Spatial variation in ecological inference

Juan Dodyk[*]

Abstract

Ecological inference is the problem of estimating individual-level behavior from aggregate data. In this article I propose and apply a new statistical model that provides a solution to the ecological inference problem under two conditions: 1) geographic space is divided into regions, and aggregate data comes from random samples of individuals from each region; 2) variation in individual-level behavior is spatially smooth, i.e., individuals who are geographically close behave similarly (in aggregate). As an application, I use the model to estimate voter transition rates in Argentina's 2015 presidential election. The results provide evidence that there is substantive spatial variation in the transition rates, and that it can be explained by structural territorial cleavages, local coalitional patterns and class segregation in urban settings.

1 Introduction

Ecological inference is the problem of estimating individual-level behavior from aggregate data. More specifically, given two discrete variables indicating distinct behaviours of individuals in a population divided into p units, the problem consists in estimating the unobserved interior cells of the contingency tables at each unit given the observed row and column marginals. This problem has applications in Voting Rights litigation, electoral behavior, epidemiology, marketing, political campaigning, and economics (King, 1997).

Ecological inference is, and will remain, an unsolvable problem in its general formulation. It is described as an "ill-posed inverse problem" by Cho and Judge (2008), plagued by indeterminacy. However, lots of methodological strategies have been devised since Goodman's landmark

[*] Please send me an email to dodykjuan@gmail.com if you are interested in the code and data used for the article.

paper (1953) to provide answers to concrete instances of it. These strategies include deterministic inferences, statistical models, information-theoretical approaches, and the incorporation of survey data (King et al, 2004; Cho and Manski, 2008; Klima et al., 2015). These methods come with trade-offs, one of which is that the inferential power required by applications comes at the cost of strong assumptions, which can be difficult to test in general.

In this paper I propose a new approach, inspired by machine learning methods, that is tailored to the Argentine electoral data but can be extended to other contexts. This model is based on the Multinomial-Dirichlet (MD) model (Rosen et al, 2001), and admits nonparametric variation in its parameters, only constrained by a very general spatial smoothness condition. I use it to model spatial dependence in a novel way, that allows the estimation of $R \times C$ tables directly[1].

To illustrate the methodology, I use the proposed model to analyze the data of the 2015 Argentine presidential elections. I show how the majority of the vote that Mauricio Macri was able to muster in the runoff was assembled incrementally in three stages. To that end I provide point-estimates of voter transition rates between these stages. Since territorial cleavages were in play in the election (Freytes and Niedzwiecki, 2016) and were a decisive factor in the transition of the vote share of Sergio Massa, a pivotal candidate, to the runoff candidates, this provides an opportunity to test the method in a real scenario in which extreme spatial heterogeneity affects the modeling. This also lets me make a substantive contribution: the model results provide evidence that structural territorial cleavages, local coalitional patterns and class segregation in urban settings had a strong impact on voter transition rates.

2 Ecological inference: models and assumptions

In this section I will introduce notation to formulate the problem and review some methods proposed to approach it, leading to the Multinomial-Dirichlet (MD) model of Rosen et al. (2001), on which I will later work. I will focus on the fundamental ideas underlying the statistical models and the assumptions on which they depend.

[1] For previous statements of the challenge of spatial dependence for ecological inference and proposed strategies see Anselin and Cho (2002), Calvo and Escolar (2003), Haneuse and Wakefield (2004).

Notation. I will follow the notation of Rosen et al (2001) and will interpret the parameters in terms of voter transition rates in the context of the 2015 Argentine presidential elections. Argentine voting-age population was divided into geographic units called *circuitos electorales* (electoral precincts). In each precinct, people were randomly assigned to voting booths[2]. Electoral authorities publish voting results for each voting booth. Thus, we have a voting-age population divided into these $p \approx 90,000$ units (voting booths), and we have the following data for each unit i ($i = 1, \dots, p$): counts N_i of people assigned to that unit, fractions X_{ir} of people who voted for candidate r ($r = 1, \dots, R-1$) in the first election, with $X_{iR} = 1 - \sum_{r=1}^{R-1} X_{ir}$ corresponding to absentees or blank votes, and fractions T_{ic} of people who voted for candidate c ($c = 1, \dots, C-1$) in the second election, with, again, $T_{iC} = 1 - \sum_{c=1}^{C-1} T_{ic}$. The unobserved quantities β^i_{rc} ($r = 1, \dots, R$, $c = 1, \dots, C$) are the fraction of candidate-r voters in the first election who voted for candidate c in the second election. We have $\sum_{c=1}^{C} \beta^i_{rc} = 1$ for all r, i, $\beta^i_{rc} \geqq 0$ for all r, c, i, and the accounting identity $T_{ic} = \sum_{r=1}^{R} \beta^i_{rc} X_{ir}$.

Let

$$B_{rc} = \frac{\sum_{i=1}^{p} \beta^i_{rc} X_{ir} N_i}{\sum_{i=1}^{p} X_{ir} N_i}$$

be the fraction of people in the population who voted for candidate r in the first election that voted for candidate c in the second election. Ecological inference is the problem of estimating these (global) quantities B_{rc} that fill the voter transition $R \times C$ table using the marginals X_{ir}, T_{ic}. We will aid inference by including covariates Z_{ik} ($k = 1, \dots, K$) measuring quantity k at unit i, and geographic information about the units (spatial coordinates, ascription to administrative regions –precinct, district, province–, contiguity relations and distances between them).

The method of bounds. Duncan and Davis (1953) proposed a deterministic approach to the estimation problem. The idea is to obtain logical bounds for the true quantities β^i_{rc} at each

[2] Electoral authorities assign people to voting booths following the lexicographical order of their surnames. I assume that this assignment is statistically independent of voting behavior at the precinct level. In other words, the order of the first letters of people's surnames is not correlated with any determinant of electoral choice. Thus, even though the assignment is deterministic, it is *as if* it were random with respect to voting behavior. Note, however, that while individuals assigned to a voting booth have electoral preferences that can be seen as random samples from the distribution of preferences among the precinct population, they are not *independent* samples, since families, having the same surname, tend to vote in the same booth, and share conditions that shape their vote choice. We can therefore expect the distribution of the election results at the ballot box level to have the same mean but higher variance than the corresponding hypergeometric distribution.

unit and use them to obtain bounds for B_{rc}. Concretely, we can easily prove that

$$\beta^i_{rc} \in \left[\frac{\max\{0, X_{ir} + T_{ic} - 1\}}{X_{ir}}, \frac{\min\{X_{ir}, T_{ic}\}}{X_{ir}}\right],$$

so

$$B_{rc} \in \left[\frac{\sum_{i=1}^p \max\{0, X_{ir} + T_{ic} - N_i\}N_i}{\sum_{i=1}^p X_{ir}N_i}, \frac{\sum_{i=1}^p \min\{X_{ir}, T_{ic}\}N_i}{\sum_{i=1}^p X_{ir}N_i}\right].$$

