A clustering view on ESS measures of political interest: An EM-MML approach

Cláudia Silvestre
Escola Superior de Comunicação Social - IPL

Margarida Cardoso
BRU_UNIDE, ISCTE-IUL
Portugal

Mário Figueiredo
Instituto de Telecomunicações, Inst. Sup. Técnico

Outline

- Objective
- Model
  - Finite Mixture Models
- Selection Criterion
  - Minimum Message Length
- Algorithm
  - EM-MML
- Results
- Conclusions

Objective

- Clustering the **regions** in the *European Social Survey* based on attitudes towards politics

  - Voted last national election (Yes; No; Not eligible)
  - Contacted politician or government official in last 12 months
  - Worked in political party or action group in last 12 months
  - Worked in another organisation or association in last 12 months
  - Worn or displayed campaign badge/sticker in last 12 months
  - Signed petition in last 12 months
  - Taken part in lawful public demonstration in last 12 months
  - Boycotted certain products last in 12 months
  - Feel closer to a particular party than all other parties

  (Y/N)

Model: Finite Mixture Models

$$f\left(\underline{y}_i|\underline{\theta}\right) = \sum_{k=1}^{K} \alpha_k f\left(\underline{y}_i|\underline{\theta}_k\right)$$

- K is the number of segments
- $\underline{y}_i$ is regarded as "incomplete data", the allocation to segments ($\underline{z}_i$) being missing

Complete data: $\left(\underline{y}_i, \underline{z}_i\right)$

Model: Finite Mixture Models

- The log of complete likelihood

$$log f\left(\underline{y}_i, \underline{z}_i | \underline{\theta}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} log\left[\alpha_k f\left(\underline{y}_i | \underline{\theta}_k\right)\right]$$

uknown

Selection Criterion

How to select the number of segments?

- Information criteria such as BIC, AIC, CAIC, AIC_3 or ICL can be used…
- We adopt Minimum Message Length criterion embedded in the model estimation (Figueiredo and Jain, 2002), which:
  - Provides estimates of all the model parameters including the number of segments
  - Is less sensitive to initialization than EM
  - Avoids the boundary of the parameters space

Selection Criterion: MML

- Shannon's Information Theory: optimally transmitting a random variable Y with probability $f\left(\underline{y}\right)$ requires about $-log_2\left[f\left(\underline{y}\right)\right]$ bits of information.

- to encode $\underline{y}$: $l\left(\underline{y}|\underline{\theta}\right) = -log_2\left[f\left(\underline{y},\underline{\theta}\right)\right]$

- to encode $\underline{y}$ and $\underline{\theta}$ the total message length is:

$$l\left(\underline{y},\underline{\theta}\right) = l\left(\underline{y}|\underline{\theta}\right) + l\left(\underline{\theta}\right)$$

Algorithm: EM-MML

- EM is a popular algorithm for finding ML parameter estimates, when unobserved (missing) data is considered in the model.

The EM-MML

- A mixture of multinomials is adopted and the MML estimates are obtained via an EM-type algorithm.

Algorithm: EM-MML

Categorical variables:

$$\underline{Y} = \{\underline{Y_1}, \dots, \underline{Y_i}, \dots \underline{Y_n}\}$$

$$\underline{Y_i} = (\underline{Y_{i1}}, \dots, \underline{Y_{iD}})$$

where variable d ($d = 1 \dots D$) has C_d categories

$$\underline{\theta} = \{\underline{\theta_1}, \dots, \underline{\theta_K}, \alpha_1, \dots, \alpha_K\},$$

α_k are the clusters' weights or mixing probabilities

$\underline{\theta_k}$ the multinomials' parameters

Algorithm: EM-MML

$$\log\ f\left(\underline{y}|\underline{\theta}\right) = \sum_{i=1}^{n} \log\ f\left(\underline{y_i}|\underline{\theta}\right)$$

Mixture of multinomials:

$$f\left(\underline{y_i}|\underline{\theta}\right) = \sum_{k=1}^{K} \alpha_k \prod_{d=1}^{D} \left[n! \prod_{c=1}^{C_d} \frac{\left(\hat{\theta}_{kdc}\right)^{y_{idc}}}{y_{idc}!} \right]$$

Algorithm: EM-MML

Assuming that:

- The segments have independent priors
- …independent from the mixing probabilities
- A noninformative Jeffreys prior for $\underline{\theta}$

$$l\left(\underline{y},\underline{\theta}\right) = l\left(\underline{y}|\underline{\theta}\right) + l(\underline{\theta})$$

$$= \frac{M}{2}\sum_{k,\alpha_k>0} log\left(\frac{n\,\alpha_k}{12}\right) + k_{nz} log\left(\frac{n}{12}\right) + \frac{k_{nz}(M+1)}{2} - \log\ f\left(\underline{y}|\underline{\theta}\right)$$

M is the number of parameters specifying each segment

k_{nz} is the number of segments with non-zero probability

Algorithm: EM-MML

E-step

$$E\left[Z_{ik}|\underline{y}_i;\underline{\hat{\theta}}^{(t)}\right] = P\left[Z_{ik}=1|\underline{y}_i;\underline{\hat{\theta}}^{(t)}\right]$$

$$= \frac{{\alpha_k}^{(t)} f\left(\underline{y}_i;\underline{\hat{\theta}}_k^{(t)}\right)}{\sum_{k=1}^{K} {\alpha_k}^{(t)} f\left(\underline{y}_i;\underline{\hat{\theta}}_k^{(t)}\right)}$$

where

$$f\left(\underline{y}_i;\underline{\hat{\theta}}_k^{(t)}\right) = \prod_{d=1}^{D}\left[n!\prod_{c=1}^{C_d}\frac{\left(\hat{\theta}_{kdc}^{(t)}\right)^{y_{idc}}}{y_{idc}!}\right]$$

