Computational Statistics and Data Analysis 70 (2014) 328–344

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A cluster analysis of vote transitions☆

Xavier Puig*, Josep Ginebra

Department of Statistics, Technical University of Catalonia, Avgda.Diagonal 647, 6ª Planta, 08028 Barcelona, Spain

ARTICLE INFO

Article history:
Received 26 November 2012
Received in revised form 30 September 2013
Accepted 7 October 2013
Available online 18 October 2013

Keywords:
Bayesian model checking
Bayesian hierarchical model
Ecological inference
Election data
Spatial data

ABSTRACT

To help settle the debate triggered the day after any election around the origin and destination of the vote of winners and losers, a Bayesian analysis of the results in a pair of consecutive elections is proposed. It is based on a model that simultaneously carries out a cluster analysis of the areas in which the results are broken into and links the results in the two elections of areas in a given cluster through a vote switch matrix. The number of clusters is chosen both through predictive checks as well as by testing whether the residuals are spatially correlated or not. The analysis is tried on the results in Barcelona of a pair of consecutive elections held just four months apart, in 2003 for the Catalan parliament and in 2004 for the Spanish parliament. The proposed approach, which reconstructs individual behavior from aggregated data, can be exported to be a solution for any ecological inference problem where one cannot assume that all the areas are exchangeable the way typically assumed by other ecological inference methods.

1. Introduction

The day after any election a debate is always triggered around the way voters switched their vote or switched to (from) not voting from (to) voting for this or that option relative to previous elections. That debate is especially poignant in Catalonia, an autonomous region in north-east Spain, due to its voters splitting across a national allegiance divide on top of the usual ideological divide, which leads to a lot of options to chose from and to individuals voting very differently depending on the kind of election at hand.

To help assess how voting age individuals change their vote, a Bayesian analysis is proposed based on a model for the results of a pair of consecutive elections broken down into small areas. The model simultaneously carries out an s-cluster analysis on 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, and it estimates s vote switch matrices, each ruling the way in which individuals in an area of a given cluster change their vote between the first and the second election. The number of clusters is chosen by checking whether the corresponding models capture the levels, the dispersion and the spatial dependence both in the election results as well as in the way these results change.

Disregarding the fact that the voting age population in the areas considered changes slightly between the first and the second election, one can consider the first election results of each area to be the row totals and the second election results of that area to be the column totals of a $k_1 \times k_2$ contingency table. Posed in these terms, the goal of the analysis is the estimation of the $k_1 k_2$ cells of these tables, which is the canonical formulation of the ecological inference problem in the social sciences, where one aims to extract information about individual behavior starting from information reported only at an aggregate level. Good overviews of the approaches considered for that problem can be found in King (1997), Freedman (2001), Freedman et al. (1991, 1998) and Glynn and Wakefield (2010).

☆ Supplementary materials are available online expanding on Sections 2 and 4.

* Corresponding author. Tel.: +34 93 4016737.
E-mail addresses: xavier.puig@upc.edu (X. Puig), josep.ginebra@upc.edu (J. Ginebra).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.10.006

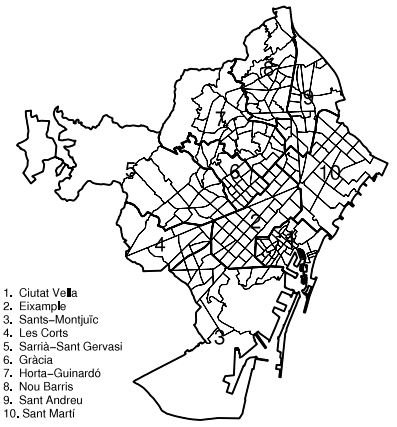

Fig. 1. Map of Barcelona divided in ten districts and 248 smaller areas.

The most serious drawback of the basic methods proposed for that problem, mostly based on ecological regression models with or without random coefficients, is their reliance on the assumption of the similarity across tables and hence that all the areas are exchangeable. That 'constancy assumption', imposed by forcing that the coefficients of the ecological regression model either be the same or share the same distribution for all areas, is often inappropriate due to the existence of contextual effects. In our case for example, one expects the vote pattern and the vote switch pattern of an area to depend on its demographic composition, and hence to be strongly associated to the location of that area. Hence, areas will not be exchangeable. One way around that drawback is to allow the coefficients of the ecological regression models to depend on area level covariates, but that requires one to know which are the relevant contextual characteristics and to be able to measure them, which most of the time is not feasible.

Embedding an ecological regression model into a cluster model the way advocated here is an alternative way around the failure of that 'constancy assumption'. In our setting the parameters of the ecological inference part of the model will be constant (or share the same distribution) only among areas in the same cluster. That allows one to estimate vote switch patterns by only pulling in the information of areas that are similar. The clusters found will typically be determined by the contextual characteristics that set areas apart, without having to make explicit and measure these covariates. In fact, the special one-cluster version of our model has a lot in common with the basic ecological regression models first considered by Goodman (1953, 1959). By considering the more general s-cluster version of the model and letting the data choose whether the one-cluster model is appropriate or not, (in our example it is clearly not appropriate), one is in effect checking whether the usual 'constancy assumption' holds or not.

The article is organized as follows. Section 2 describes the results of a pair of recent elections in Barcelona that will be used as a showcase example. Section 3 presents the model and briefly relates it to the models used by the social sciences literature on ecological inference problems. Section 4 describes how one can decide on the number of clusters and check the final model based first on predictive checks comparing the levels and the variability of the actual election results with simulations from the models, and second by testing whether the residuals of the models are spatially correlated or not. Special care is devoted to picking up the statistics, graphics and residuals that best fit the main objective of the analysis.

Section 5 presents the results of the analysis of the pair of elections in Barcelona, where it is found that individuals vote and switch their vote according to four different voting patterns and vote switch patterns depending on the area where they live. Even though the model is blind with regard to the location of the areas, the four-cluster structure uncovered has a rather strong spatial structure, which simplifies the interpretation of the results. In settings like the one of the example, where individuals vote and switch their vote very differently depending on where they live, using models that do not take into account the cluster structure of the data can lead to different, and potentially misleading conclusions. Section 6 discusses variations of the analysis, and ponders the limitations inherent to ecological inference methods.

2. Description of the example

To illustrate our approach, we will consider the results of a pair of consecutive elections in Barcelona, which is the capital of Catalonia and holds about 20% of its population. Politics in Catalonia are unusual because of the existence of two cleavages, due to voters splitting across a national allegiance divide on top of the usual ideological divide. As a consequence, the results of an election change a lot depending on whether it is for the Catalan parliament or for the Spanish parliament. The first election of the pair considered here was held on November 16th, 2003 and it was for the Catalan parliament, while the second election was held on March 14th, 2004 and it was for the Spanish parliament.

Barcelona is organized in the 10 districts and 248 areas shown in Fig. 1. The results of these elections in these areas are grouped into 7 categories. The first five correspond to the parties or coalition of parties with seats in both parliaments, labeled CIU, PSOE, PP, ERC and ICV. All the votes for parties obtaining less than 1% of the vote, null votes, and votes casted