Goodman's model. The first statistical approach to the ecological inference problem was proposed by Goodman (1953, 1959). It rests on the *constancy assumption*: $\mathbb{E}(\beta^i_{rc}|X_i) = \mathfrak{B}_{rc}$ for all units i, where $\mathfrak{B}_{rc} =_{\text{def}} \mathbb{E}(\beta^i_{rc})$. Under that assumption we have

$$\mathbb{E}(T_{ic}|X_i) = \mathbb{E}\left(\sum_{r=1}^R \beta^i_{rc}X_{ir}|X_i\right) = \sum_{r=1}^R \mathbb{E}(\beta^i_{rc}X_{ir}|X_i) = \sum_{r=1}^R \mathfrak{B}_{rc}X_{ir},$$

and therefore the estimator $\hat{\mathfrak{B}} = (X^tX)^{-1}X^tT$ is unbiased, provided X^tX is nonsingular. Moreover, under the following (mildly) stronger assumptions $\hat{\mathfrak{B}}$ is also consistent and asymptotically normal[3] (a) independence of β^i for different i; (b) regularity: $\operatorname{plim}_{p\to\infty} \frac{1}{p}X^tX = Q$ and $\operatorname{plim}_{p\to\infty} \frac{1}{p}X^tDX = R$ are nonsingular matrices, where $D \in \mathbb{R}^{p\times p}$ is diagonal with $D_{ii} = X_iV(\beta^i_{\bullet c}|X_i)X^t_i$, $V(\cdot)$ is the variance operator, and $\beta^i_{\bullet c} = (\beta^i_{1c}, \ldots, \beta^i_{Rc})$. Also, although $\hat{\mathfrak{B}}_{rc}$ estimates the mean voter transition rate $\mathbb{E}(\beta^i_{rc})$, not the population transition rate B_{rc}, we have $\operatorname{plim} B_{rc} = \mathfrak{B}_{rc}$ by the weak law of numbers as long as (β^i, X_i, N_i) are independent, identically distributed and N_i are bounded.

If we restrict our population to an Argentine electoral precinct, in which voters are randomly assigned to voting booths, the main assumption of Goodman's model is justified. In effect, the distribution of the number of candidate-c voters among the candidate-r voters assigned to voting booth i, given that the number of candidate-r voters is $X_{ir}N_i$, is hypergeometric with mean $B_{rc}X_{ir}N_i$. In other words, once we fix the fraction of candidate-r voters in the sample as X_{ir}, the expected fraction of candidate-c voters among those will be B_{rc}. Therefore, $\mathbb{E}(\beta^i_{rc}|X_{ir}) = B_{rc}$, which implies the constancy assumption and that $\mathfrak{B}_{rc} = B_{rc}$ for all r, c. Moreover, while the stronger assumptions needed for statistical inference may not hold exactly,

[3] The proof is a straightforward variation of the proof of consistency and asymptotic normality of the OLS estimator (see, e.g., Greene, 2011). Asimptotic normality means that $\sqrt{p}(\hat{\mathfrak{B}}_{\bullet c} - \mathfrak{B}_{\bullet c})$ converges in distribution to a multivariate normal with zero mean and variance $Q^{-1}RQ^{-1}$.

independence (condition a) is approximately correct and regularity (condition b) is reasonable.

Why might the constancy assumption fail? I identify two general reasons. First, there may be a causal connection linking X_i to β^i_{rc}. Concretely, for the second election local political leaders might pursue different campaigns targeted to voters from different electoral precincts depending on the results of the first election at each place. For example, they might attempt to increase turnout in precincts where the first-election results were favorable and to decrease it elsewhere. Another possible causal link is the following: voters might make strategic voting decisions for the second election conditional on first-election results at the local level (see Calvo and Escolar, 2003). I identify a second general reason: there may be a joint effect of sociodemographic and local-political variables Z_i both on X_i and β^i_{rc}, leading to a spurious correlation:

$$\begin{tikzcd}
X_i \arrow[rr, dash, dotted, bend left] & & \beta^i_{rc} \\
& Z_i \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

In addition, independence between units (condition a) may fail as a result of spatial autocorrelation.

Goodman (1959) offers ways to model violations to the constancy assumption. He allows a covariate Z_{ir} to explain the variation of β^i_{rc} between units, leading to the assumption $\mathbb{E}(\beta^i_{rc}|X_i) = \mathfrak{B}_{rc} + \delta_{rc}Z_{ir}$ and the linear model

$$\mathbb{E}(T_{ic}|X_i) = \sum_{r=1}^{R} X_{ir}\mathfrak{B}_{rc} + \sum_{r=1}^{R} X_{ir}Z_{ir}\delta_{rc}.$$

Note, however, that we can not model the variation of $\mathbb{E}(\beta^i_{rc}|X_i)$ directly as a linear function of X_{ir} because the model becomes non-identifiable (see Goodman, 1959, p. 623).

The Multinomial-Dirichlet model. In 2001, Rosen, Jiang, King and Tanner published an article proposing the Multinomial-Dirichlet (MD) model, which is an extension of the Binomial-Beta model of King, Rosen and Tanner (1999) to $R \times C$ tables. It uses the information on bounds for the local parameters β^i_{rc} (in contrast to Goodman's model), admits and explicitly models the variation among the $\beta^i_{rc}|X_i$ for different i, and incorporates covariates in a natural way. It is a hierarchical model. At the first level, for each unit i, $(T_{1i}N_i, \ldots, T_{Ci}N_i)$ is modeled as following a multinomial distribution (independent across units) with count N_i and parameters $(\theta_1^i, \ldots, \theta_C^i)$, with $\theta_c^i = \sum_{r=1}^R \beta_{rc}^i X_{ir}$. At the second level, for each unit i and each row r, the vector $(\beta_{r1}^i, \ldots, \beta_{rC}^i)$ is modeled as following a Dirichlet distribution (independent across units and rows, i.e., different values of i and r) with parameters $(d_r \exp(\gamma_{rc}+\delta_{rc}Z_i), \ldots, d_r \exp(\gamma_{r,C-1}+\delta_{r,C-1}Z_i), d_r)$.[4] Finally, at the third level, the variables d_r $(r = 1, \ldots, R)$ follow independent exponential distributions with means $1/\lambda$ for a fixed λ (or, in the version implemented by Lau, Moore and Kellerman (2007), Gamma distributions with fixed parameters λ_1, λ_2). If one uses Bayesian inference to derive posterior distributions of the parameters of interest γ_{rc} (and δ_{rc} if there is a covariate), one can put flat or normal priors on them.

