Romero, R.; Pavía, JM.; Martín Marín, J.; Romero, G. (2020). Assessing uncertainty of voter transitions estimated from aggregated data. Application to the 2017 French presidential election. Journal of Applied Statistics. 47(13-15):2711-2736. https://doi.org/10.1080/02664763.2020.1804842

Assessing uncertainty of voter transitions estimated from aggregated data. Application to the 2017 French presidential elections

Romero, Rafael
Universidad Politécnica de Valencia, email: rromero@eio.upv.es
Pavía, Jose M.
GIPEyOP, Universitat de Valencia, email: pavia@uv.es
Martín, Jorge
Universidad Politécnica de Valencia, email: jmartinm@eio.upv.es
Romero, Gerardo
EST2 Consultoría, Innovación y Desarrollo S.L., email: gerardo@est2.es

Abstract: Inferring electoral individual behaviour from aggregated data is a very active research area, with ramifications in sociology and political science. In this paper, a new approach based on linear programming is proposed to estimate voter transitions among parties (or candidates) between two elections. In contrast to other similar models previously suggested in the literature, our approach presents two important innovations. First, it explicitly deals with new entries and exits in the election census without assuming unrealistic hypotheses, enabling a reasonable estimation of vote transfers for young electors voting for the first time. We illustrate this in a real instance. Secondly, by exploiting the information contained in the model residuals, we develop a procedure to assess the level of uncertainty in the estimates. This significantly distinguishes our model from other linear and quadratic programming methods previously published. The method is illustrated estimating the vote transfer matrix between the first and second round of the 2017 French Presidential election and measuring its level of uncertainty. Likewise, compared to the most current alternatives based on ecological regression, our approach is considerably simpler and faster, and has provided reasonable results in all the actual elections to which it has been applied. Interested scholars can easily use our procedure with the aid of the *R*-function described at the bottom of this paper.
Keywords: Ecological Inference; Linear Programming; Voter transitions; R × C contingency tables; French elections.

Acknowledgements: The authors wish to thank the special issue editor and a reviewer for their valuable comments and suggestions. Thanks are also due to M. Hodkinson for revising the English of the paper. This piece of research has been supported by the Spanish Ministry of Science, Innovation and Universities and the Spanish Agency of Research, co-funded with FEDER funds, grant ECO2017-87245-R, and by Consellería d'Innovació, Universitats, Ciència i Societat Digital, Generalitat Valenciana, grant AICO/2019/053.

Assessing uncertainty of voter transitions estimated from aggregated data. Application to the 2017 French presidential elections

1. Introduction

The analysis of voter transitions that occur between two elections from a set of parties (or candidates) to another is a relevant studying topic of political sociology. Disposing of appropriate estimates is relevant for many agents, including the media, political scientists and party teams (Abou-Chadi and Stoetzer, 2020). Hence, for decades the issue has attracted the interest of many authors who have tried to exploit survey data and/or aggregate election results to produce accurate estimates (e.g., Hawkes, 1969; Miller, 1972; McCarthy and Ryan, 1977; Upon, 1978; Brown and Payne, 1986; Payne, Brown, and Hanna, 1986; Forcina and Marchetti, 1989, 2011; Füle, 1994; Vázquez and Romero, 2001; Park, 2008; van der Ploeg, 2008; Romero, 2014; Corominas, Lusa, and Calvet, 2015; Puig and Ginebra, 2015; Klima et al., 2016; Nuñez, 2016; Pavía, Bodoque, and Martín, 2016; Plescia and De Sio, 2018; Klima et al., 2019; Baydoğan, 2019; or Pavía and Aybar, 2020).

In poll-based approaches, electoral mobility is estimated using vote recall of exit-polls or post-election surveys, or via panel surveys, as an aggregation of individual vote displacements. This strategy, however, presents serious concerns that question both their efficiency and effectiveness. Firstly, issues emerge of complexity and of sample size. Large, complex samples are required to reach reasonably accurate rate estimates given that, from a statistical point of view, a single population is not being sampled, but as many populations as election options are contemplated in the first election. Secondly, even more disturbing is the challenge posed by nonresponse bias and measurement error. On the one hand, nonresponse is not randomly distributed among political options. The individual probability of nonresponse depends on the context, on the voter's own vote, and even on the propensity to change it (Pavía, Badal and García-Cárceles, 2016). On the other hand, retrospective answers are not very reliable. When asked about their past electoral behaviour, electors frequently cannot recall their vote or are concerned with social desirability issues (Dassonneville and Hooghe, 2017). Hence, combined, both issues raise doubts about the variance and bias of poll-based vote transfer estimates.

Indeed, many authors have reported actual cases in which the bias induced by measurement error and nonresponse can lead to results that are far from the reality. As an example, we can compare actual data and raw answers collected in a survey conducted in November 2014 by the most prestigious Spanish survey organization (Centro de Investigaciones Sociológicas). In that survey, just 28% of the respondents claimed to have voted for the Conservative Party (PP) in the 2011 Spanish General election (CIS, 2014), when actually 45% of voters supported PP in that election. Thus, it is not surprising Miller (1972, p. 122) claims that: "Surveys dealing with voting change are especially unreliable". As a consequence of the above limitations, authors who have studied this issue in depth conclude that, in these types of surveys, imprecision and bias can be large and represent obstacles difficult to overcome (Klima et al., 2016).

The existence of these flaws has motivated several authors to try to estimate voter transitions using either statistical or mathematical algorithms that exploit recorded aggregate official results, which are certainly more reliable. All these methods, examples of the so-called ecological inference procedures, can be grouped into two main sets: ecological regression methods and mathematical programming procedures. The ecological regression literature has been more prolific, producing a larger number of proposals undoubtedly fuelled by the US legal ramifications related to the electoral redistricting processes (King, Rosen, and Tanner, 2004) and by their use in epidemiology (e.g., Fisher and Wakefield, 2020). Electoral studies and epidemiology are not, however, the only disciplines where

these approaches can be used. They are useful in many situations where the goal is to infer individual-level behaviour from aggregate data (e.g., Caughey and Wang, 2019). In this paper, we propose a new ecological inference approach to estimate voter transitions, but based on linear programming. Compared to other procedures, our method presents two important innovations. On the one hand, it explicitly considers new entries and exits in the election census. On the other hand, by exploiting the information contained in the model residuals, it proposes a procedure to assess the level of uncertainty in the estimates. This second innovation is the main contribution of our paper. No other previously published method based on linear or quadratic programming (e.g., McCarthy and Ryan, 1977; Tziafetas, 1986; Vázquez and Romero, 2001; Romero, 2014; Corominas et al., 2015) measures this issue. Uncertainty is routinely estimated in ecological regression approaches.