Algorithm: EM-MML

M-step

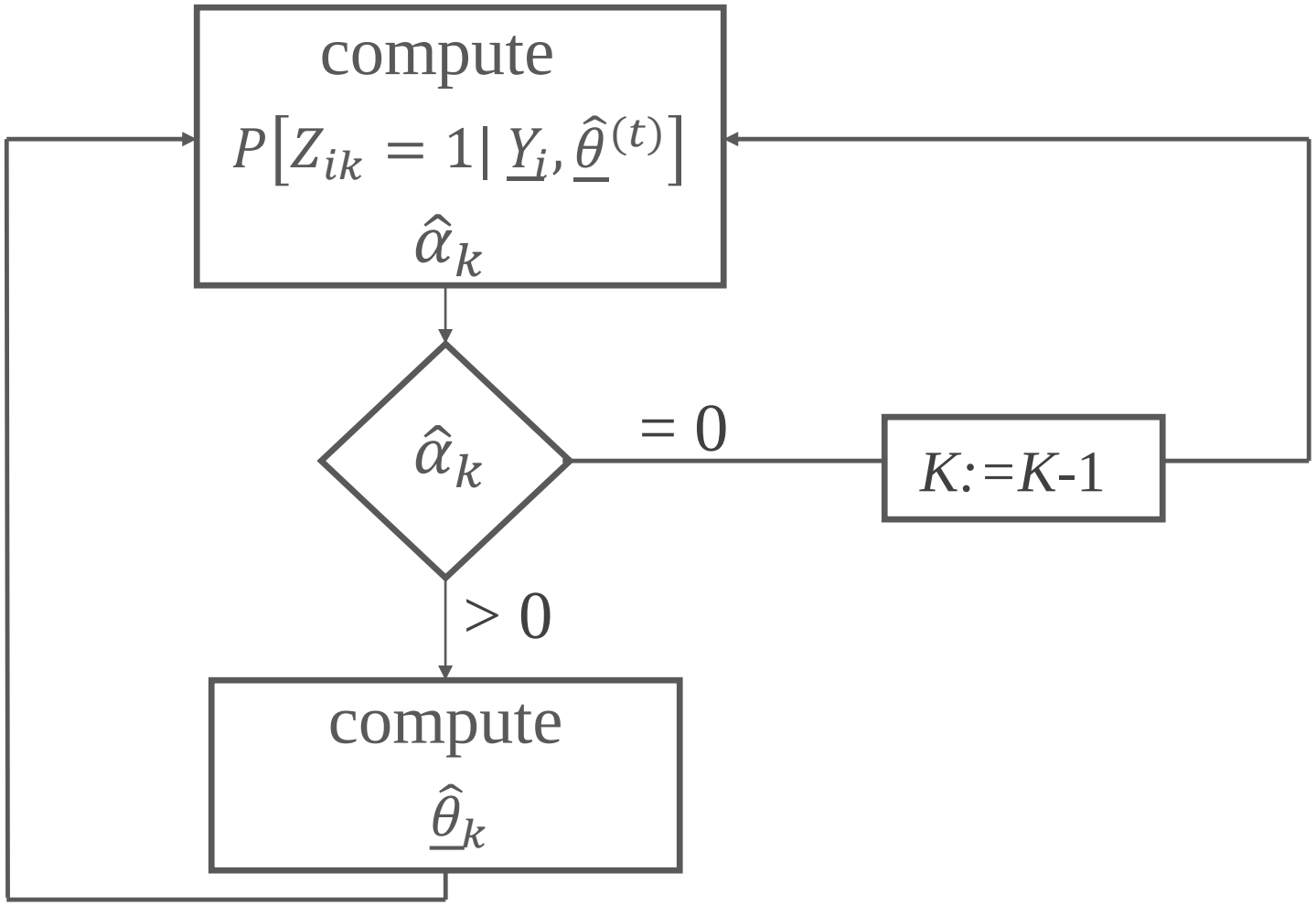
- Update the estimates of mixing probabilities

$$\hat{\alpha}_k^{(t+1)} = \frac{max\left\{0, \sum_{i=1}^{n} P\left[Z_{ik} = 1 | \underline{y}_i; \underline{\hat{\theta}}^{(t)}\right] - \frac{M}{2}\right\}}{\sum_{k=1}^{K} max\left\{0, \sum_{i=1}^{n} P\left[Z_{ik} = 1 | \underline{y}_i; \underline{\hat{\theta}}^{(t)}\right] - \frac{M}{2}\right\}}$$

- Update the estimates of multinomial parameters

$$\hat{\theta}_{kdc}^{(t+1)} = \frac{\sum_{i=1}^{n} P\left[Z_{ik} = 1 | \underline{y}_i; \underline{\hat{\theta}}^{(t)}\right] y_{idc}}{n! \sum_{i=1}^{n} P\left[Z_{ik} = 1 | \underline{y}_i; \underline{\hat{\theta}}^{(t)}\right]}$$

Algorithm: EM-MML



Results

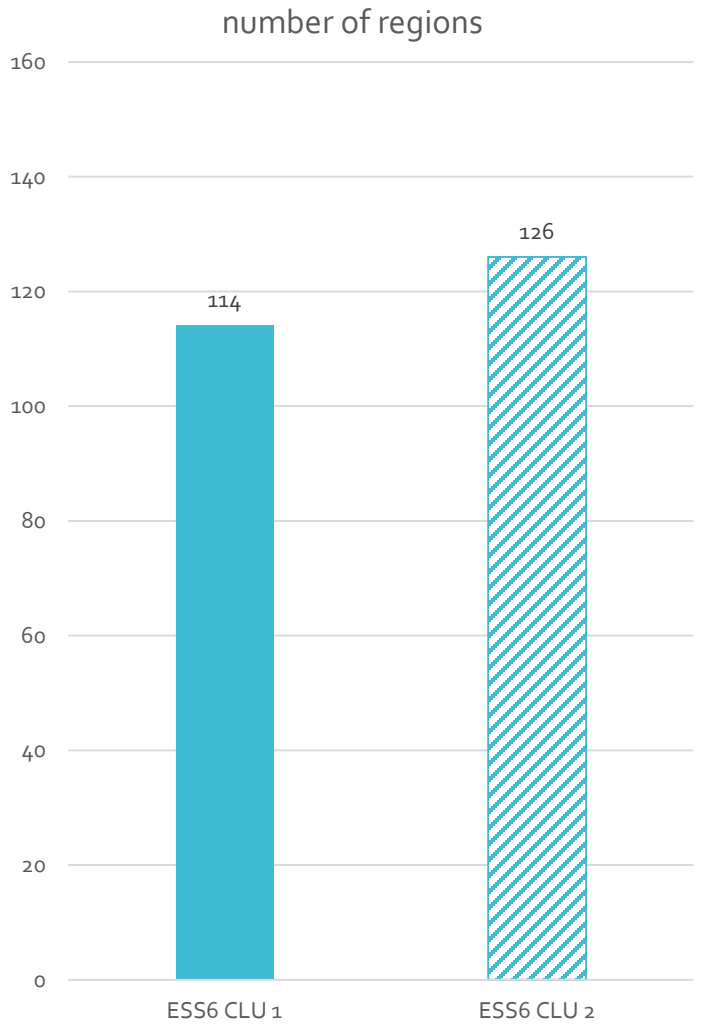
- The clustering of Regions in the European Social Survey based on attitudes towards politics, using EM-MML, yields 2 clusters