Note that the assumptions on which Goodman's model rested, i.e., constancy, independence between units, and regularity, are also needed to guarantee inferences based on the Multinomial-Dirichlet model (for the regularity condition, see Proposition 1 in Rosen et al, 2001, p. 144). The constancy assumption can only be relaxed if the variation of $\beta_{rc}^i | X_i$ is precisely modeled through a measured covariate, as in Goodman (1959). Klima et al (2015) provide evidence that the assumptions are to some extent necessary: under extreme spatial heterogeneity or aggregation bias (which breaks the constancy assumption) the estimates of the quantities B_{rc} are severely off the mark. However, they also show that the MD model compares favorably to other statistical models for $R \times C$ tables on voter transition synthetic data, even under extreme model violations.

Nonlinear least squares. The nonlinear least squares estimator for the MD model is based on first moments: if η is defined as $(\gamma_{rc}, \delta_{rc})_{r,c=1,1}^{R,C}$, then $\mathbb{E}(T_{ci}|\eta) = \sum_{r=1}^R \mathbb{E}(\beta_{rc}^i|\eta) X_{ir}$, where

$$\mathbb{E}(\beta_{rc}^i|\eta) = \frac{\exp(\gamma_{rc} + \delta_{rc}Z_i)}{1 + \sum_{j=1}^{C-1} \exp(\gamma_{rj} + \delta_{rj}Z_i)}.$$

[4] If there is not a covariate, plug $Z_i = 0$ in the parameter vector. Note that, unlike in Goodman (1959) model, the covariate has a linear effect not directly on β_{rc}^i but on the log-odds ratio: $\log\left(\frac{\mathbb{E}(\beta_{rc}^i)}{\mathbb{E}(\beta_{rC}^i)}\right) = \gamma_{rc} + \delta_{rc}Z_i$; this makes the model identifiable for almost all parameters (in the sense of Lebesgue) even if one takes X_{ir} as the covariate. However, Rosen et al (2001, p. 145) recommend to use a nonlinear transformation, since non-identification occurs in the crucial case where $\delta = 0$, and thus the model cannot be used to test the null hypothesis of no effect of X_i.

This leads to the least squares estimate $\hat{\eta} = \text{argmin}_\eta SS(\eta)$, where

$$SS(\eta) = \sum_{i=1}^{p} \sum_{c=1}^{C-1} (T_{ci} - m_c^i(\eta))^2$$

and $m_c^i(\eta) = \mathbb{E}(T_{ci}|\eta)$. Note that although the model allows for random variation in the β_{rc}^i for different units i, that variation is only captured by this first-moment method through Z_i. Moreover, if there is no covariate (or, equivalently, $Z_i = 0$ for all units i), the estimator is exactly equal (if uniquely determined by the optimization problem) to a constrained *linear* least squares estimator, i.e., a Goodman least squares estimator constrained so that $\mathfrak{B}_{rc} \in (0,1)$ and $\sum_{c=1}^{C} \mathfrak{B}_{rc} = 1$. In effect, in that case we have

$$SS(\eta) = \sum_{i=1}^{p} \sum_{c=1}^{C-1} \left(T_{ci} - \sum_{r=1}^{R} \mathfrak{B}_{rc} X_{ri} \right)^2 ,$$

where $\mathfrak{B}_{rc} = \frac{\exp(\gamma_{rc})}{1+\sum_{j=1}^{C} \exp(\gamma_{rj})}$, but the function $f(\eta) = (\mathfrak{B}_{rc})_{r,c=1,1}^{R,C}$ is a bijection between $\mathbb{R}^{R\times(C-1)}$, the space of possible values of η, and the set

$$\mathcal{B} = \left\{ \mathfrak{B} \in (0,1)^{R\times(C-1)} \mid 1 - \sum_{c=1}^{C-1} \mathfrak{B}_{rc} \in (0,1) \text{ for } r = 1, \ldots, R \right\},$$

the space of possible values of $\mathfrak{B}$. Indeed, a trivial computation shows that $g : \mathcal{B} \to \mathbb{R}^{R\times(C-1)}$ given by $g(\mathfrak{B}) = \left(\log \left(\frac{\mathfrak{B}_{rc}}{1-\sum_{j=1}^{C-1} \mathfrak{B}_{rj}} \right) \right)_{r,c=1,1}^{R,C}$ is the inverse of f. This lets us formulate the nonlinear least squares estimator when no covariate is present as a solution to the following quadratic programming problem:

$$\text{minimize} \sum_{i=1}^{p} \sum_{c=1}^{C-1} \left(T_{ci} - \sum_{r=1}^{R} \mathfrak{B}_{rc} X_{ir} \right)^2$$

$$\text{subject to } \mathfrak{B}_{rc} \in (0,1), 1 - \sum_{c=1}^{C-1} \mathfrak{B}_{rc} \in (0,1).$$

There are efficient algorithms to solve this problem, and are preferable to general iterative optimization methods for numerical reasons (e.g. methods like L-BFGS are sensitive to the starting point).[5]

[5] I implemented this algorithm in R. Please request the code if you are interested.

3 Models for spatial variation

The statistical methods described in the previous section attempt to model the variation in $\mathbb{E}(\beta^i|X_i)$ parametrically in terms of measured covariates or nonlinear transformations of X_i. In this section I will propose an alternative strategy. If we assume that the link between X_i and β^i_{rc} is a consequence of common determinants that are a function of spatial location (local politics, sociodemographic variables) then we can break that link conditioning by the geographic position of the units i. In fact, as I argued for the case of Argentine electoral precincts, the constancy assumption holds if we maintain the location of voters constant. This justifies obtaining estimators for voter transition rates for each location using Gooodman's or the MD model, and then aggregating them to estimate the population parameters. However, this has two obvious limitations. First, there may not be enough observations at each location to obtain reliable estimates, or even an estimate at all in the (reasonable) case of less than R observations. Second, it prevents the model from "borrowing strength" across space. Now, under the assumption that the variables that determine and link X_i and β^i_{rc} are in some sense a *smooth* function of the geographic coordinates of the unit i, i.e., that units that are contiguous or close behave similarly, we can devise semiparametric models that admit variation in $\mathbb{E}(\beta^i_{rc}|X_i)$ but do not require a specific functional form for it nor measurements for the relevant covariates.

That is the strategy that I will follow in the rest of the article. I will describe three models in this section that implement these ideas.