The rest of the paper is organised as follows. Section 2 briefly reviews the ecological inference literature. In Section 3, we present the model, introduce the mathematical conditions that must be fulfilled and state conditions under which the hypothesis of electoral homogeneity on which the model rests can be applied. In this section, we also address the problem of census changes, suggesting a personal solution that allows the estimation of new electors' votes. This is illustrated with an actual example in Section 4. Section 5 is devoted to analysing the uncertainty associated with the model estimates, which is case-dependent on both data and model structure. We propose an original procedure to estimate it. The procedure is based on quantifying, by simulation, the relationship between true error rates and the degree of non-compliance of the homogeneity hypothesis. In actual applications, we can estimate the latter from the model's residuals. In Section 6, we illustrate our procedure by estimating uncertainty in voter transitions between the first and second round of 2017 French presidential elections. Section 7 summarizes the conclusions obtained and suggests directions for further research. An *R* function to apply the methodology is described in an Appendix and its code provided in the Online Supplemental Materials.

2. Ecological inference methods. A brief revision of the literature

Polls are not always available (e.g., in local elections) and, when available, they are, as stated in the previous section, exposed to significant sources of bias. They also give rise to voter transition estimates with large variances. Hence, as an alternative, many methods just rely on recorded official outcomes. In this case, the basic strategy to reach estimates consists of applying a statistical or mathematical procedure to the results tailed in a set of territorial units. The drawback of this approach, which is exposed to the presence of the so-called ecological fallacy (Robinson, 1950), lies in the fact that the underlying mathematical problem is indeterminate, since it depicts a system with more unknowns than equations. This forces the inclusion of additional hypotheses to attain a solution (Greiner and Quinn, 2009): usually the assumption that the voter transition matrices in the different units are, in some sense, "similar". Two basic strategies have been followed in the literature. One based on ecological regression and another grounded on mathematical programming.

The ecological regression literature has been very fertile since the seminal papers of Duncan and Davis (1953) and Goodman (1953, 1959) and has experienced a resurgence since King (1997), despite the criticisms of Freedman et al. (1998) and Cho (1998). Indeed, King (1997) and notably King et al. (2004) represent a tipping point in this literature, with some of the key references including King, Rosen, and Tanner (1999), Rosen et al. (2001), Wakefield (2004), Greiner and Quinn (2010), Glynn and Wakefield (2010), Puig and Ginebra (2015), Plescia and De Sio (2018), Klima et al. (2019) and Forcina and Pellegrino (2019). Klima et al. (2016) discuss some of the main methods developed under the ecological regression framework and show some of the difficulties that these kinds of procedures entail. The principal problem with the most current approaches, apart from their high computational demand, lies in their complexity; mainly for those methods relying on the Bayesian framework. They

require the intervention of high skilled experts to properly specify the setting parameters and hypothesis for the distributions of the quantities to be estimated on which they rest heavily on. Indeed, given specific election outcomes, the different methods can lead to sensitive different results and even a single method leads to quite different results as different values for certain operational parameters are set. The situation is worsened by the fact that the relevance of some hypotheses is sometimes difficult to establish and to gauge from the point of view of political science.

A different way followed by other authors is to approach the subject as a mathematical programming problem, looking for the values p_{jk} of the vote transition matrix that, fulfilling certain restrictions, minimize, in some sense, the discrepancy with the outcomes recorded in the different territorial units. McCarthy and Ryan (1977) propose a quadratic program model to minimize the sum of squares of these discrepancies, Tziafetas (1986) suggests minimizing the sum of their absolute values, which transforms the model into a linear program, and Corominas et al. (2015) explore four possible optimality criteria to estimate the p_{jk}. Although the mathematical programming approaches share some similarities with the ecological regression methods when the sum of squares of the discrepancies is used as loss function, these have the advantage of not requiring assuming any particular probability distribution to guarantee that the p_{jk} estimates are logically consistent. In mathematical programming, constraints are introduced in a natural way, making possible to reach distribution-free estimates.

The abovementioned proposals, however, present two main drawbacks. First, the way all of them handle census changes is questionable. Secondly, none of the mathematical programming methods incorporates a procedure to measure the levels of uncertainty of the estimates provided by the model. Despite being well known that the electoral behaviour of young electors just entitled to vote is different to experienced voters (see, e.g., Henn and Foard, 2012; Snelling, 2016), none of these methods takes this into account when estimating the electoral behaviour of young electors. Vázquez and Romero (2001) propose an initial solution to this issue, with Romero (2014, 2015, 2016) expanding and exemplifying its use in three actual election processes. In this paper, we deal with these two drawbacks. On the one hand, we deepen on the solution proposed by Romero (2014). On the other hand, as main contribution, we develop within the mathematical programming approach a procedure to estimate in actual studies the margins of uncertainty associated with the estimated transition rates.

3. The LPHOM model

3.1 The basic model

The application of the proposed methodology requires having, as in all the models referred to in the previous section, the results of the two elections in a set of I territorial units (which we shall refer to hereinafter as units) in which the overall area of the study is partitioned.

Let x_{ij} be the votes gained in unit i by the election option j ($j = 1,\dots,J$) of election 1, and let y_{ik} be the votes obtained in the same unit by the election option k ($k = 1,\dots,K$) of election 2. In both elections non-voters (abstentions), including perhaps null and blank votes, are considered as an additional election option. As is usual, we group the minor parties (or candidates) in a rest option. We discuss the issues related to census changes between the two elections later, in subsection 3.2, and in subsection 3.3 we introduce our whole model, referred to as LPHOM.

The objective is to estimate the $J \times K$ unknown values p_{jk}, defined as the proportion of voters in the overall analysed territory who having chosen option j in election 1, choose option k in election 2. According to this definition, the p_{jk} proportions must inevitably fulfil constraints (1), (2) and (3).

$$p_{jk} \geq 0 \quad for \quad j = 1, \dots, J \quad k = 1, \dots, K \tag{1}$$

$$\sum_{k=1}^{K} p_{jk} = 1 \quad for \quad j = 1, \dots, J \tag{2}$$

$$\sum_{j=1}^{J} \left(\sum_{i=1}^{I} x_{ij} \right) p_{jk} = \left(\sum_{i=1}^{I} y_{ik} \right) \quad for \quad k = 1, \dots, K \tag{3}$$

The mathematical programming models proposed in McCarthy and Ryan (1977) and Tziafetas (1986) include constraints (1) and (3). Constraints (2) are similar to those proposed in Johnston and Hay (1983), Romero (2014) and Corominas et al. (2015). The problem that arises is that the above system, having more unknowns than equations, is indeterminate, with infinite possible solutions.

At this point, denoting p_{jk}^{i} as the proportion of voters in unit i that having chosen option j in election 1 choose option k in election 2, we have that the p_{jk}^{i} proportions must exactly fulfil (4).

$$\sum_{j=1}^{J} x_{ij} p_{jk}^{i} = y_{ik} \quad for \quad i = 1, \dots, I \quad k = 1, \dots, K \tag{4}$$

Including these additional unknowns, p_{jk}^{i}, and constraints (4), however, does not solve indeterminacy: the system remains indeterminate. To solve this indeterminacy it is necessary to include some hypothesis. As in all procedures referred to in the previous section, our hypothesis is that the voter transition matrices in the different territorial units are in some sense "similar" to each other, and, therefore, similar to the global matrix.