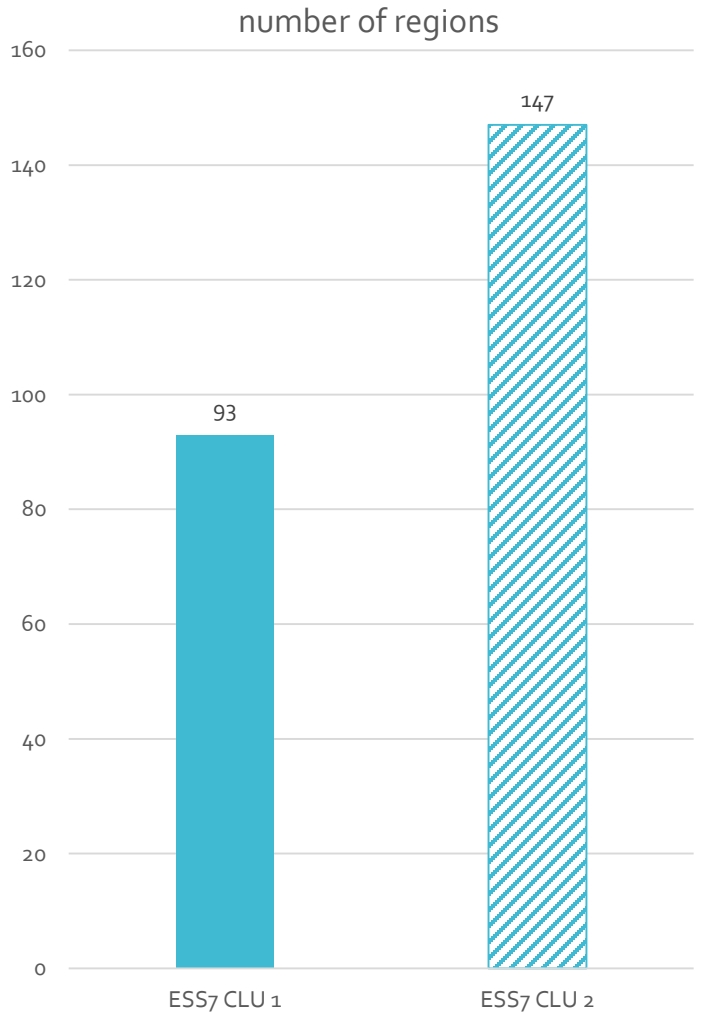
Results:

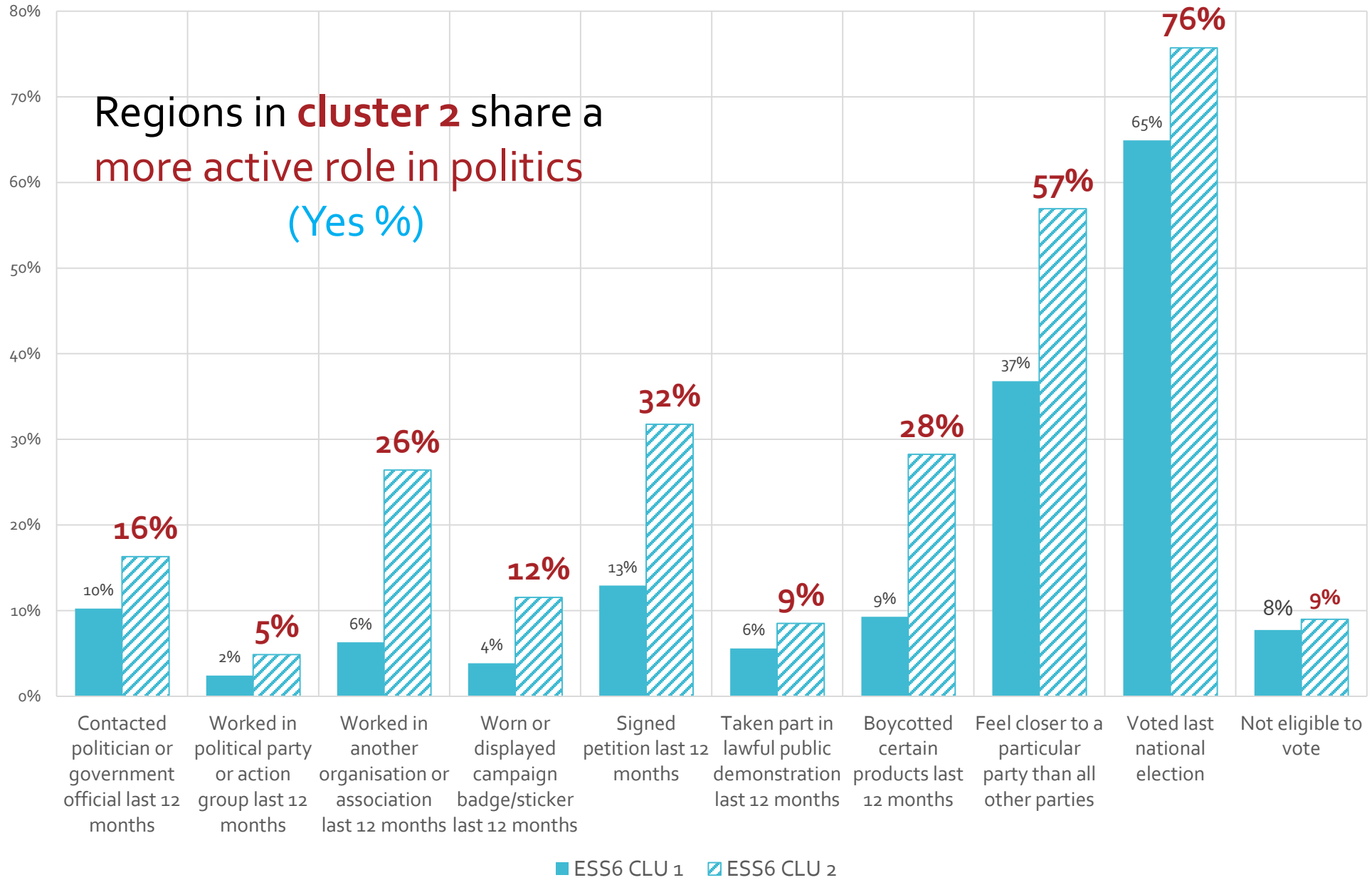
cohesion-separation
stability
computation time