Regularization. Suppose the geographic space is divided into regions $\mathcal{R}_1, \ldots, \mathcal{R}_S$ that form a partition of the set of units: $\{1, \ldots, p\} = \bigsqcup_{s=1}^{S} \mathcal{R}_s$. In the case of Argentine elections, the regions are electoral precincts (*circuitos*), and the units are voting booths (*mesas*). We model vote transitions in each region as separate MD models without covariates. Let γ_s be the parameter γ of the MD model for region s. We assume that the total variation between the parameters γ_s for regions that are contiguous or close is bounded. Concretely, we define

$$TV(\gamma) = \sum_{s,t=1}^{S} w_{st} \|\gamma_s - \gamma_t\|^2,$$

where $\|\cdot\|$ is the Frobenius norm, i.e., $\|\gamma\|^2 = \sum_{r=1}^{R}\sum_{c=1}^{C}\gamma_{rc}^2$, and the weights w_{st} measure how much we penalize the difference between regions s and t. We can take the weights to be a decreasing function of the distance between regions (for example $w_{st} = e^{-d_{st}^2/h^2}$, where d_{st} measures the distance between s and t, and h is the bandwidth), or to comprise the adjacency matrix of a graph where the vertices are the regions and the edges connect contiguous regions, or a region to its K nearest regions. The *spatial smoothness* assumption is that $TV(\gamma) \leqq C$, where C is a positive number, a hyper-parameter.

We can provide a nonlinear least squares estimator for the quantities γ_{src} by minimizing $SS(\gamma) = \frac{1}{p}\sum_{i=1}^{p}\sum_{c=1}^{C-1}(T_{ic} - m_c^i(\gamma))^2$ under the constraint $TV(\gamma) \leqq C$, where

$$m_c^i(\gamma) = \mathbb{E}(T_{ic}|\gamma) = \sum_{r=1}^{R}\frac{\exp(\gamma_{s_i rc})}{1+\sum_{j=1}^{C}\exp(\gamma_{s_i rj})}X_{ir}$$

and s_i is unit i's region, so that $i \in \mathcal{R}_{s_i}$. In order to enforce the smoothness constraint $TV(\gamma) \leqq C$, we minimize $SS(\gamma) + \lambda TV(\gamma)$ for a certain λ that depends on C and the data. Since choosing C and choosing λ is equivalent[6], we will focus on λ. We can see $TV(\gamma)$ as a *regularization term*. In order to choose the best λ we can use K-fold cross-validation. To that end, we subdivide the observations located at each region s into K equally sized samples, and we then run the model K times, each time leaving aside one of these samples, obtaining estimators $\hat{\gamma}_1, \ldots, \hat{\gamma}_K$. The cross-validation error is then computed as the average of $SS_k(\hat{\gamma}_k)$, for $k = 1, \ldots, K$, where SS_k is computed over the left-out sample for iteration k (i.e., the observations that were not used for obtaining the estimator $\hat{\gamma}_k$).

I implemented this method in Python using TensorFlow, an open source software library for machine learning[7], for fast and scalable computation. I use the L-BFGS algorithm to solve the optimization problem.

Discussion. In this approach we trade bias for variance. We assume (conservatively) that in each region Goodman's estimator is unbiased, although it may have high variance. In order to reduce it, we do two things. First, we employ instead the nonlinear least squares estimator for the MD model, which guarantees that the parameter estimates lie inside their plausible

[6] If $\hat{\gamma}$ minimizes $SS(\gamma) + \lambda TV(\gamma)$, $\hat{\gamma}$ also minimizes $SS(\gamma)$ under the constraint $TV(\gamma) \leqq C$, for $C = TV(\hat{\gamma})$. Conversely, varying λ in the first problem we cover all instances of the second problem for all C.

[7] See tensorflow.org for details.

domain. Second, we bound the total spatial variation, thus enforcing a smoothness condition and letting the estimation in region s to borrow strength from its connected regions' data. The hyper-parameter λ controls how much spatial structure we impose on the local parameters: if $\lambda = 0$, no spatial structure is imposed, and separate MD models are fit for each region; if $\lambda \to +\infty$, spatial variation is totally constrained, and the model is equivalent to a single global MD model; for intermediate values of λ we get a compromise.

To define the measure of spatial variation I take inspiration from the literature on machine learning over graphs (Herbster, Pontil and Wainer, 2005). The resulting strategy is extremely flexible, and can be used to model nonparametrically the effects of location in more general definitions of space. Thus, the "spatial coordinates" may be given, for example, by sociodemographic variables or by the configuration of local political coalitions. All that is needed is a graph (V, E), where V is a partition of the set of units (such that inside each $\mathcal{R} \in V$ the constancy assumption holds) and E is a set of connections of those subdivisions in "space". Optionally, we can incorporate a measure of the "distance" between those regions in this generalized notion of space.

Geographically weighted least squares. An alternative to the preceding approach is to obtain estimates for the local parameters γ_s in region s by weighting the importance of data from nearby regions according to their distance to s. Thus we obtain $\hat{\gamma}_s$ by minimizing

$$\sum_{t=1}^{S} \sum_{i \in \mathcal{R}_t} \sum_{c=1}^{C-1} w_{st} \left(T_{ic} - \sum_{r=1}^{R} \frac{\exp(\gamma_{src})}{1 + \sum_{j=1}^{C} \exp(\gamma_{srj})} X_{ir} \right)^2 ,$$

where w_{st} is a decreasing function of distance between regions s and t. This is a nonlinear least squares version of Geographically Weighted Regression (Fotheringham, Brunsdon and Charlton, 2002).

Multilevel Bayesian model. We can model the spatial variation in the MD model as follows:

$$\begin{aligned}
(T_{i1}N_i, \dots, T_{iC}N_i) &\sim \text{Multinomial}\left(N_i, \sum_{r=1}^{R} \beta^i_{r1} X_{ir}, \dots, \sum_{r=1}^{R} \beta^i_{rC} X_{ir}\right) \\
(\beta^i_{r1}, \dots, \beta^i_{rC}) &\sim \text{Dirichlet}\left(d_r \exp(\gamma_{s_i r 1}), \dots, d_r \exp(\gamma_{s_i, r, C-1}), d_r\right) \\
\gamma_{\bullet rc} &\sim \text{MultivariateNormal}(\mu_{rc}, \tau^2_{rc}\Sigma) \\
d_r &\sim \text{Exponential}(\lambda)
\end{aligned}$$

Here s_i is the region where unit i belongs (i.e., $i \in \mathcal{R}_{s_i}$), $\gamma_{\bullet rc}$ is the vector $(\gamma_{1rc}, \dots, \gamma_{Src})$, and $\Sigma \in \mathbb{R}^{S \times S}$ is a symmetric positive definite matrix that specifies the structure of spatial correlation. We can take, for example, $\Sigma_{st} = e^{-d^2_{st}/h^2}$, where d_{st} measures the distance between regions s and t, and h is the bandwidth.