It should be noted that the homogeneity hypothesis does not imply that the different units have voted in a similar way, but that the matrix of voter transitions between parties has been in all of them "similar" to the global average matrix. For example, in the 2017 French elections it is obvious that there are regions that relatively voted more for Macron and others that did so for Le Pen. What the hypothesis of homogeneity implies is that, for example, if at the national level the majority of those who voted for Macron in the first round also went on to do so in the second, this phenomenon of fidelity will have occurred in a similar way in all regions.

For the hypothesis of homogeneity to be reasonable, we need the study area to be electorally homogeneous in the considered elections. This means that: (i) the main options presented, on the one hand, in election 1 and, on the other hand, in election 2 have been basically the same in all the units; and, (ii) the motivations that may have influenced voters' behaviour between elections 1 and 2 have been similar throughout the whole territory analysed, i.e., that voters' motivations have not varied too much in the different units, with global trends weighting more than local trends (Pavía-Miralles, 2005). In addition, for the homogeneity hypothesis to be adequate, it is advisable that the size of the units and also the size of the election options considered not be too small.

Thus, according to the hypothesis of homogeneity, equations (4) will be fulfilled approximately if the p_{jk}^{i} proportions are replaced by the p_{jk}; an issue which is expressed through equation (5), where the error terms e_{ik} should be "small".

$$\sum_{j=1}^{J} x_{ij} p_{jk} = y_{ik} + e_{ik} \quad for \quad i = 1, \dots, I \quad k = 1, \dots, K \qquad (5)$$

The basic model, therefore, consists of obtaining the values of p_{jk} that, fulfilling constraints (1), (2), (3) and (5), verify (6).

$$minimize \sum_{i,k} |e_{ik}| \qquad (6)$$

An advantage of this basic model for sociologists and political scientists is that it easily allows the inclusion of constraints to force the result to fulfil certain conditions that the expert considers appropriate. The problem of acting like this, however, is that the results can lose in part their objective character; depending on the validity of the subjective hypotheses imposed. As an example, additional restrictions are imposed to the p_{jk}'s in the model suggested in Romero (2014). In particular, after establishing a correspondence between some of the J political options of election 1 and some of the K political options of election 2, Romero (2014) imposes two additional conditions. On the one hand, he imposes that those parties that improve their election results retain most, at least a minimum percentage w_s, of their previous voters. On the other hand, he assumes that the greater part of the votes that those parties which had a worse result in the second election comes, at least in a minimum percentage w_l, from voters who already voted for them in the first election. These hypotheses, in principle reasonable, can be included in the model by adding the corresponding constraints. We have not included them in our model because we have found in most actual studies that they are usually automatically fulfilled by the estimates.

3.2 The problem of changes in the election census

It is unrealistic to maintain the hypothesis of stationary electorates for any pair of elections separated by a period of time. Almost certainly, there will be changes in the composition of the election censuses of the different units as a consequence of entries and exits. On the one hand, there will be new electors included in the unit censuses of election 2 who did not appear in the lists of election 1. They correspond to young people, n_i, who reached the voting age between the two elections and new residents, m_i, with the right to vote coming from other places. On the other hand, some voters included in the election censuses of election 1 will have exited from the lists of the election 2. Exit voters can be divided between voters who, between both elections, moved outside the given territorial unit i, e_i, and those voters who died, d_i. Unfortunately, this detailed information is almost never available for the average analyst. Even having access to the deanonymized, detailed census lists of both elections and linking them, it is impossible to separate exit voters into emigrants and deceased. Deanonymized lists of deceased would also be required to do that.

Demographic figures broken down into (single or five-year) age groups, nevertheless, are regularly published by official statistical agencies; therefore, accurate estimates of the number of new young voters (if they are not made available by the election authorities) can be easily obtained in each unit (Pavía and Veres-Ferrer, 2016a, b). In a similar fashion, and depending on the size of the units, rough estimates of deceased voters could be computed applying age death probabilities to population figures.

Finally, the balance of immigrants and emigrants, who cancel each other out, could be computed in each territorial unit as a residue.

With regard to new young voters, which generally represent a significant part of the new voters, their size depends on the time elapsed between the two elections and the age structure of the population pyramid. For instance, currently in Spain for each year elapsed between two elections, these new voters represent, on average, slightly less than 1.1% of the population over 18 years. On the other hand, regarding deceased voters, we see that again their size depends on the time elapsed between the two elections and the population age pyramid as well as on mortality rates. Currently, in Spain, this rate, expressed as a percentage of the population over the age of 18, is on average more than 1.5% for each year elapsed between the elections.

It is noteworthy that, although it could be assumable that mortality and migration flows would have a similar effect on the different options competing in election 1, i.e., proportionally to their relative weight, it seems questionable to assume that young voters will be behave in election 2 similarly to the older electorate.

In the references consulted, however, it is always assumed, more or less implicitly, that new voters behave similarly to those who go out of the census would have done or in a similar fashion to those remaining in the census. For example, McCarthy and Ryan (1977) compute for each unit the difference between entries and exits and define, for each party k in the election 2, a new parameter, γ_k, to capture the behaviour of these differences. This makes it impossible to estimate the vote of the new electors separately. They define γ_k as the proportion of vote to option k of the net balances between entries and exits. Defined this way, it is easy to verify that the γ_k proportions are a complicated combination of the share of votes for party k of new and previous voters with weights that depend on the ratio between entries and exits in each unit, an issue that makes the hypothesis of territorial homogeneity for the γ_k's strongly questionable. A similar criticism can be made of the approach of Corominas et al. (2015), who assume that the number of voters in each unit is the same in both elections. This implies, as the authors themselves indicate, "*that the behavior of the electors not belonging to the intersection of both censuses is not different from those that belong to it*", not permitting an estimation of the young electors' vote and making the hypothesis of homogeneity implicit in their model more questionable.

3.3 Extending the basic model: The LPHOM model

With detailed election census lists available, it is theoretically possible to know in each unit, or at least roughly estimate and benchmark, the number of young voters (n_i), of immigrant entries (m_i) and of exits ($e_i + d_i$) between the two elections. Therefore, in this case, a more correct specification of the model would entail considering both kind of entries as additional election options ($J-1$ and J) of the first election and exits as a possible destination (K) of the votes in the second election and to include as additional constraints $p_{J-1,K} = 0$ and $p_{JK} = 0$.

The above scenario, however, is quite data-demanding. A more typical scenario is one in which only accurate estimates of young voters are available in each unit. In this case, assuming that the counted votes in each election correspond to the $J-1$ and $K-1$ first categories of the respective elections, net exits ($b_i = d_i + e_i - m_i$) can be easily computed from the available data $b_i = \sum_{j=1}^{J-1} x_{ij} + n_i - \sum_{k=1}^{K-1} y_{ik}$ and both n_i and b_i figures introduced in the problem, respectively, as option J of election 1 and option K of election 2. It should be note that in this situation, with units of sufficient size, net

exits will be always positive and in great part will coincide with the number of exits due to mortality because of the compensating effect of immigration and emigration.