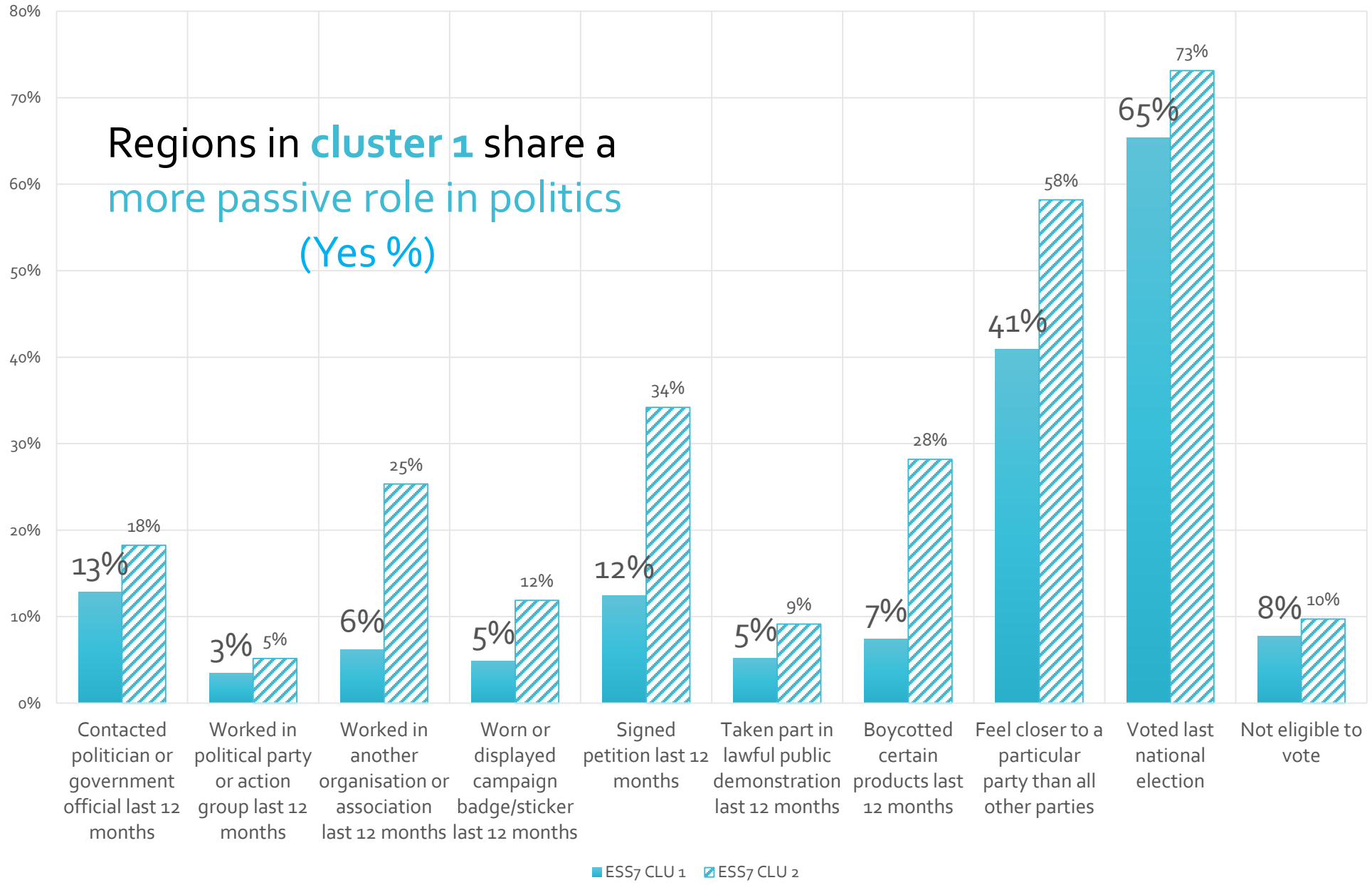
		BIC; CAIC; ICL	AIC; AIC_3	EM-MML
2012	Number of clusters	7	7	**2**
	Silhouette index	0.213	0.191	**0.361**
	Calinski-Harabasz	83.327	74.977	**190.825**
	Computation time (seconds)	109	109	**2**
2014	Number of clusters	7	8	**2**
	Silhouette index	0.152	0.164	**0.367**
	Calinski-Harabasz	80.766	78.477	**189.552**
	Computation time (seconds)	91	91	**2**
2012 vs 2014	Adjusted Rand	0.377	0.499	**0.707**
	Normalized mutual information	0.523	0.591	**0.598**