In the next section I will show an application of the first model. Since this is still work in progress, I have not tested models 2 and 3 on real or synthetic data sets yet.

4 Application: 2015 Argentine elections

In 2015, Argentina held a three-stage presidential election leading to the victory of Mauricio Macri from the *Cambiemos* (Let's Change) alliance, integrated by the parties PRO, UCR and CC-ARI. Cristina Fernández de Kirchner, the incumbent president, had governed for two terms and was prevented from running again. Her coalition, the *kirchnerist* FPV (Victory Front), lost the race after 12 years in power; the candidate was Daniel Scioli, the Province of Buenos Aires governor. The election was momentous for various reasons. The center-left segment of the Peronist party, the FPV, lost the presidency to a pro-business coalition (Cambiemos) which was "programmatically in its antipodes"; moreover, it was the first time a programmatically centre-right party won democratic elections (Freytes and Niedzwiecki, 2016). It was also the first time a democratically elected president since World War Two was neither Radical (from the UCR party) nor Peronist: indeed, "Macri's PRO is arguably the first Argentine political party in more than sixty years to establish a true national presence" (Alles, Jones, Tchintian, 2016).

The first stage of the election was a single-day all party/alliance federal primary that was mandatory for political parties and alliances, and compulsory for voters (enforced with a small fine). 15 candidates participated and 3 alliances held competitive primaries (between more than one candidate). Six candidates won their primaries and got the mandatory 1.5% of the vote necessary to pass to the next stage. Since no candidate in the second stage (the general election) got 45% of the vote nor 40% and a margin of 10% over the second place, a runoff was held (for the first time since it was included in the 1994 Constitution). In this third stage, Mauricio Macri was finally able to muster a majority of the (valid) votes, winning the election to Daniel Scioli (FPV) by 51.34%–48.66%.

Table 1: Presidential Vote Shares in Argentina's 2015 Primary and General Elections

	Primary Vote Share (%)	First-Round Vote Share (%)	Runoff Vote Share (%)
Scioli	38.7	37.1	48.7
Macri	24.5	34.2	51.3
Sanz	3.3	–	–
Carrió	2.3	–	–
Massa	14.3	21.4	–
De la Sota	6.3	–	–
Stolbizer	3.5	2.5	–
Del Caño	1.7	3.2	–
Altamira	1.6	–	–
Rodriguez Saá	2.1	1.6	–
Others	1.8	–	–

Note: Turnout for the primary, first round, and second round was 74.9%, 81.2%, and 80.9% respectively.
Source: Lupu (2016), from Dirección Nacional Electoral.

The first transition: from the primaries to the first round. Macri's vote majority was assembled incrementally in this three-stage process. In the first transition, from the first to the second stage, Macri not only retained his votes and a majority of the votes received by his coalition allies (Sanz and Carrió) but also almost a half of De la Sota's and Stolbizer's votes (see Table 2). De la Sota, the Province of Cordoba governor and a center-right Peronist distanced from the *kirchnerist* government, competed against Sergio Massa in the UNA primary and lost.

Stolbizer, an ex-UCR congresswoman, ran a center-left campaign on "honesty" which resonated on middle-class voters who were perceptive to allegations of widespread corruption affecting the FPV government. Massa increased his vote share, retaining his vote and almost half of his primary competitor's vote, and also benefiting from the higher turnout. Scioli did not see an increase in his share of the valid votes (see Table 1).

Table 2: Voter transition rates from the Primary to the First-Round

	Scioli	Macri	Massa	Del Caño	Stolbizer	Rodriguez Saá	No Candidate
Scioli	0.87	–	0.04	–	–	–	0.09
Macri	–	0.96	–	–	–	–	0.04
Sanz	0.01	0.66	–	–	–	–	0.32
Carrió	–	0.55	–	–	0.07	–	0.37
Massa	–	–	0.91	0.03	–	–	0.05
De la Sota	0.01	0.46	0.39	–	–	–	0.13
Stolbizer	–	0.40	–	0.02	0.31	–	0.27
Del Caño	–	0.35	–	0.64	–	–	0.01
Altamira	–	0.27	0.02	0.37	0.13	–	0.21
Rodriguez Saá	–	0.08	0.03	–	–	0.81	0.08
Others	0.13	–	0.24	0.10	0.09	–	0.43
No Candidate	0.18	0.13	0.15	0.03	0.03	–	0.48

Zero-valued cells were omitted.

In order to estimate the voter transition rates in Table 2, I computed the nonlinear least squares estimator for the MD model. I used the quadratic programming formulation of the problem (Section 2), which was convenient for its numerical robustness and its time performance. Running the Bayesian MD model (Section 2) or fitting a more complex model from Section 3 over the full data set (91,719 voting booths) and such a big table (12×7) is problematic for performance and numerical limitations of current implementations (Lau, Moore and Kellermann, 2007 for the Bayesian MD model, and mine for the first model in Section 3).

The second transition: from the first round to the runoff. The position of Sergio Massa in the first round was pivotal (third place with 21.4% of the vote), but he did not endorse any of the two candidates for strategic reasons (Murillo, Rubio and Mangonnet, 2016). The main question, therefore, is how and why the Massa's votes in the first round were transferred to Scioli and Macri at the runoff. I will first argue, based on the literature, that spatial effects are crucial. Second, I will run the model described in Section 3 to estimate the voter transition rates in a way that admits spatial variation.

The presidential election was not defined in programmatic terms, i.e., by voters' ideological preferences and candidates' placement along the left-right dimension. In effect, as Calvo and Escolar (2016) show, voters placed themselves and the runoff candidates' parties (FPV and PRO) at the center of the ideological scale. Although differing programmatically (Freytes and Niedzwiecki, 2016), Scioli and Macri placed themselves in the center during the campaign; e.g. Scioli promised to gradually lift capital controls implemented by Cristina Fernández de Kirchner and Macri vowed to maintain popular social policies such as conditional cash transfers. As Lupu (2016) shows, this was successful: "when the Argentine Panel Election Study (APES), a national public-opinion survey, asked respondents to place Macri's PRO on a left-to-right ideological scale from 0 to 10, they put the party, on average, at 5.7, just right of center. They placed Scioli's FPV at 5.3, indistinguishably further to the left." And, as he says later, "voters relied on the classic 'valence issue' of incumbent performance in making what they viewed as a choice between continuity and change" (p. 42).