Another scenario occurs with electoral processes very close in time, as is the case of, for example, the first and second rounds of the French presidential elections, where the changes in the electorate are really very small. In these cases, we can compute for each unit the net exits between the two elections as $b_i = \sum_{j=1}^{J-1} x_{ij} - \sum_{k=1}^{K-1} y_{ik}$ (considering again the same notation as in the previous scenario) and define in each unit the quantities given by equation (7). These new J and K categories will be irrelevant and, therefore, they could be omitted for presentation purposes.

$$\begin{array}{llll} y_{iK} = b_i & x_{iJ} = 0 & if & b_i \geq 0 \\ y_{ik} = 0 & x_{iJ} = -b_i & if & b_i < 0 \end{array} \qquad (7)$$

When no information about new young voters is available and the time elapsed between both elections is significant, the values on equation (7) will not be irrelevant. Even so, they can still be computed and our method applied after introducing them in the system, although at the cost of a loss of interpretability in some of the p_{jk} coefficients related to both categories J and K.

In a typical scenario, the rate transfers p_{jK} corresponding to net exits are less relevant than the other rates. What is more, as some simulation studies have shown us, their estimates are, as expected, quite volatile for the smallest election options. Hence, taking into account that they are mainly a consequence of mortality, we will force them, for the first $J-1$ election options considered in election 1, to be equal. This constraint might seem reasonable, as initially there is no reason to consider that mortality affects the older voters differently in the different political options. In addition, we will also impose the obvious condition $p_{JK} = 0$. These constraints are included in the model using equations (8) and (9).

$$p_{jK} = \left(\sum_{i=1}^{I} y_{iK}\right) \Big/ \left(\sum_{j=1}^{J-1}\sum_{i=1}^{I} y_{iK}\right) \quad j = 1, \ldots, J-1 \qquad (8)$$

$$p_{JK} = 0 \qquad (9)$$

The model we propose is to obtain the $J \times K$ values of the p_{jk} that, fulfilling constraints (1), (2), (3), (5), (8) and (9), minimize the sum of absolute values of the e_{ik}. Given that this model is ultimately a linear program, we propose to name it: LPHOM (acronym for **L**ineal **P**rogramme based on **HOM**ogeneity hypothesis). In the Appendix we describe an *R* function (whose code is provided in the Online Supplemental Materials) to apply LPHOM procedure to actual data in all the possible discussed scenarios.

3.4 Additional considerations

LPHOM offers a tool to estimate the transfer of votes between two elections separated by a certain period of time. It is also possible, however, to use LPHOM in formally analogous problems, but with no changes in the electoral census. This would be the case, for example, of a single election where voters are partitioned into J "groups" (based on criteria such as sex, race and/or social class), the x_{ij} are the numbers of electors in unit i belonging to group j and the y_{ik} are the votes gained by electoral option k in unit i, the objective being to estimate the proportions p_{jk} of voters of the different groups voting for the different options. This is a typical ecological inference problem. In these situations,

assuming that the hypothesis of homogeneity in electoral behaviour by group is reasonable (i.e., that the values of p_{jk} are "similar" in different units), the LPHOM model can be applied directly, but without including restrictions (8) and (9).

Another situation in which there are no changes in electoral censuses arises in simultaneous elections. This could be the case, for example, in Spain when general elections and regional elections are held simultaneously in a given autonomous region. In this scenario, the LPHOM model can be applied directly, obviously not including constraints (8) and (9). The challenge here arises in deciding which election should be considered as "origin" and which "destination".

In situations where they are changes in the electoral census and the J option in election 1 corresponds to the new young electors and the K option in election 2 as net exits, it is necessary to include in the model the restriction (9), being not indispensable, although reasonable, to consider constraint (8).

4. Assessing LPHOM with new voters

In this section, we exemplify the use of the method in a scenario where new voters are explicitly considered by applying LPHOM to the 2015 Aragon regional elections. The regional elections held in 2015 in Spain were very interesting for being the first time in which the then two new big emerging parties of Spanish politics, Podemos (POD) and Ciudadanos (C's), presented candidatures. POD is a left populist party, which has its roots in the so-called "15M movement" (Galais, 2014). C's is a centre-right party born in Catalonia to oppose the independent nationalism that in 2015 decided to expand throughout Spain. Both parties presented themselves as new, regenerative options opposed to the two traditional big parties, PP and PSOE, under fire due to their problems with corruption and the economic crisis (Bosch and Durán, 2017). In this scenario, almost all the analysts agreed that new voters were going to turn their backs to traditional main-stream parties. Despite new voters being a relatively small group (a 4% of the census in the 2015 Aragon regional elections), we aim to know whether LPHOM is able to properly capture their behaviour.

Aragon is chosen as our case of study because it is considered a swing territory that perfectly reflects the particular mood that the Spanish politics is living at the moment, like an electoral thermometer (Piedras de Papel, 2015). Aragon is one of seventeen autonomous regions in Spain. It is divided into three constituencies: Huesca, Teruel and Zaragoza; the latter holding the capital of the region where half of the total regional inhabitants live.

In 2011 regional elections, only six of the seventeen parties competing surpassed 1% of the total votes: PP (the conservative party), PSOE (the socialist party), IU (a left party with a heavy weight of communists), CHA (a left nationalist Aragonese party), PAR (a moderate nationalist conservative party) and UPyD (a party just created a few years earlier that was the largest party with no representation in the regional parliament). Table 1 provides a summary of the results of the regional elections held in Aragon in 2011 and 2015. In the table, blank and null votes have been added to non-voters (ABST), while REST groups the remaining minority parties.

Table 1. Summary of election outcomes for 2011 and 2015 Aragonese region elections.

	ABST	PP	PSOE	PAR	CHA	IU	UPyD	POD[(1)]	C's[(1)]	REST
2011	340,020	269,729	197,189	62,193	55,932	41,874	15,667	-	-	20,214
2015	332,911	181,757	141,528	45,577	30,334	27,936	5,637	135,554	62,188	17,357

(1) Podemos and Ciudadanos did not compete in the 2011 election.

To estimate the matrix of transfer votes between the 2011 and 2015 regional elections, we split Aragon into 15 territorial units: the provinces of Huesca and Teruel, the twelve election districts of the capital of the region and the rest of the province of Zaragoza. Although it is not a requirement of the approach to split the electoral space into a relative small number of spatial units, this practice shows three real-world benefits. First, it makes easier the homogeneous hypothesis to be verified (Pavía et al., 2008). Second, it avoids the problem of stablishing the correspondence between small-area election units of different periods (Pavía and López-Quilez, 2013; Pavía and Cantarino, 2017). Third, it significantly reduces the computational burden. The outcomes recorded in both elections in each of the 15 units considered are available in the Online Supplemental Materials (Tables S.1 and S.2).