Results:
round 6 vs round 7



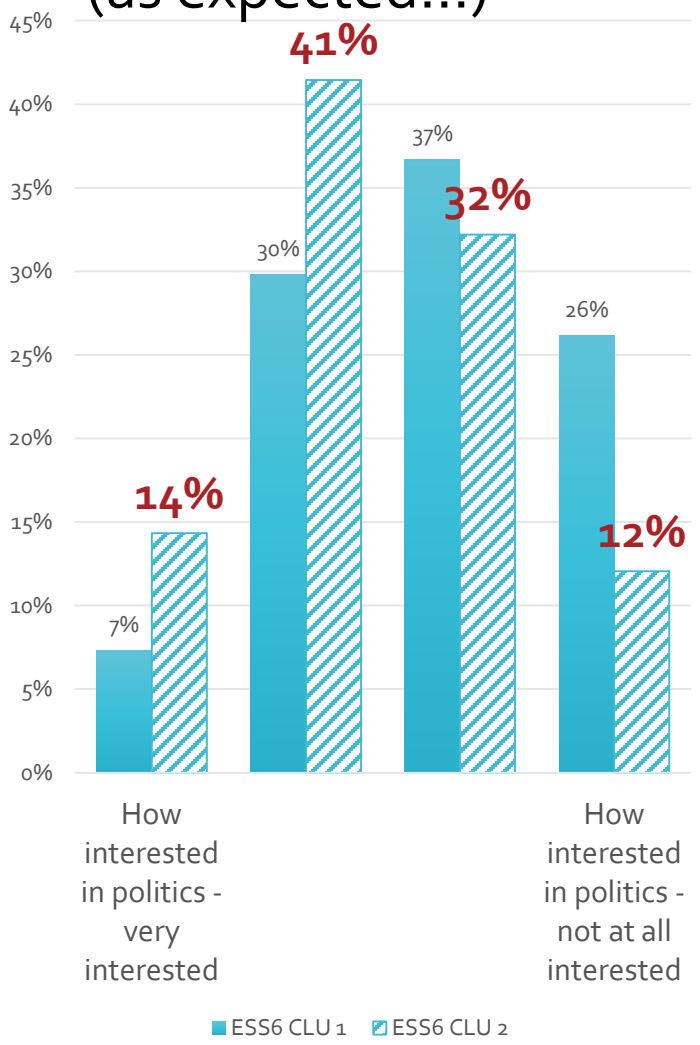


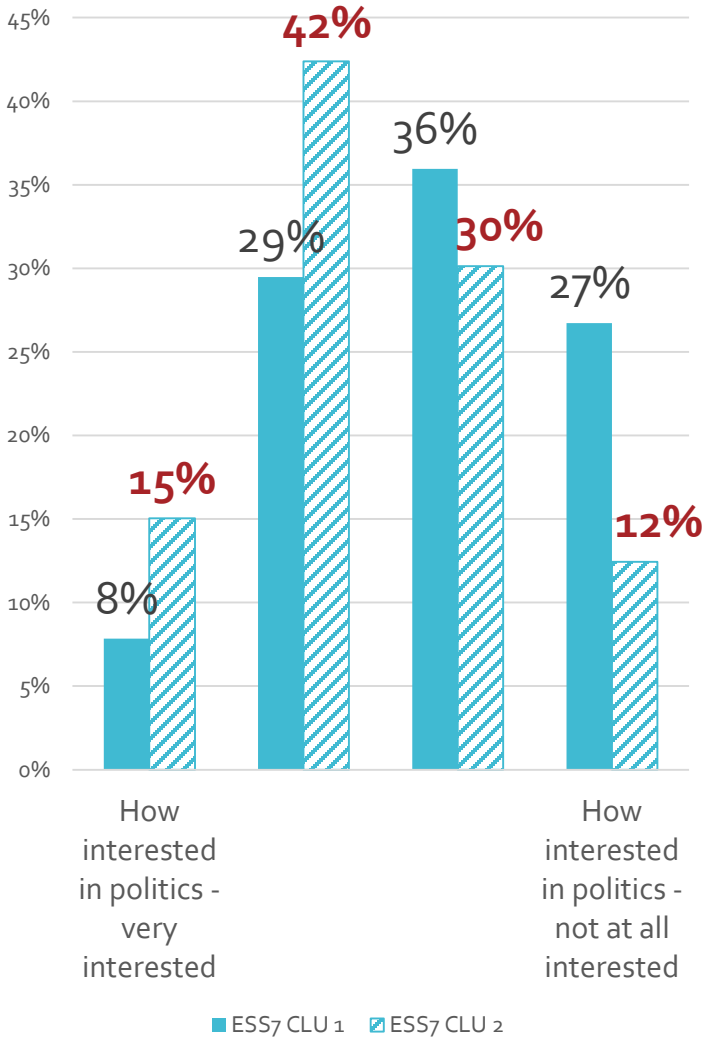

Regions in cluster 2 share a more active role in politics
(Yes %)
80%
70%
60%
50%
40%
30%
20%
10%
0%
10%
16%
2%
5%
6%
26%
4%
12%
13%
32%
6%
9%
9%
28%
37%
57%
65%
76%
8%
9%
Contacted politician or government official last 12 months
Worked in political party or action group last 12 months
Worked in another organisation or association last 12 months
Worn or displayed campaign badge/sticker last 12 months
Signed petition last 12 months
Taken part in lawful public demonstration last 12 months
Boycotted certain products last 12 months
Feel closer to a particular party than all other parties
Voted last national election
Not eligible to vote
ESS6 CLU 1
ESS6 CLU 2


Regions in cluster 1 share a more passive role in politics (Yes %)
80%
70%
60%
50%
40%
30%
20%
10%
0%
13%
18%
3%
5%
6%
25%
5%
12%
12%
34%
5%
9%
7%
28%
41%
58%
65%
73%
8%
10%
Contacted politician or government official last 12 months
Worked in political party or action group last 12 months
Worked in another organisation or association last 12 months
Worn or displayed campaign badge/sticker last 12 months
Signed petition last 12 months
Taken part in lawful public demonstration last 12 months
Boycotted certain products last 12 months
Feel closer to a particular party than all other parties
Voted last national election
Not eligible to vote
ESS7 CLU 1
ESS7 CLU 2

Results

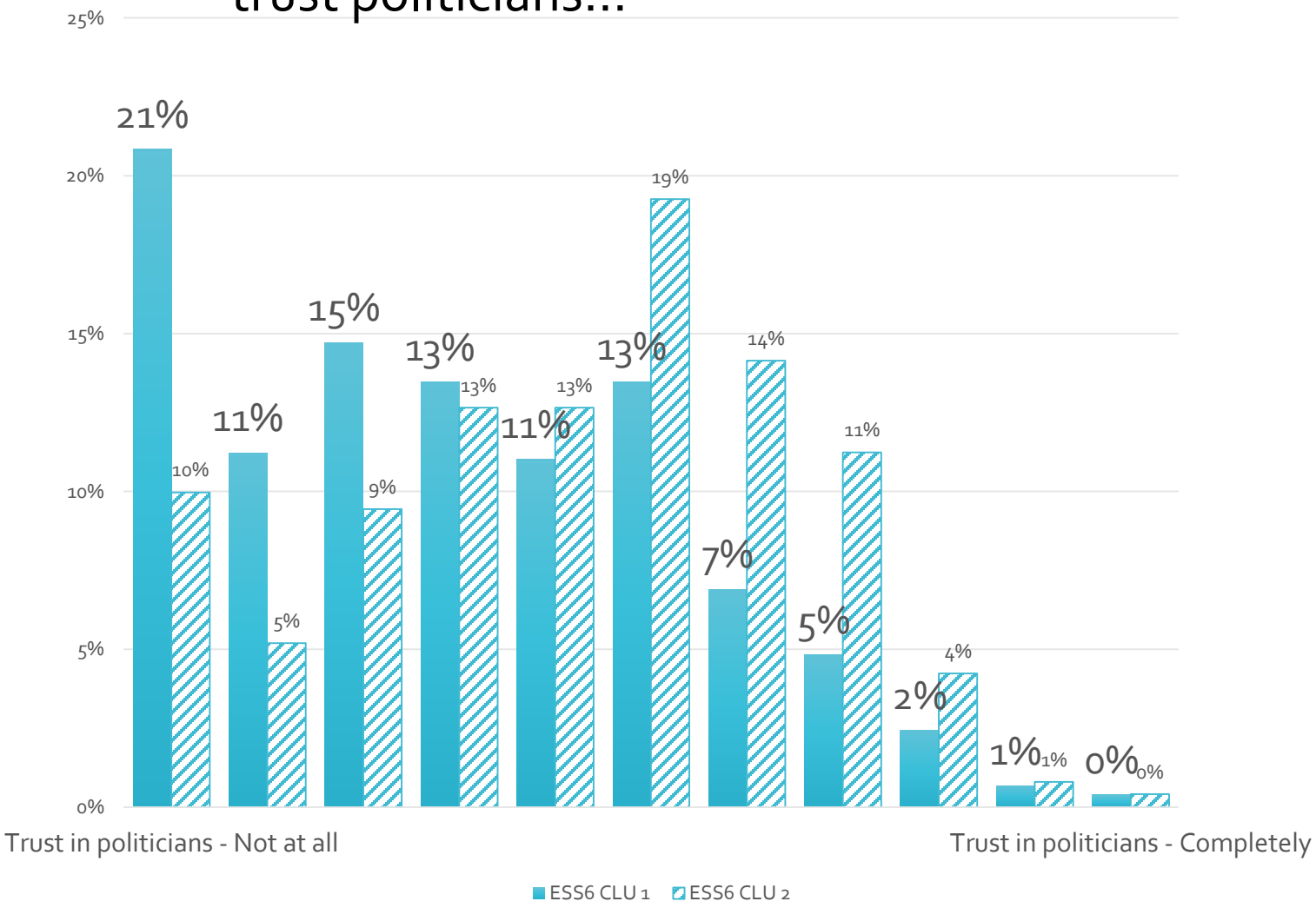
Regions in **cluster 2** are clearly more interested in politics (as expected…)