Voters evaluated incumbent performance mainly in economic terms, as Lupu showed (2016). And, as Freytes and Niedzwiecki (2016) argue, this, plus mounting economic challenges (sluggish economic growth, high inflation, foreign currency shortage), "deepened territorial cleavages between the agricultural central region and the peripheral provinces" (p. 3). The voter coalition that supported the FPV was based on the urban poor and the peripheral, less-developed and fiscally dependent provinces, the "low-maintenance peripheral coalition" (Gibson and Calvo, 2001). This coalition was maintained by redistributive policies paid for by export taxes on agricultural rents. "As the economy stagnated, the electorate in the central provinces perceived the redistribution of agricultural rents to the periphery as a zero-sum game that subtracted regional wealth without commensurable gains" (Freytes and Niedzwiecki, 2016, p. 8).

The impact of these territorial cleavages can be expected to produce strong spatial effects on the vote of a pivotal candidate such as Massa, who sought an intermediate position between "continuity" and "change". Therefore, in order to estimate the voter transition rates I ran the first model outlined in Section 3 over the election data. In the rest of the section I will describe the details of the implementation and the results.

Implementation details. Argentina is divided into provinces (including the City of Buenos Aires, counted here as a province); provinces are divided into *departamentos* (municipalities, except for the City of Buenos Aires, where they are the *comunas*); and *departamentos* are divided into *circuitos* (electoral precincts). There are 96339 ballot boxes distributed into 13883 polling stations (schools) from 5703 precincts. There is valid data only for 91671 ballot boxes[8]. Since I need multiple observations per precinct in order to perform stratified cross-validation, I leave out precincts with less than 10 ballot boxes, resulting in 82190 units ($\sim$ 90% of the total)[9]. I use polling stations' geolocation provided by La Nación Data's team. I use it to compute the distances d_{st} between precincts s and t defined as the minimum distance between polling stations u and v from s and t, respectively.

In order to construct the graph that models the spatial relations between precincts, I connect precinct s with its K nearest precincts, where $K \in \mathbb{N}$ is a parameter that I will choose later. I then define the weights w_{st} for precincts s and t as $e^{-d_{st}^2/h^2}$ if s and t are adjacent, and 0 otherwise; h, the bandwidth, is another parameter. I choose this mix between nearest neighbours and gaussian decay because of the huge heterogeneity of population density across the country. Polling stations in sparsely populated rural areas require a big bandwidth in order to borrow strength from near locations' data. In contrast, metropolitan areas require a smaller bandwidth in order for the model to capture variation within them. A global bandwidth is thus not appropriate. Penalizing variation just for the K nearest neighbours introduces the necessary compromise, and is less costly than finding an optimal adaptive bandwidth for each location. I ran the model for various values of K and h, under the constraint of the resulting graph being connected. I found that the model is robust to the specification of K and h. I chose $K = 25$ and $h = \infty$ (i.e., $w_{st} = 1$ if t is among the nearest 25 precincts to s, and 0 otherwise).

As Figure 1 shows, when we let λ grow, and thus penalization for spatial variation increase, the in-sample error grows (dashed line). This is expected: as we enforce stricter constraints, the model is less able to fit the data. However, the cross-validation (CV) error, which is an estimate

[8] The data corresponds to the *escrutinio provisorio* (preliminary counting) and is provided by the Dirección Nacional Electoral (DINE).

[9] Eliminating 10% of the data may bias the estimation. Indeed, I am leaving out not a random sample, but data from peripheral or sparsely populated precincts. Inferences will still be valid for the rest of the population.

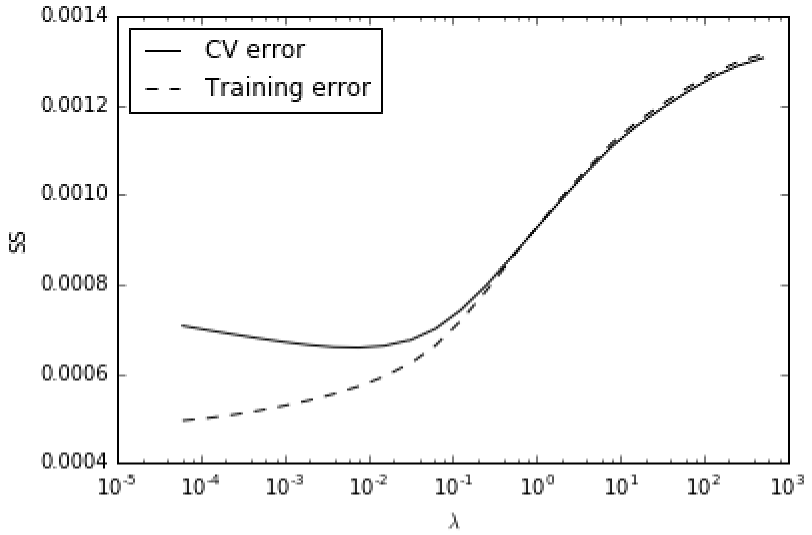

Figure 1: 5-fold Cross-Validation errors for $\lambda = 2^k$, $k = -14, \ldots, 9$.

of the out-of-sample error, decreases until it reaches $\lambda = 2^{-7}$, and then it starts growing. That is because when λ is too small the model *overfits* the data: it is able to represent very well the sample, but is incapable of distinguishing between signal and noise. When $\lambda = 0$ the model estimates the local parameters using only the data from voting booths within each precinct, ignoring the spatial structure. When $\lambda \to +\infty$ the model assumes that the parameters are constant across precincts, and we get the nonlinear least squares estimator for the MD model. The fact that the optimal λ (in terms of out-of-sample error) is between these two extremes provides evidence in favor of the model's assumptions.

Results. Table 3 shows the estimates for the global voter transition rates. Consistent with estimates from panel data (Lupu, 2016), a majority of Massa's votes went to Macri, defining the election. Are these estimates reliable? In order to partially answer that question, I calculated stratified Bootstrap confidence intervals. The results are remarkably precise. For example, the share of Massa's vote that went to Macri at the runoff varied in the interval $[0.522, 0.537]$; in the case of Stolbizer's vote, the interval was $[0.540, 0.589]$.

A by-product of modeling the spatial variation is that we get estimates for voter transition rates at each location. In Figure 2 I plot Massa-to-Macri transition rates at each polling station (dark red means a higher-than-average transition rate; soft yellow means the contrary, i.e., a high transition rate to Scioli). It can be seen that in the agricultural central region a majority of Massa's voters chose Macri at the runoff, while in the peripheral provinces this was reversed. A

Table 3: Voter transition rates from the First Round to the Runoff

	Scioli	Macri	No Candidate
Scioli	0.90	0	0.10
Macri	0	0.91	0.09
Massa	0.33	0.53	0.14
Del Caño	0.44	0.34	0.22
Stolbizer	0.21	0.56	0.23
Rodriguez Saa	0.35	0.52	0.13
No Candidate	0.22	0.22	0.56

notable exception is Jujuy Province (at the north-west extreme), where there is a concentration of Massa's voters that preferred Macri by majority at the runoff. This can be explained by the political coalition that Massa forged with Cambiemos to sustain the candidacy of Gerardo Morales (UCR) for governor. Thus, the structural territorial cleavages discussed earlier are not the sole determinants of spatial variation in the transition rates. Local coalitional dynamics also shape electoral behavior. Another example of this phenomenon is La Rioja Province, where Massa also forged a coalition with the UCR's candidate for governor.