The censuses of both 2011 and 2015 elections and the numbers of new young voters incorporated into the 2015 election census of each province as a result of having reached the legal age to vote (18 years old) since 2011 election, made public by the Spanish Official Statistical Agency (INE), were combined to estimate new entries and net exits between both elections. New electors were 6,836 in Huesca, 4,457 in Teruel and 29,224 in Zaragoza. New voters in the province of Zaragoza were divided among the thirteen territorial units in which we split this constituency in a fashion proportional to their total election populations. Given that the total population of the region decreased between 2011 and 2015, net exits were positive. We assume that net exits (mainly due to mortality) affected in a similar way he different options competing in the 2011 election (constraint (8)). Net exits accounted for 6.2% of 2011 census.

Table 2, where new young voters are referred to as ENTR and net exits as EXIT, shows the estimated transition probabilities between the options considered in 2011 and 2015 elections obtained after applying LPHOM procedure. From the data in Tables 1 and 2 it is easy to obtain Table S.3 in the Online Supplemental Materials that shows the origin of the votes obtained by the different options competing in Aragon regional election in 2015.

Table 2. Estimated vote transfer matrix (in percentages) between 2011 and 2015 Aragon regional elections.

	ABST	PP	PSOE	POD	C's	PAR	CHA	IU	UPyD	REST	EXIT
ABST	70.7	*	*	19.5	0.9	*	*	0.1	0.2	2.2	6.2
PP	13.6	65.2	*	*	13.6	0.1	*	*	0.2	0.7	6.2
PSOE	20.0	*	65.5	3.3	*	*	*	*	*	*	6.2
PAR	*	9.1	19.7	*	*	64.8	*	*	*	*	6.2
CHA	*	*	*	56.7	*	*	37.0	*	*	*	6.2
IU	*	*	*	36.9	*	*	*	54.6	*	2.1	6.2
UPyD	*	*	*	*	75.2	*	*	*	18.5	*	6.2
REST	44.5	*	*	*	*	23.1	*	2.9	*	23.1	6.2
ENTR	17.5	*	*	37.9	26.0	*	8.0	2.5	3.0	4.8	0.0

Note: Since the solution of a linear program is always a basic solution, LPHOM tends to make exactly 1 or 0 the results very close to these values. Therefore, we prefer to substitute ones, if they exist, for 0.999 and zeros for the asterisk symbol indicating a very low value.

Despite the purely mathematical nature of LPHOM procedure, which does not include any consideration regarding the ideological proximity between the different options competing in both elections, outcomes in Tables 2 and S.3 are extremely clear and simple to interpret from the point of view of political sociology. For example, we can observe that: (i) the most important source of the votes gained for any party in 2015 are the voters who voted for that same party in 2011, if the party competed at that election; (ii) the votes lost by the two main traditional parties were mostly to abstention and to the two new parties following an ideological alignment, C's in the case of PP and POD in the case of PSOE; (iii) the new party POD received most of its votes from former abstentions, from previous left-wing party voters (CHA, IU and PSOE) and from new young voters; and, (iv) the

new centre-right party C's gained its votes mainly from previous PP (conservative) and UPyD (a party very close ideologically to C's) voters, from new young voters and from former abstentions. Interested readers on the subject can consult a more detailed analysis in this regard in the Online Supplemental Materials.

For the purpose of this paper, we focus on the transition probabilities estimated for new young voters, whose behaviour is clearly different from those of previous election voters. The new parties POD (37.9%) and C's (26.0%), followed by abstention (17.5%), were the preferred choices of new young voters; whereas, the two traditional parties, PP and PSOE, had almost no success among this electorate. It is important to emphasize that this differential estimate of the new voters' voting pattern is not possible through the procedures proposed by other authors. Contrary to what is assumed in those procedures, the voting patterns of young electors are distinct from that found at a global level in the region, where PP and PSOE were the most voted parties.

5. Estimating the uncertainty of model results

5.1 Introduction

The fundamental problem in scientifically establishing the validity of the methodology comes from the fact that it is (almost) impossible to compare LPHOM outcomes with actual transition probabilities. Aside from extraordinary circumstances, such as in simultaneous elections where the same ballot paper is used to vote for the different political contests and individual votes are available, actual voter transition probabilities are impossible to know. Likewise, due to the lack of reliability of retrospective answers and poll data for these kinds of studies, comparing LPHOM outcomes to survey approximations does not seem to be a valid alternative.

Faced with this impossibility, we can conceive, in principle, two possible approaches for judging the validity of the proposed method. One alternative is to assess the logic and rationality of the process followed to estimate the vote transfer matrices. The other alternative is to analyse whether the results provided by the method are "reasonable" when applied to actual elections. With respect to the first point, that of the rationality of the process, we have already discussed in Section 3 the soundness and logic of the hypothesis of homogeneity of electoral mobility in the different units, provided that the conditions indicated therein were fulfilled in the definition of the territorial units. Regarding the second alternative, a former and simpler version of PLHOM procedure has been used in a number of recent electoral processes held in several Spanish regions (see Table 3). There is no particular reason to choose these elections beyond opportunity and easy access to the data for the authors. In our opinion, which is also shared by many Spanish experts in political sociology, in all cases the results obtained (which can be consulted, in Spanish, in the Online Supplemental Materials and in the references indicated in Table 3) have been "reasonable", in the sense of being logical and clearly interpretable in sociological terms. As an example, we have always obtained that the most important source of votes of any party competing in election 2 was the voters of the same party in the previous election (if the party contested at that election). This remark can look surprising because we should remember that the results obtained are based on a purely mathematical manipulation of the data that does not take into account the possible ideological proximity among the different options, nor even between a party in election 1 and the same party when competing in election 2. The fact that the model, despite its purely mathematical nature, has always provided reasonable results in actual studies seems to be a certain guarantee of its validity, that is, of the validity of the homogeneity hypothesis on which it rests.

Table 3. Some studies performed using a former version of LPHOM procedure.

Region	Number of units (I)	Election 1	Options in Election 1* (J)	Election 2	Options in Election 2* (K)	Source**
Aragon	15	2011 regional election	8	2015 regional election	10	Online Suppl. Materials
Madrid	13	2011 regional election	5	2015 regional election	7	Online Suppl. Materials
Valencian Region	10	2011 regional election	7	2015 regional election	9	Online Suppl. Materials
Andalusia	8	2012 regional election	7	2015 regional election	9	Online Suppl. Materials
Catalonia	10	2012 regional election	9	2015 regional election	8	Romero (2015)
Andalusia	8	2015 regional election	7	2015 General elections	7	Online Suppl. Materials
Valencian Region	14	2015 regional election	8	2015 General elections	7	Online Suppl. Materials
Andalusia	8	2015 General elections	7	2016 General elections	6	Romero (2016)
Madrid	23	2015 General elections	7	2016 General elections	6	Romero (2016)
Valencian Region	14	2015 General elections	7	2016 General elections	6	Romero (2016)
Basque Country	8	2016 General elections	6	2016 regional election	6	Online Suppl. Materials

* Entries and exits in census were not included in these studies.
** Associated documents in Spanish.

At this point, therefore, the question is: What is the margin of uncertainty of the results obtained when applying LPHOM procedure to a specific study? As we discuss in following subsections, the model outcomes provide information to calculate a heterogeneity index that allows the adequacy of the homogeneity hypothesis in each actual study to be quantified as well as the margin of uncertainty of the results achieved. This is the main contribution of our paper: a procedure to assess the level of uncertainty of the estimates. No other previously published method based on linear or quadratic programming method provides such a measure.