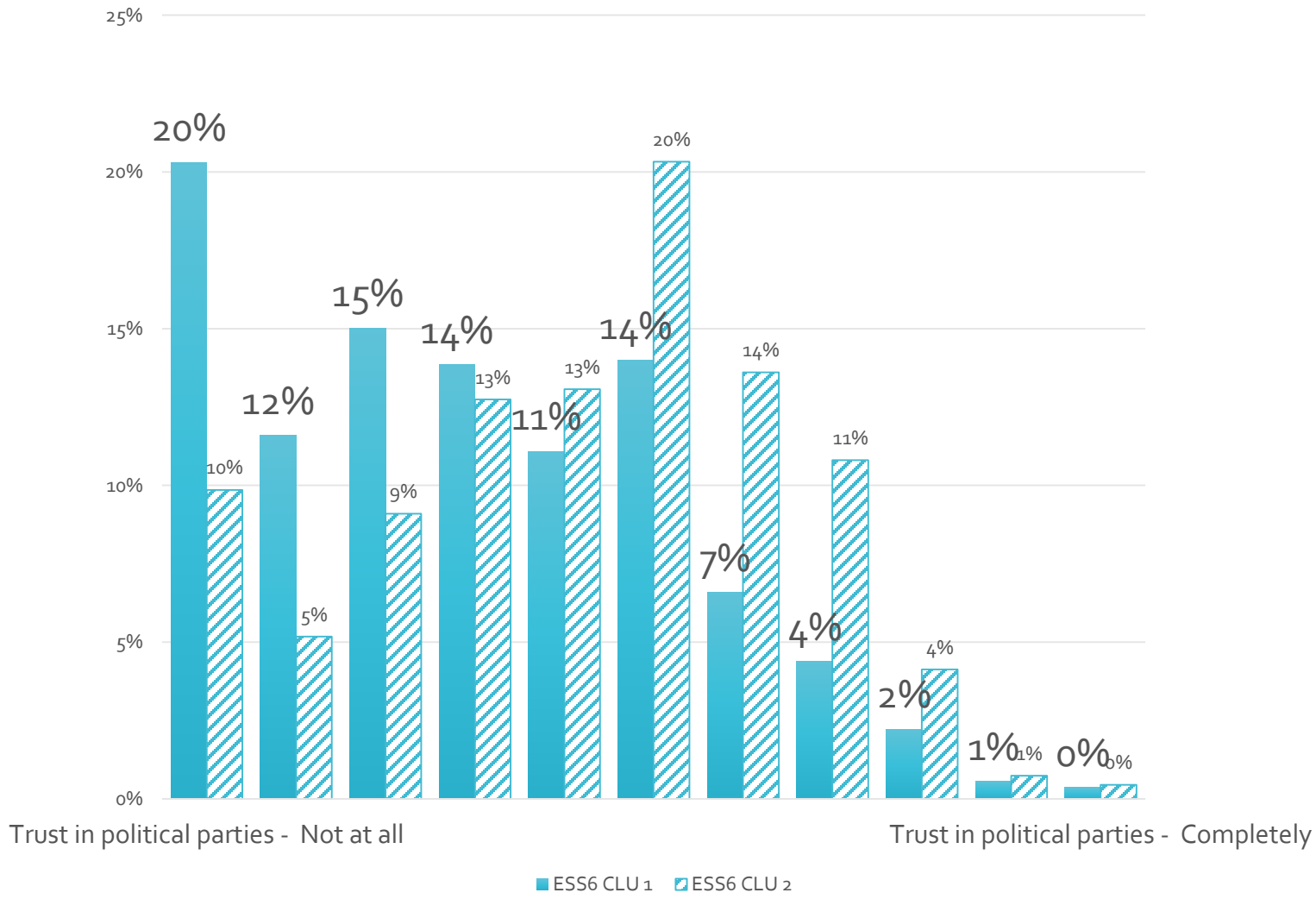
Results

Most respondents in Cluster 1 do not trust politicians…

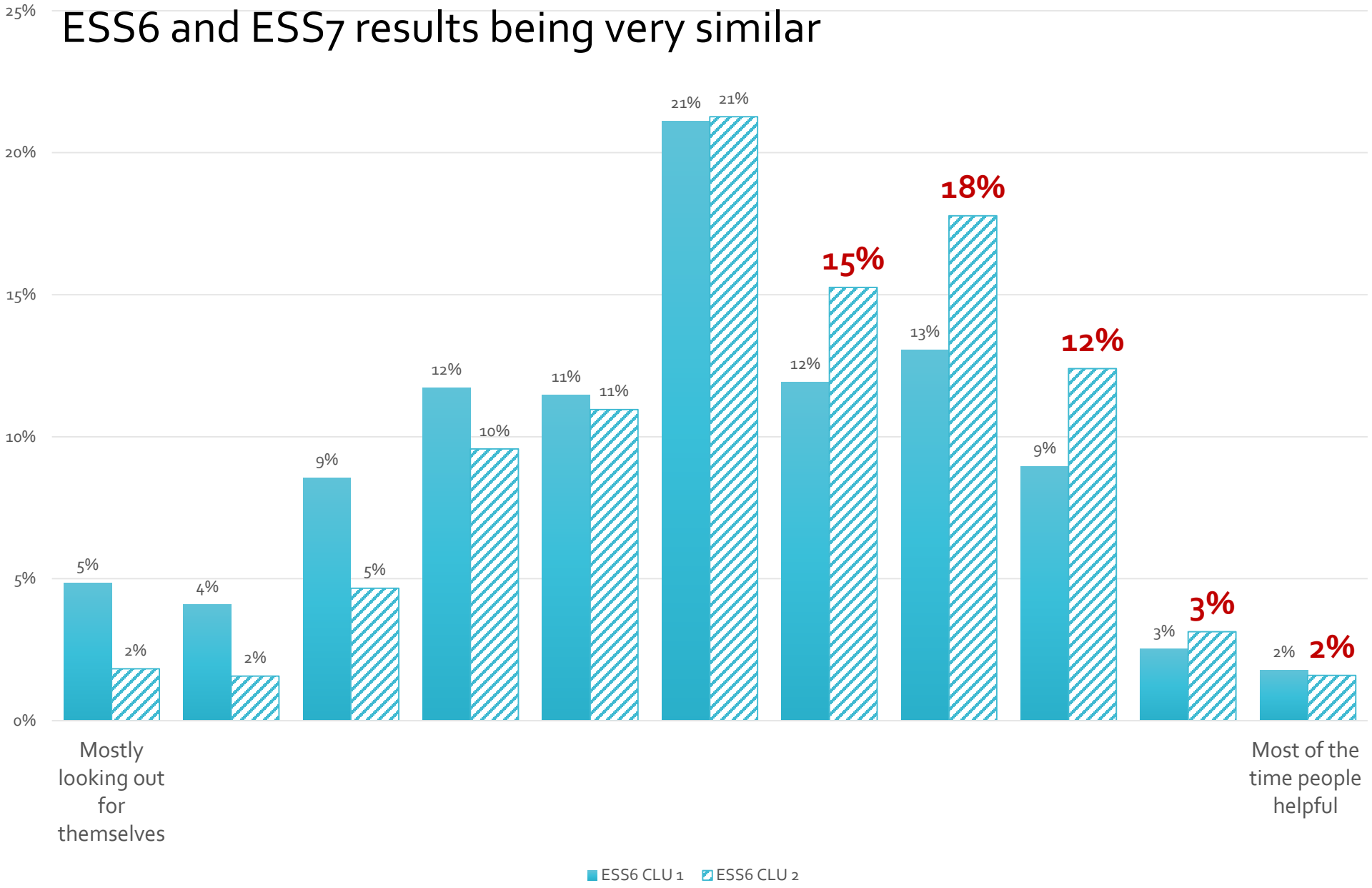


Results

