Cluster Analysis of Vote Transitions;

How do people switch vote between consecutive elections?

Josep Ginebra
Xavi Puig

Department of Statistics and O.R.
Universitat Politècnica de Catalunya
February of 2013

1) Data
2) Goal
3) Model
4) Model Selection
5) Results

1. Data

Barcelona broken down into 248 small areas

In the poster you will see an implementation on Catalonia, broken down into 1463 small areas

1. Data

Catalan Parliament 2003									
District	area	CIU	PSOE	PP	ERC	ICV	*others*	*abs*	*N*
1	1	195	375	76	86	58	19	701	1510
1	2	208	333	75	97	70	26	790	1599
...	...	...	...	...	...	...	...	...	...
10	248	441	1535	592	245	229	82	2202	5326
Total		227783	249020	123163	126626	69234	19295	407294	1222415
Spanish Parliament 2004									
District	area	CIU	PSOE	PP	ERC	ICV	*others*	*abs*	*N*
1	1	141	488	127	156	52	28	496	1488
1	2	154	498	110	183	57	25	564	1591
...	...	...	...	...	...	...	...	...	...
10	248	375	2037	814	282	267	125	1372	5272
Total		188386	359254	171102	138762	65001	24489	268393	1215387

Table 1: Part of the results in the 2003 and 2004 elections in Barcelona.

1. Data

An observation $y_i = (y_i^1, y_i^2)$ is a set of two seven dimensional vectors of categorical data, each with the result in one of the two elections of the pair.

Two vectors ordered in time and located in space.

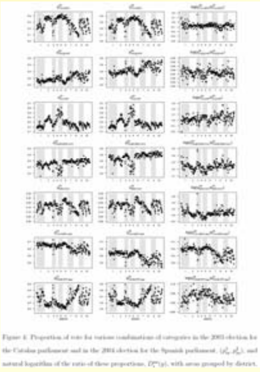
The data will have a strong spatial dependence.

We will need meaningful ways of summarizing the two tables with election results and the way these results change in each area.

1. Data

·CIU
·PSC
·PPC
·ERC
·ICV
·Others
·abs

1. Data



2. Goal

To estimate how do people switch their vote between two (consecutive) elections.

The two election results in an area are the two marginal distributions of a 7X7 contingency table, and the goal is to estimate the corresponding joint distribution (i.e., the 49 table cells).

We need to reconstruct individual behavior from aggregated data. Our problem is a special instance of an ecological inference problem.

Our approach to the problem can be exported to be a solution for any ecological inference problem.

2. Goal

Our approach consists in:

- Carrying out an s-cluster analysis of the areas, assuming that both the average voting behavior as well as the way in which individuals switch their vote in areas of the same cluster are similar.
- Estimating s vote switch matrices, each ruling the way in which individuals in an area of a given cluster change their vote between the two elections.

The cluster analysis and ecological inference analysis are carried out simultaneously through a Bayesian model.

3. Model

The model includes:

- The cluster analysis part is based on a finite mixture of Dirichlet-Multinomial models that groups areas into s clusters,
- The ecological inference part links the two elections through vote switch matrices determining the average voting behavior of an area in the second election starting from its first election result.

3. Model

The model is Bayesian.

One can update it in the light of data and simulate from it using Markov Chain Monte Carlo methods.

The actual implementation is made using WinBugs.

4. Model Selection (Number of Clusters)

One has a different model for each # of clusters.

The number of clusters, s, is chosen by:

A. Looking for the smallest s that makes it plausible that the s-cluster model could simulate data similar to the actual election results.

B. Checking whether the s-cluster model captures most of the spatial dependency in the actual results by testing whether its residuals are spatially dependent or not.

4. Model Selection (Number of Clusters)

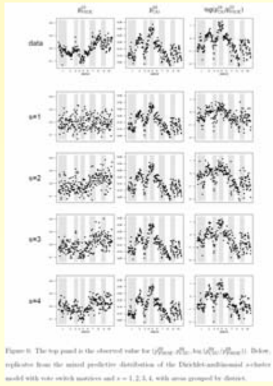
Election results are a set of two 248X7 tables.

How does one compare the tables with actual results with the tables with results simulated through models?

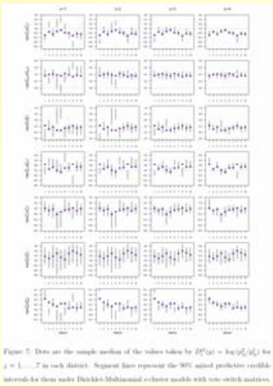
A lot of care is devoted to finding efficient ways to summarize the election results graphically in a way that one captures all the relevant details.

We have used 63 different statistics for that, and various different kinds of graphical displays.

4. Model Selection (Number of Clusters)



4. Model Selection (Number of Clusters)



4. Model Selection (Number of Clusters)

To check whether their "residuals" are spatially correlated or not, one needs to agree first on a definition of a residual.

An observation is two 7 dimensional vectors of categorical data. What do we use as a residual for that?

$$p_{ia}^{P2} = \frac{p_{ia}^2 - E[p_{ia}^2|y]}{\sqrt{Var[p_{ia}^2|y]}}, \quad i=1,2,\ldots,248.$$

We implement that on 63 different residuals

4. Model Selection (Number of Clusters)

As a measure of spatial dependency in the residuals we use as a statistic the Moran Index

$$I_M(p_a^{P2}) = \frac{248}{\sum_{i=1}^{248}\sum_{j=1}^{248}\lambda_{ij}} \frac{\sum_{i=1}^{248}\sum_{j=1}^{248}\lambda_{ij}(p_{ia}^{P2}-\overline{p_a^{P2}})(p_{ja}^{P2}-\overline{p_a^{P2}})}{\sum_{i=1}^{248}(p_{ia}^{P2}-\overline{p_a^{P2}})^2},$$

λ_{ij} is 1 if the zrp are neighbors and it is 0 if they are not.