5.2 Model residuals: estimating the heterogeneity

If the hypothesis of electoral homogeneity was fulfilled exactly in a given study, that is, if all p_{jk}^{i} were exactly equal to their average value p_{jk} in the whole territory, LPHOM would yield as output the unknown true values of p_{jk} with all the e_{ik} errors being zero. The departure of homogeneity hypothesis in each unit is therefore captured somehow in the residuals. In this and the following subsections we show how these can be used to estimate the uncertainty of LPHOM outputs.

To address the problem of quantifying the uncertainty associated with LPHOM outcomes, it is therefore important to estimate in each real instance the extent to which the homogeneity hypothesis is verified. If all the true vote transfer matrices in each unit were known, the degree of non-compliance of the homogeneity hypothesis can be easily quantified using, for instance, the heterogeneity index HET defined in equation (10), where v_{jk}^{i} denotes the number of voters that, in unit i, choose option j in election 1 and option k in election 2.

$$HET = 100 \cdot \frac{0.5 \sum_{ijk} |v_{jk}^{i} - x_{ij} p_{jk}|}{\sum_{ij} x_{ij}} \qquad (10)$$

As can be clearly seen, HET accounts for the percentage of voters which should be shifted to match perfectly the homogeneity hypothesis. The problem with HET lies in the impossibility of computing it in real studies, given that actual values for v_{jk}^{i}, and also for p_{jk}, are unknown. Nevertheless, since if HET were zero all the e_{ik} residuals would also be null, it makes sense to define an estimated heterogeneity index $HETe$ through equation (11).

$$HETe = 100 \cdot \frac{\sum_{ik} |e_{ik}|}{\sum_{ij} x_{ij}} \qquad (11)$$

Unlike HET, the $HETe$ value can be calculated in any actual study from the results provided by LPHOM.

As we show in the next subsection, we have carried out a set of simulation studies in five different scenarios to analyse, among other issues, the relationship between HET and $HETe$. Considering together the results of the 6900 simulations performed, 1380 in each one of the five scenarios, we obtain an almost perfect linear relationship between $log(HETe)$ and $log(HET)$, with a Pearson correlation of 0.989 and $HET = 1.626 \cdot (HETe)^{0.921}$ as fitted equation. These results reveal HETe as being a good predictor of the real heterogeneity index HET. In next subsections, we exploit this relationship to quantify the uncertainty associated with the results provided by LPHOM in actual studies.

5.3 Relationship between error index and heterogeneity index

In order to assess the relationship between the estimated heterogeneity index (HETe) and the uncertainty of the results provided by LPHOM, we have carried out a set of simulation studies. These studies have been implemented in five different scenarios characterized by two matrices X and Q.

- The matrix $X = [x_{ij}]$ collects the results achieved in election 1 by the different options in the different territorial units. This matrix accounts for the impact of numbers I and J (number of units and options considered in the first election) and for the degree of electoral diversity in the different units in election 1.

- The basic matrix $Q = [q_{jk}]$ of global voting transitions between the options presented in both elections. This matrix accounts for the impact of number K (the number of options considered in the election 2) as well as for the basic structure of the voter transitions.

To generate randomly a certain degree of heterogeneity in the transition matrices of the different units, the p_{jk}^{i} values have been obtained by adding to the q_{jk} values a uniform random variable between $-d$ and $+d$. These initial p_{jk}^{i} values are subsequently readjusted to be non-negative and verifying $\sum_k p_{jk}^{i} = 1$ for all i and j. The level of heterogeneity is regulated by d. In all the simulated scenarios, we attained a correlation coefficient between d and HET higher than 0.99. More details of the

simulations performed are available in the Online Supplemental Materials that accompanies this paper.

From p_{jk}^{i} and x_{ij} we build the hypermatrix $W = [w_{jk}^{i}]$, whose generic element is the number of voters that swing from option j in election 1 to option k in election 2 in unit i. From W it follows the matrix $V = [v_{jk} = \sum_i w_{jk}^{i}]$ of transition votes in the overall territory and the matrix $P = [p_{jk}]$ of voter transition probabilities at the global level. In general P is close to Q.

From W we also obtain the matrix $Y = [y_{ik} = \sum_j w_{jk}^{i}]$ whose generic element represents the number of votes gained for each option j of election 2 in unit i. The matrices X and Y are given as inputs to LPHOM, from which we obtain the estimated matrices of vote transitions $V^* = [v_{jk}^*]$ and of transition probabilities $P^* = [p_{jk}^*]$. These matrices are compared to V and P in order to assess the degree of proximity between estimated and "actual" results. LPHOM also provides the e_{ik} values of the residuals and the value of the estimated heterogeneity index $HETe$.

An overall measure of the discrepancy between V and V^* is the error index, EI, defined by equation (12).

$$EI = 100 \cdot \frac{0.5 \sum_{jk} |v_{jk} - v_{jk}^*|}{\sum_{jk} v_{jk}} \qquad (12)$$

It is easy to verify that EI is the percentage of votes whose destination has been erroneously estimated by the model.

For each scenario, we have considered 46 possible values of d, chosen between 0.001 and 0.1, and performed 30 simulations for each value. The characteristics of the five scenarios and the results obtained for HETe and EI in each one of the 1380 simulations performed in each scenario can be consulted in the Online Supplemental Materials. In all scenarios analysed, we find a close relationship between the error indexes, EI, and the estimated heterogeneity indexes, $HETe$. In the five cases this relationship is satisfactorily modelled by a regression equation using $log(EI)$ as dependent variable and $log(HETe)$ and its square as predictor variables. The high values attained for the multiple correlation coefficient (0.931 in Scenario 1, 0.976 in Scenario 2, 0.965 in Scenario 3, 0.972 in Scenario 4 and 0.976 in Scenario 5) show that $HETe$ is a good predictor of EI.

Figure 1 displays the mean values predicted for EI as a function of $HETe$ in the five scenarios analysed. As can be observed in the figure, although the general shape of the relationships are very similar in all the scenarios, the particular values of the corresponding equations noticeably differ among scenarios.

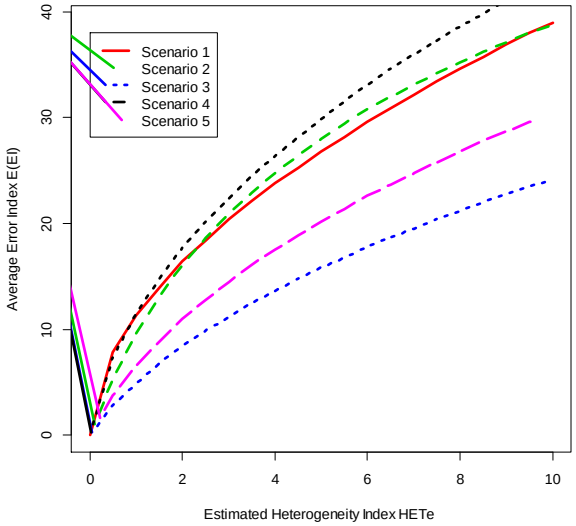

Figure 1. Relationships between averages of EI and $HETe$ in five simulated scenarios.