Finally, in Figure 3 I plot Massa-to-Macri transition rates in the Buenos Aires Metropolitan Area. These transition rates are higher in CABA (at the center) than in the Province of Buenos Aires districts that surround it, except the affluent municipalities at its north. In the poorest locations, the share of Massa's voters who voted for Macri is among the lowest in the country (~ 0.25). We can see that there is a huge variation at this level, even across contiguous precincts. This highlights the class cleavage inside urban areas and its spatial pattern. It also shows that the model is able to capture this fine-grained variation.

5 Conclusion

The statistical models reviewed in Section 2 assume in their basic form the constancy assumption, i.e., that transition rates β^i_{rc} are mean-independent with respect to the first election results X_i. This assumption can only be relaxed if the dependence between β^i_{rc} and X_i disappears modeling β^i_{rc} explicitly in terms of a covariate Z_i (which can be a nonlinear transformation of X_i). In the case of Argentina's elections, voters are randomly assigned to voting booths inside their

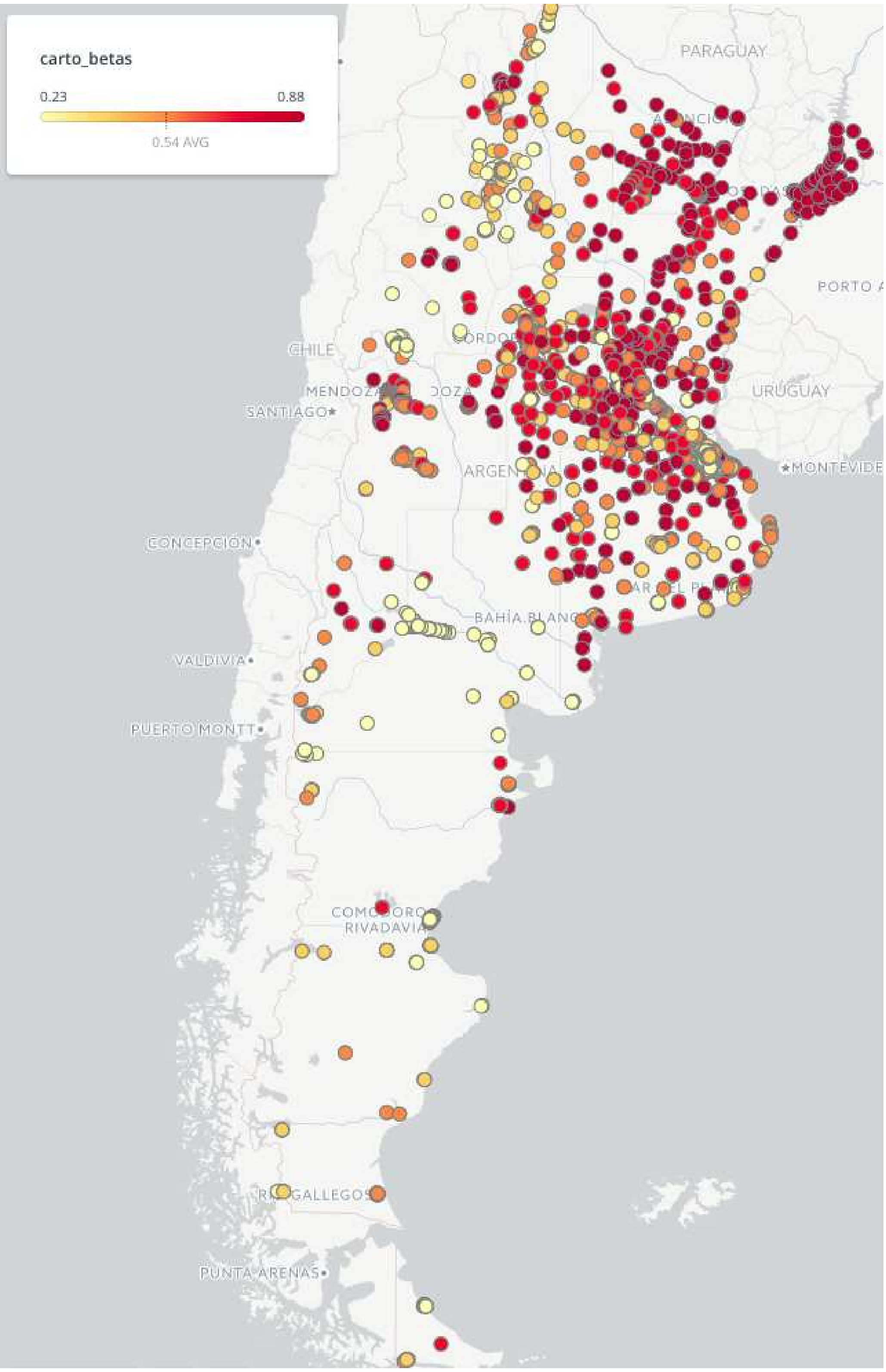

Figure 2: Share of Massa's voters in the First-Round that voted for Macri at the Runoff.