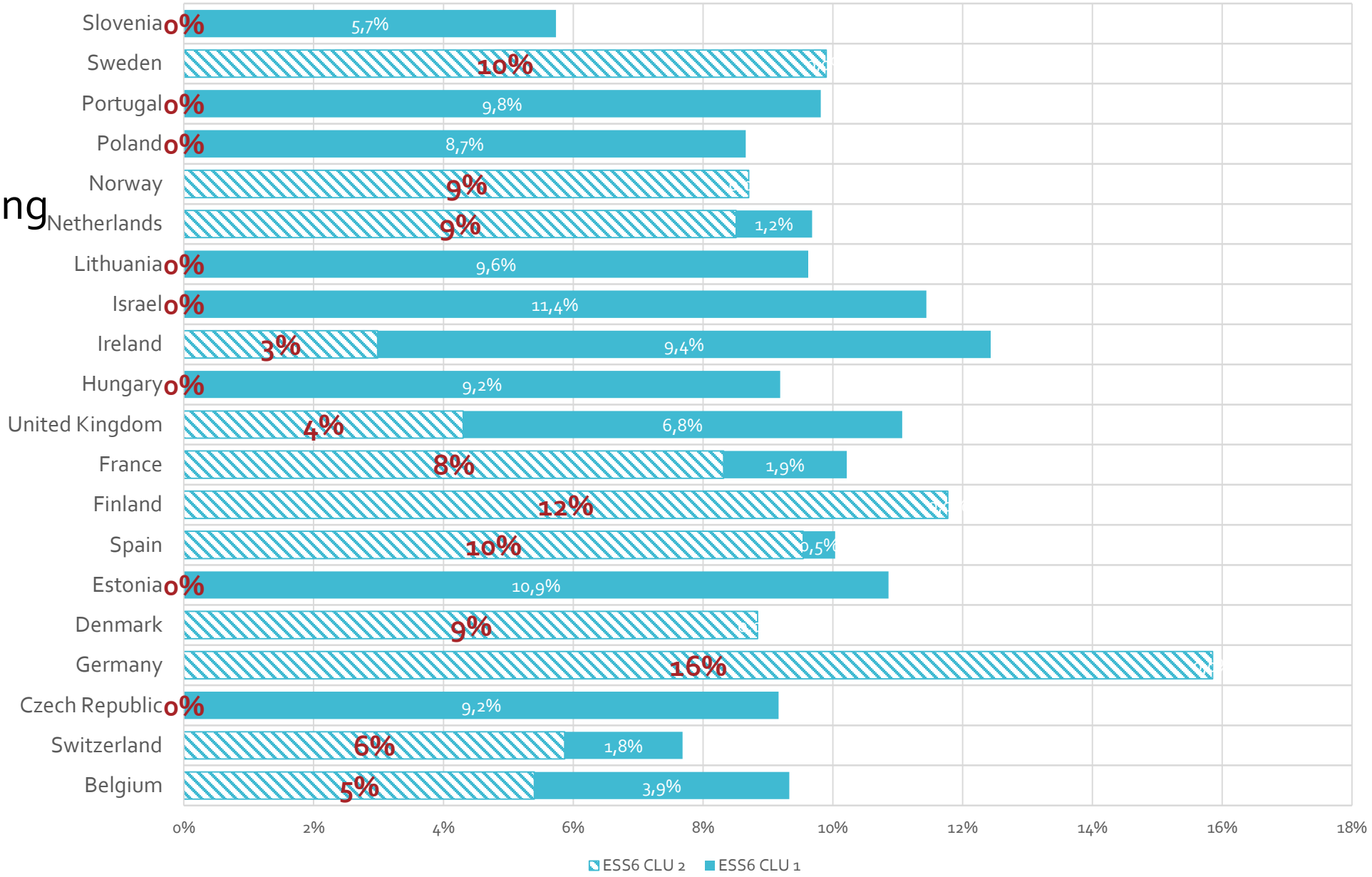
...or political parties



Regions in **cluster 2** share a more positive view of other people
ESS6 and ESS7 results being very similar