Preliminary simulation studies that we have undertaken seem to indicate that some of the factors that influence the relationship between EI and $HETe$ are the ratio between the number of equations in the model and the number of unknowns p_{jk} and the degree of diversity of the results in election 1 in the I territorial units.

5.4 A procedure to estimate outcomes' uncertainty in actual studies

Outcomes from the previous subsection show that to evaluate the degree of validity of PLHOM results it is necessary to estimate in each specific study the particular function that relates EI and $HETe$. This can be performed by means of a simulation study similar to those carried out in this research, but using the particular scenario defined by the data corresponding to the study. That scenario will be characterized by the X matrix with the results obtained in election 1 in the I different territorial units and by the P matrix of voter transition estimated by PLHOM.

From the relationship estimated in this way and from the particular value for $HETe$ computed by PLHOM, it is possible to estimate the predicted value for the error index EI in the instance under study, and also to establish confidence limits for this index. In the next section, the operative of this procedure is illustrated applying it to the study of the voter transfers between the first and second rounds of the 2017 French Presidential election.

6. Voter transitions between first and second round of the 2017 French Presidential election

To illustrate the simplicity of the LPHOM procedure, we estimate and analyse voter transitions between the first and second rounds of 2017 French Presidential election, measuring its level of uncertainty. This an interesting case of study due to their political relevance and because, as can be deduced from Table 4, at least 20 million French people changed their vote between 23 April and 7 May 2017, the dates of the first and second rounds. Knowing how votes of first and second rounds relate is undoubtedly relevant to understanding the main drivers that operated during that election.

Table 4 provides the votes gained at a national level for the main candidates (Emmanuel Macron, Marine Le Pen, François Fillon, Jean-Luc Mélenchon, Benoît Hamon and Nicolas Dupont-Aignan) in both rounds, with "Others" grouping the remaining candidates; those who received less than half

a million votes. Abstaining (NonVoters) and voting either blank or null (BlaNull) complete the electors' alternatives.

Table 4. National results of first and second rounds of the 2017 Presidential French election.

Round	Census	NonVoters	BlaNull	Macron	Le Pen	Fillon	Mélechon	Hamon	Dupont	Others
First	47,582,183	10,578,455	949,334	8,656,346	7,678,491	7,212,995	7,059,951	2,291,288	1,695,000	1,460,323
Second	47,568,693	12,101,366	4,085,724	20,743,128	10,638,475	-	-	-	-	-

Source: Official results from https://www.conseil-constitutionnel.fr/ Retrieved 3 March 2020.

To run LPHOM, we need a partition of the territory under study and the election outcomes recorded in such a set of territorial units. In this research, we consider the official results recorded in the 107 departments in which France was divided plus an artificial department that grouped the French electors living abroad. The election results of both rounds at department level can be consulted in the Online Supplemental Materials (Tables S.4 and S.5). As expected, given the temporal proximity of the two elections, the changes in the censuses between them have been minimal. However, as LPHOM still requires that $\sum_j x_{ij}$ matches exactly $\sum_k y_{ik}\ \forall i$, a column with (net) census entries and another one with (net) census exits need to be added to the respective matrices X and Y to account for the differences. Before applying LPHOM, these columns can be calculated following any of the approaches pointed out in subsection 3.2. In our solution, we use the R-function described in the Appendix and provided in the Online Supplemental Materials (with the option new_and_exit_voters = "raw"), which implicitly implements equation (7) in this circumstance. Hence, our LPHOM implemented model to estimate the transitions of votes between first and second rounds of the 2017 French Presidential election is a linear program with 1130 variables and 565 restrictions. Its solution, by means of our R function, takes less than 0.6 seconds on a standard PC.

Table 5 shows the results attained. According to our estimates, 86.6% of first-round non-voters did the same and not vote in the second round either, while 9.5% of them voted for Macron and about 3.8% for Le Pen. As expected, virtually all the electors who voted for either Macron or Le Pen in the first round again chose the same candidate in the second round. We want to emphasize that although this output seems very logical, at no time has LPHOM been informed that "Macron in the second round" is the same candidate as "Macron in the first round", nor the equivalent information concerning Le Pen.

Regarding the behaviour of the voters of the remaining candidates, it seems that LPHOM was able to capture the logical transfers according to the ideology of the candidates. Thus, Fillon's centre-right supporters mostly voted in the second round for the centrist, independent candidate Emmanuel Macron, with the remaining of his voters split between abstentions and blank or null votes. Just a few of these voters chose the far-right Front National Leader Marine Le Pen. Likewise, practically all voters of the socialist Hamon decided to vote Macron in the second round. On the other hand, the first-round voters of the populist Mélenchon shared out much of their vote in the second round: 48% of them deciding to vote Macron and 11% to vote Le Pen, with the rest of Mélenchon's voters split between abstention and blank or null votes. Equally logical is that the majority of the first-round voters of the ultranationalist DuPont decided to vote Le Pen in the second round. Finally, we find that 90% of the nearly one and a half million voters who voted for other minority candidates in the first round voted blank or null in the second round, with the remaining 10% of them voting for Le Pen.

Our results can be assessed by comparing them with the outcomes obtained using ecological regression and with the estimates derived from several polls conducted between the first- and second-round of the election (see Tables S.7 to S.10 in the Online Supplemental Materials).

Table 5. Estimated swings between first- and second round of 2017 French presidential election.

	Non Voters	Blank and Null	Emmanuel Macron	Marine Le Pen
Non Voters	86.6	*	9.5	3.8
Blank and Null	*	61.3	*	38.6
Macron	*	*	99.9	*
Le Pen	*	*	*	99.9
Fillon	16.4	5.8	74.5	3.3
Mélechon	24.6	16.2	48.5	10.7
Hamon	*	*	99.9	*
Dupont	*	37.7	*	62.2
Others	*	89.4	*	10.5

Notes: In the first round, blank and null votes are included in Non-voters. See also note under Table 2.

Regarding the first issue, we have compared our vote transfer estimates (see Table 5) to (i) the estimates published in Pons (2017), who applies King's method, and (ii) the transfers that can be reached after applying to our data the function ei.MD.bayes. The function ei.MD.bayes of the R-library eiPack (Lau, Moore and Kellermann, 2018) implements a version of the Bayesian hierarchical model suggested in Rosen et al. (2001) and, according to Klima et al. (2016), exhibits the best overall estimation performance among the most commonly used approaches. We have found that our results are quite similar to the ones attained by Pons (2017). For example, Pons estimates 9.0% of first-round non-voters voting for Macron in the second round or 21% first-round Fillon's voters deciding to abstain in the second-round. It should be noted, nevertheless, that we only used 108 units and spent less than a second of computation, whereas Pons (2017) used 69,241 units (*bureaux de votes*) and spent several hours of computation. However, after applying (with default options and with the help of the MDtune function) the function ei.MD.bayes to our data, we obtained nonsense estimates. For example, with default options, ei.MD.bayes estimates about 43% of first-round Le Pen voters choosing Macron in the second-round. It seems that ei.MD.bayes needs many units, a really proper tune of its key parameters and a lot of computational time to reach reasonable estimates. Even so, according to Plescia and De Sio (2018) and Klein (2019), the coverages of its resulting credible intervals are well below the target credible levels.