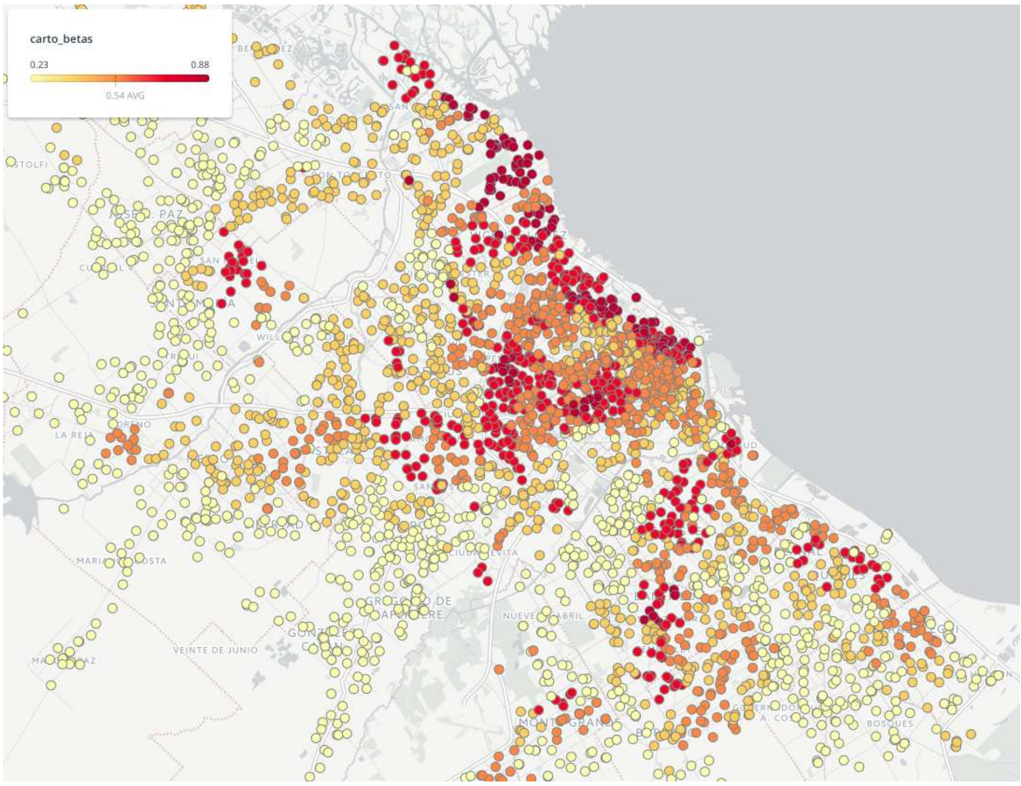

Figure 3: Share of Massa's voters in the First-Round that voted for Macri at the Runoff. Buenos Aires Metropolitan Area.

electoral precinct, which implies that the constancy assumption holds—at the precinct level. Therefore, the basic statistical models' estimators are unbiased inside each precinct, although their variance may be as high as to make them unreliable. Now, assuming some form of spatial smoothness in the parameters' variation leads to a promising strategy: to look for estimators at each precinct but borrowing strength from observations located sufficiently close (how close can vary with the population density of the location). In Section 3 I presented three statistical models that implement this strategy, and applied one of them to the problem of estimating voter transition rates in the Argentine 2015 presidential election.

The results are promising. The model outperforms the MD nonlinear least squares estimator in terms of out-of-sample error, achieving a balance between total spatial homogeneity and total spatial heterogeneity (i.e., when each district is modeled as a separate spatial regime). Moreover, the model provides estimates of the voter transition rates at the precinct level, which can be interpreted geographically. The variation across space of these quantities provides evidence of three distinct patterns that lead to interesting substantive hypotheses. First, the structural territorial cleavage between the agricultural center (inclined for Macri) and the peripheral provinces (with a strong linkage to Peronism) had a significant impact on the destination of Massa's votes at the runoff. Second, in the peripheral provinces in which Massa formed a coalition with the UCR the majority of his votes benefited Macri. Third, in urban areas the class cleavage had a strong effect on the transition of Massa's vote. Incorporating sociodemographic and local-coalition variables into the model is a future avenue of work that should provide more evidence for these hypotheses.

References

Alles, S., Jones, M.P. and Tchintian, C., 2016. "The 2015 Argentine presidential and legislative elections." *Electoral Studies.*

Anselin, L. and Cho, W.K.T., 2002. "Spatial effects and ecological inference." *Political Analysis*, 10(3), pp.276-297.

Calvo, E. and Escolar, M., 2003. "The local voter: A geographically weighted approach to ecological inference." *American Journal of Political Science*, 47(1), pp.189-204.

Calvo, E. and Escolar, M., 2016. "La grieta es un espejismo." *El Estadista*, http://elestadista.com.ar/?p=10512

Cho, W.K.T. and Judge, G.G., 2008. "Recovering vote choice from partial incomplete data." *Journal of Data Science*, 6(2), pp.155-171.

Cho, W.K.T. and Manski, C.F., 2008. "Cross level/ecological inference." *Oxford handbook of political methodology*, pp.530-569.

Dirección Nacional Electoral, http://www.elecciones.gob.ar/

Duncan, O.D. and Davis, B., 1953. "An alternative to ecological correlation." *American sociological review*, 18(6), pp.665-666.

Fotheringham, A.S., Brunsdon, C. and Charlton, M., 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* Wiley.

Freytes, C. and Niedzwiecki, S., 2016. "A turning point in Argentine politics: demands for change and territorial cleavages in the 2015 presidential election." *Regional & Federal Studies*, pp.1-14.

Gibson, E.L. and Calvo, E., 2001. "Federalism and low-maintenance constituencies: Territorial dimensions of economic reform in Argentina." *Studies in Comparative International Development*, 35(3), pp.32-55.

Goodman, L.A., 1953. "Ecological regressions and behavior of individuals." *American sociological review.* 18:663–664.

Goodman, L.A., 1959. "Some alternatives to ecological correlation." *American Journal of Sociology*, pp.610-625.

Greene, W. H., 2011. *Econometric Analysis* (7 ed.). Pearson.

Haneuse, S. and Wakefield, J., 2004. Ecological inference incorporating spatial dependence. In G. King, M. Tanner, and O. Rosen (Eds.), *Ecological Inference: New Methodological Strategies*, Chapter 12, pp. 266–300. Cambridge University Press.

Herbster, M., Pontil, M. and Wainer, L., 2005. "Online learning over graphs." In *Proceedings of the 22nd international conference on Machine learning* (pp. 305-312). ACM.

King, G., 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton, NJ: Princeton University Press.

King, G., Rosen, O. and Tanner, M.A., 1999. "Binomial-beta hierarchical models for ecological inference." *Sociological Methods & Research*, 28(1), pp.61-90.

King, G., Tanner, M.A. and Rosen, O. eds., 2004. *Ecological inference: New methodological strategies.* Cambridge University Press.

Klima, A., Thurner, P.W., Molnar, C., Schlesinger, T. and Küchenhoff, H., 2015. "Estimation of voter transitions based on ecological inference: an empirical assessment of different approaches." *AStA Advances in Statistical Analysis*, 100(2), pp.133-159.

Lau, O., Moore, R.T., Kellermann, M., 2007. "eiPack: RxC ecological inference and higher-dimension data management." *R News* 7(2), 43–47

Lupu, N., 2016. "The End of the Kirchner Era." *Journal of Democracy*, 27(2), pp.35-49.

Murillo, M.V., Rubio, J.M. and Mangonnet, J., 2016. "Argentina: Voters' Influence and Electoral Alternation." *Revista de Ciencia Política*, 36(1), pp.3-26.

Rosen, O., Jiang, W., King, G. and Tanner, M.A., 2001. "Bayesian and frequentist inference for ecological inference: The R×C case." *Statistica Neerlandica*, 55(2), pp.134-156.