL21	25	1	0	0	0	1	.	.	Netherlands
8	W Germany	CL21	25	1	0	0	0	1	.	.	W Germany
9	Denmark	CL20	25	1	0	0	0	1	.	.	Denmark
10	Norway										Norway
VIEWTABLE: Work.Clust_r_height			VIEWTABLE: Work.Clust_r_merge			VIEWTABLE: Work.Clust_r_order		VIEWTABLE: Work.Clust_r_labels			
	V1		V1	V2		V1		V1			
11	CL22										
12	Uk	1	4.875448697	1	-4	-25	1	18	1	Albania	Uk
13	Greece	2	5.097057975	2	-12	-21	2	4	2	Austria	Greece
14	Italy	3	6.007495318	3	-3	-9	3	25	3	Belgium	Italy
15	Czechoslovakia	4	6.52533524	4	-14	-24	4	11	4	Bulgaria	Czechoslovakia
16	Poland	5	6.688049043	5	-6	-15	5	5	5	Czechoslovakia	Poland
17	CL24	6	7.184474001	6	-22	3	6	16	6	Denmark	
18	Romania	7	7.97684148	7	-10	-13	7	1	7	E Germany	Romania
19	Portugal	8	8.255301327	8	-5	-16	8	23	8	Finland	Portugal
20	Spain	9	8.380334122	9	-18	1	9	10	9	France	Spain
21	CL18	10	8.788629017	10	-17	-19	10	13	10	Greece	
22	Ussr	11	9.112445702	11	-23	7	11	17	11	Hungary	Ussr
23	Austria	12	9.205433178	12	-2	4	12	19	12	Ireland	Austria
		13	9.943339479	13	-7	-20	13	7	13	Italy	
		14	10.29970873	14	2	12	14	20	14	Netherlands	
		15	11.37292106	15	-11	8	15	8	15	Norway	
		16	12.42148676	16	-1	11	16	6	16	Poland	
		17	13.27416036	17	-8	5	17	15	17	Portugal	
		18	15.28125758	18	6	14	18	22	18	Romania	
		19	16.86193939	19	10	13	19	3	19	Spain	
		20	18.31746499	20	15	16	20	9	20	Sweden	
		21	20.57229124	21	17	18	21	12	21	Switzerland	
		22	32.61124258	22	9	20	22	21	22	Uk	
		23	41.19060977	23	19	21	23	2	23	Ussr	
		24	72.62540786	24	22	23	24	14	24	W Germany	
							25	24	25	Yugoslavia	

Manipulating the R data to match SAS ...

- Merge the n-row tables
- Combine clade info
- Merge clade and label
- Similar structure to SAS
- ... height information ...
 - ... does not match

match

VIEWTABLE: Work.Clust_r_labelsorder					
	COUNTRY	SEQ	ORDER		
1	Albania	1	18		
2	Austria	2	4		
3	Belgium	3	25		
VIEWTABLE: Work.Clust_r_pcheight					
	INDEX	NODE	SEQ	HEIGHT	
1	1	.	4	4.875	
2	1	.	25	4.875	
3	2	.	12	5.097	
4	2	.	21	5.097	
VIEWTABLE: Work.Tree_r_step2					
	INDEX	NODE	SEQ	HEIGHT	COUNTRY
1	1	.	4	4.875	Bulgaria
2	1	.	25	4.875	Yugoslavia
3	2	.	12	5.097	Ireland
4	2	.	21	5.097	Switzerland
5	3	.	3	6.007	Belgium
6	3	.	9	6.007	France
7	4	.	14	6.525	Netherlands
8	4	.	24	6.525	W Germany
9	5	.	6	6.688	Denmark
10	5	.	15	6.688	Norway
11	6	3	.	7.184	.
12	6	.	22	7.184	Uk
13	7	.	10	7.977	Greece
14	7	.	13	7.977	Italy
15	8	.	5	8.255	Czechoslovakia
16	8	.	16	8.255	Poland
17	9	1	.	8.38	.
18	9	.	18	8.38	Romania
19	10	.	17	8.789	Portugal
20	10	.	19	8.789	Spain
21	11	7	.	9.112	.
22	11	.	23	9.112	Ussr
23	12	4	.	9.205	.
24	12	.	2	9.205	Austria
25	13	.	7	9.943	E Germany
26	13	.	20	9.943	Sweden