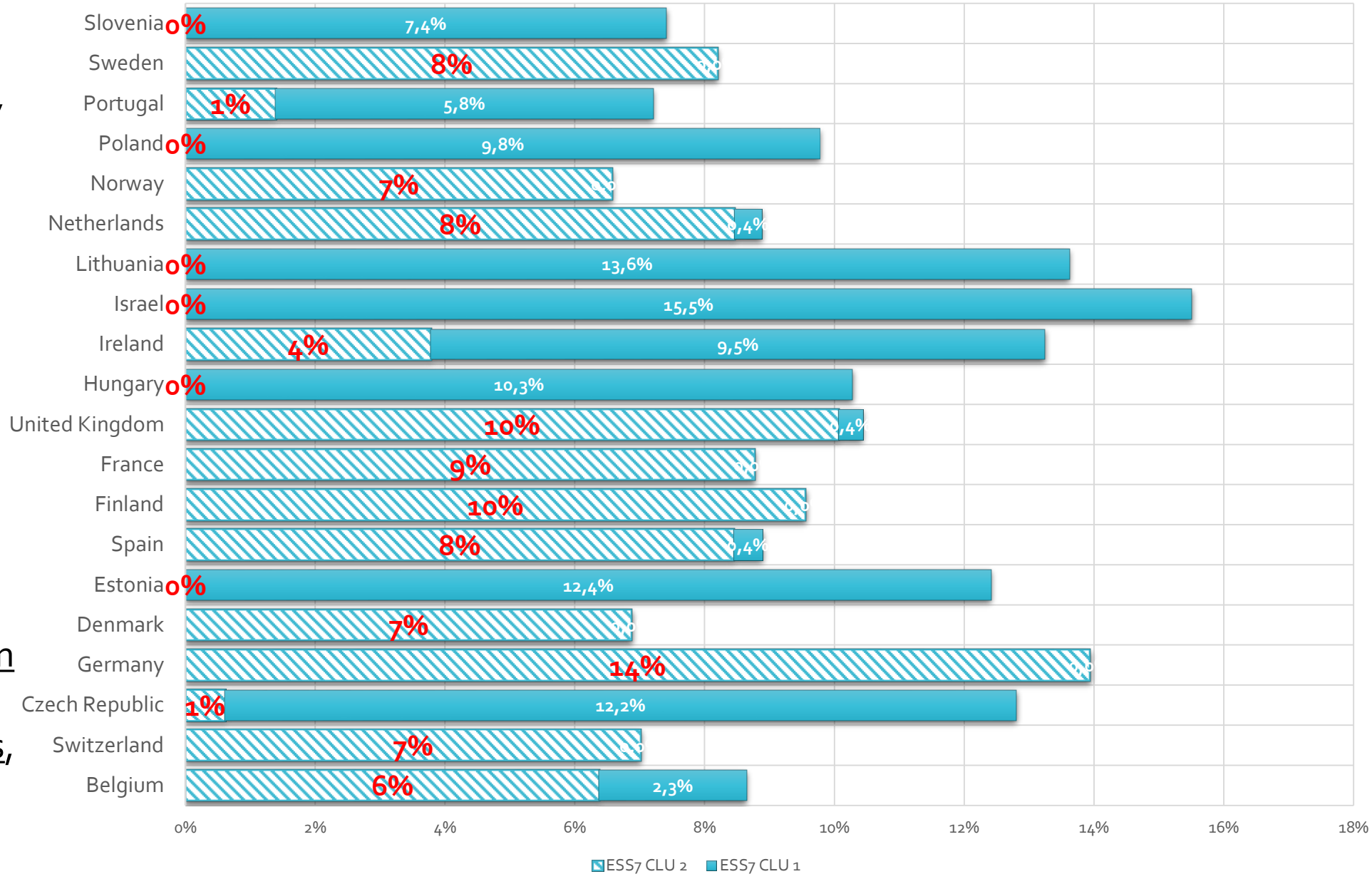
All regions in Sweden, Norway, Finland, Denmark and Germany belong to **cluster 2**



All regions in Sweden, Norway, Finland, Denmark and Germany belong to **cluster 2**

25 regions change to **cluster 2**, e.g. Lisbon (in Portugal) Jihoceský kraj (in Czech Republic)

4 regions change to **cluster 1**: Prov. West-Vlaanderen (in Belgium), Principado de Asturias, La Rioja (in Spain) and Drenthe (in Netherlands)



Conclusions

- A new EM variant – the EM-MML – was used to cluster categorical aggregated data and estimate the number of clusters simultaneously.
- It estimates parameters of a finite mixture of multinomials, using a Minimum Message Length criterion.
- EM-MML shows better performance when compared with traditional EM-ML combined with BIC, AIC and ICL: more parsimonious and robust solutions; better cohesion-separation and stability
- Abrief profiling of the segments showing that the main changes occurred between rounds 6 and 7

References

Biernacki, C., Celeux, G. and Govaert, G., 2000. Assessing a Mixture model for Clustering with the integrated Completed Likelihood. IEEE Transactions on Pattern analysis and Machine Intelligence, 22: 719–725.

Figueiredo, M. A. T., and Jain, A. K., 2002. Unsupervised learning of finite mixture models", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, pp. 381-396.

Fonseca, J. R. & Cardoso, M. G. (2007). Mixture-model cluster analysis using information theoretical criteria. Intelligent Data Analysis 11(2): 155-173.

Silvestre, C., Cardoso, M. and Figueiredo, M., 2008. Clustering with finite mixture models and categorical variables. Contributed Papers to the International Conference on Computacional Statistics, Porto, Portugal, pp. 109-116.

Silvestre, C., Cardoso, M., and Figueiredo, M., 2012. Categorical Data Clustering Using a Minimum Message Length Criterion. IDA 2012 - The eleventh International Symposium on Intelligent Data Analysis. Helsínquia, Finlândia, 25-27 de outubro, 2012.

Silvestre, C., Cardoso, M., and Figueiredo, M., 2013. Determining the Number of Groups while Clustering Categorical Data. IFCS 2013 – The International Federation os Classification Societies. Tilburg, the Netherlands, 14-17 July (Book of Abstracts p. 158)