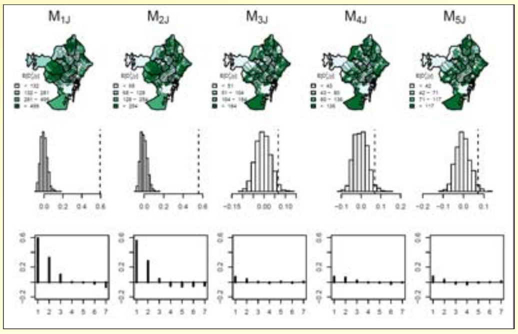
4. Model Selection (Number of Clusters)

We test whether the residuals of the models are spatially independent through permutation tests.

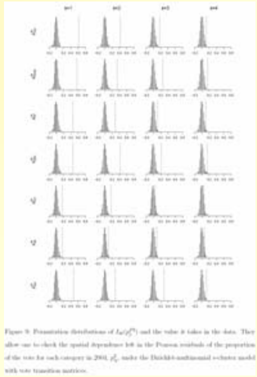
The idea is that if they were independent and one randomly shuffled their values on the map without changing the area labels and re-measured the spatial dependence in them, one would obtain a value similar to the spatial dependence measured in the actual results.

4. Model Selection (Number of Clusters)

Spatial distribution of a residual, Moran index and correlogram



4. Model Selection (Number of Clusters)



5. Results

In this case, we settle with a 4-cluster model.

The results of the analysis are presented through:

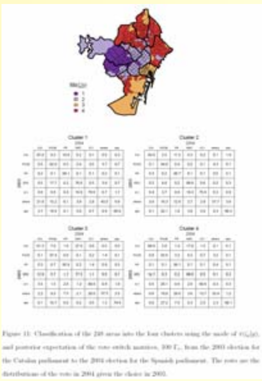
1. A table with the voting pattern, the relative size in # of areas and in pop., and a measure of the heterogeneity of each one of the 4 clusters,
2. a map classifying areas into clusters,
3. one vote transition matrix for each cluster, and an overall vote transition matrix obtained through a weighed average of the four cluster matrices.

5. Results

Elect	Cluster	*CIU*	*PSOE*	*PP*	*ERC*	*ICV*	*others*	*abs*	ω	% Pop	τ
2003	1	0.301	0.120	0.180	0.087	0.038	0.017	0.257	0.116	10.6	262.13
	2	0.223	0.175	0.090	0.135	0.064	0.016	0.298	0.337	38.9	428.23
	3	0.120	0.199	0.068	0.070	0.054	0.019	0.471	0.185	7.5	145.40
	4	0.132	0.250	0.094	0.082	0.060	0.019	0.363	0.361	43.0	155.05
2004	1	0.279	0.173	0.239	0.082	0.033	0.018	0.176	0.116	10.5	
	2	0.189	0.262	0.124	0.145	0.058	0.018	0.204	0.337	38.9	
	3	0.086	0.289	0.103	0.090	0.058	0.020	0.354	0.185	7.5	
	4	0.105	0.355	0.135	0.098	0.056	0.023	0.228	0.361	43.1	

Table 2: Posterior expected value of $\mu_r = E[\theta_i^1|\zeta_i = r]$ and of $E[\theta_i^2|\zeta_i = r]$, determining the voting patterns of the four clusters, and of ω_r, % Pop and τ_r, determining their relative size in terms of number of areas and of voting age individuals, and their heterogeneity.

5. Results



5. Results

Destination of the 2003 vote

2003 \ 2004	CIU	PSOE	PP	ERC	ICV	others	abs
CIU	78.6	3.2	7.8	6.8	0.6	0.2	2.8
PSOE	0.1	93.7	1.9	0.3	0.4	3.5	0.1
PP	0.2	0.2	98.8	0.1	0.2	0.4	0.1
ERC	6.0	6.4	0.3	85.9	0.6	0.5	0.3
ICV	0.7	13.6	0.7	9.7	74.3	0.3	0.7
others	4.2	16.3	19.1	3.2	9.0	45.3	2.9
abs	0.3	24.5	5.6	1.7	2.2	1.3	64.3

Origin of the 2004 vote

2003 \ 2004	CIU	PSOE	PP	ERC	ICV	others	abs
CIU	94.3	2.0	10.3	11.1	2.1	1.8	2.4
PSOE	0.1	64.6	2.8	0.5	1.7	35.0	0.1
PP	0.1	0.1	71.0	0.1	0.3	2.1	0.1
ERC	4.0	2.2	0.2	78.0	1.1	2.4	0.1
ICV	0.2	2.6	0.3	4.8	78.6	0.9	0.2
others	0.4	0.9	2.1	0.4	2.7	35.5	0.2
abs	0.7	27.6	13.3	5.0	13.5	22.2	96.9

Figure 12: The rows of the first matrix are the overall distributions of the vote in 2004 of all the individuals with a given choice in 2003. The columns of the second matrix are the overall distributions of the vote in 2003 of all the individuals with a given choice in 2004.