- Non-matching height - two issues
 - structure (parent / child)

VIEWTABLE: Work.Tree_sas (Ward's Minimum Variance Cluster Analysis)

	NAME	_PARENT_	_NCL_	_HEIGHT_	COUNTRY
1	Bulgaria	CL24	25	0	Bulgaria
2	Yugoslavia	CL24	25	0	Yugoslavia
3	Ireland	CL23	25	0	Ireland
4	Switzerland	CL23	25	0	Switzerland

VIEWTABLE: Work.Tree_sas (Ward's Minimum Variance Cluster Analysis)

	NAME	_PARENT_	_NCL_	_HEIGHT_	COUNTRY
1	Bulgaria	CL24	25	0	Bulgaria
2	Yugoslavia	CL24	25	0	Yugoslavia
3	Ireland	CL23	25	0	Ireland
4	Switzerland	CL23	25	0	Switzerland
5	Belgium	CL22	25	0	Belgium
6	France	CL22	25	0	France
7	Netherlands	CL21	25	0	Netherlands
8	W Germany	CL21	25	0	W Germany
9	Denmark	CL20	25	0	Denmark
10	Norway	CL20	25	0	Norway
11	CL22	CL19	22	0.0033894366	
12	UK	CL19	25	0	UK
13	Greece	CL18	25	0	Greece
14	Italy	CL18	25	0	Italy
15	Czechoslovakia	CL17	25	0	Czechoslovakia
16	Poland	CL17	25	0	Poland
17	CL24	CL16	24	0.0022323887	

VIEWTABLE: Work.Tree_test5

	NAME	_PARENT_	_NCL_	COUNTRY	HEIGHT	PHEIGHT
1	Bulgaria	CL24	25	Bulgaria	4.8754486973	.
2	Yugoslavia	CL24	25	Yugoslavia	4.8754486973	.
3	Ireland	CL23	25	Ireland	5.0970579749	.
4	Switzerland	CL23	25	Switzerland	5.0970579749	.

VIEWTABLE: Work.Tree_test5

	NAME	_PARENT_	_NCL_	COUNTRY	PHEIGHT	_HEIGHT_
1	Bulgaria	CL24	25	Bulgaria	.	0
2	Yugoslavia	CL24	25	Yugoslavia	.	0
3	Ireland	CL23	25	Ireland	.	0
4	Switzerland	CL23	25	Switzerland	.	0
5	Belgium	CL22	25	Belgium	.	0
6	France	CL22	25	France	.	0
7	Netherlands	CL21	25	Netherlands	.	0
8	W Germany	CL21	25	W Germany	.	0
9	Denmark	CL20	25	Denmark	.	0
10	Norway	CL20	25	Norway	.	0
11	CL22	CL19	22		6.0074953184	0.0033894366
12	UK	CL19	25	UK	.	0
13	Greece	CL18	25	Greece	.	0
14	Italy	CL18	25	Italy	.	0
15	Czechoslovakia	CL17	25	Czechoslovakia	.	0
16	Poland	CL17	25	Poland	.	0
17	CL24	CL16	24		4.8754486973	0.0022323887

- Main challenges when trying to align
 - Online resources
 - Analysis complexity
 - How to resolve validation differences?
- Was our rationale for choosing this approach justified?
- Is this approach always applicable/relevant?
 - Context is important
 - No “one size fits all” recommendation
- Possible ways to improve future alignment
 - Improved R documentation?
 - Development of an online SAS<->R equivalency guide?

Your comments and questions are valued and encouraged.

michael.cooper@elderbrooksolutions.com

stephen.prust@elderbrooksolutions.com