Regarding the second issue, comparing LPHOM outcomes (Table 5) and polls estimates (Tables S.7 to S.10), we see that both sets of estimates exhibit the same patterns, but each of them with their own nuances. In our view, our results are superior to survey estimates because they are fully consistent with actual outcomes (fulfilling all the constraints) and they offer vote transfer estimates between all the relevant election options. Furthermore, they are not exposed to sources of error such as nonresponse bias, social desirability, measurement error or changes of opinion. One drawback to our solution is that it probably underestimates slightly the electoral mobility. This drawback could be significantly reduced using more detailed data, for example, using outcomes at municipality or, even better, at precinct level.

In addition to an estimate of the vote transfer matrix, LPHOM also provides the estimated heterogeneity index, $HETe$, which reaching 4.21% for this study indicates the degree of compliance of the hypothesis of homogeneity. Indeed, this index can be observed as the average of the discrepancies between the global transition matrix and the corresponding transition matrices in each territorial unit. In particular, computing heterogeneity indexes for each unit, we find that these range between a minimum of 0.46% for the department of Tarn and a maximum of 23.8% for French Polynesia.

As we stated in Section 5, once one disposes of an estimate of the heterogeneity index, it is possible to approximate the uncertainty linked to the estimated vote transfer matrix. To compute this, we carry out a simulation study similar to those described in subsection 5.4, but using the data corresponding to the current scenario, which is defined by the matrix X available in Table S.5 in the Online Supplemental Materials and the Q matrix of voter transitions of Table 5. From this, we simulate 46 possible values of d between 0.001 and 0.1 and run 30 simulations for each value.

Figure 2 displays the values attained for $HETe$ and EI in the 1380 simulations completed. In the Online Supplemental Materials (simulations.csv), interested readers can consult the series of $HETe$ and EI obtained. As can be seen in Figure 2, a close relationship links $log(EI)$ and $log(HETe)$, also for this dataset. The black line in Figure 2 depicts the equation relating both statistics and Table S.6 in the Online Supplemental Materials offers the details of the model fitted. At this point, and plugging the value 4.21 reached for $HETe$ in the estimated equation, we obtain the amount 8.7% as estimate for EI, with an upper confidence limit $(1 - \alpha = 0.90)$ of 11.2%. Remember that EI can be interpreted as the percentage of votes whose destination has been erroneously estimated by the model.

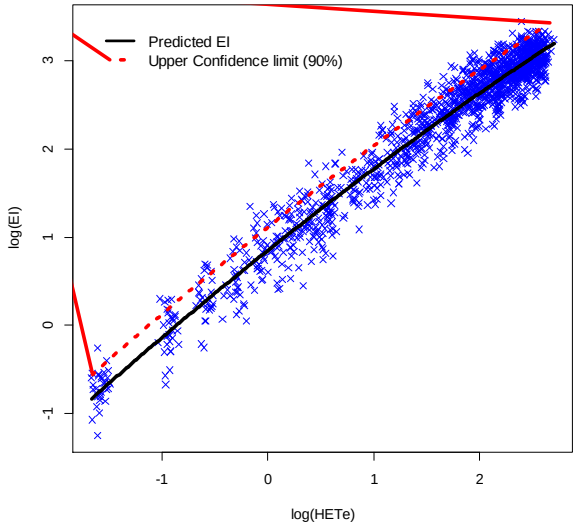

Figure 2. Relationship between $log(EI)$ and $log(HETe)$ for the 2017 French presidential election.

At a first glance, the estimated EI looks high. In order to contextualize it, we compare this to the numbers reached by other methods in similar problems. In this sense, Klima et al. (2016) assess the performance of five different ecological inference procedures estimating by simulation the voter transfer matrix in five different scenarios. In that research, Klima and colleagues evaluate performance using as, measure of dissimilarity, an index, AD, defined as two times our EI index. The average values they obtain for these AD indices depend on the scenario and the procedure considered, ranging from a minimum of 10% to a maximum of 60%, most of them being between 20% and 30%. Although comparisons could be debatable, given that results strongly depend on the scenario considered, it seems that the performance obtained in our study is better than that observed when using other much more complicated procedures.

7. Concluding remarks and further research

The deficient reliability of responses to retrospective questions, the challenge posed by nonresponse bias together with the financial costs and large sample sizes required to reach suitable estimates of vote transition probabilities have encouraged many authors to look for an alternative to polls to solve the problem of estimating voter transfer matrices. In this vein, several authors have taken the route of

estimating vote transitions between two elections using exclusively the undisputable records available in the election aggregate outcomes. The determination of the transfer vote matrix based exclusively on these aggregate data is, however, a basic indeterminate problem, whose resolution requires the imposition of additional hypotheses. Hence, the validity of the estimates, that is, their closeness to the unknown true values, depends on the extent to which these hypotheses are satisfied in the electoral processes under scrutiny.

Both in ecological regression procedures and in mathematical programming approaches, the basic idea behind such hypotheses is that vote transfer matrices in the different territorial units in which the whole area is partitioned are, in some sense, "similar" to each other, and therefore similar to the global matrix. Mathematical program procedures, such as LPHOM, have the advantage of being much simpler to apply than equivalent ecological regression methods. This is mostly true when we consider Bayesian ecological inference approaches, which require specific training as well as previous knowledge and expertise that many analysts lack. Furthermore, this higher simplicity of mathematical program procedures is reached without impairment to the quality of estimates. LPHOM has provided reasonable results in all the actual studies where it has been tested.

Furthermore, from a computational point of view, (Bayesian) ecological inference approaches are computationally very intense, demanding really huge amounts of processing times in models involving many variables. The hierarchical distributional structures that characterize Bayesian approaches mean that, on the one hand, the Markov Chain Monte Carlo (MCMC) procedures routinely employed require very long computation times and that, on the other hand, the analysts must face convergence problems even when parameter transformations are performed (Klima et al., 2016). And, although the advent of Stan language (Carpenter et al., 2017) is speeding up many Bayesian problems, its use still remains quite complex for the average analyst.

Likewise, compared to other procedures based on mathematical programming, we find some important advantages in our approach. First, LPHOM considers explicitly new voters, making it possible to estimate their behaviour differentially. Secondly, LPHOM offers a way to estimate, from the model results, the degree of non-compliance of the basic hypothesis of homogeneity. And lastly and more importantly, as shown in subsections 5.2 to 5.4, we provide a procedure to assess the level of uncertainty of the estimates.

Regarding further research, we have started a new line in order to investigate, by simulation, the factors that influence the accuracy of the outcomes provided by LPHOM and to compare them with those obtained by ecological inference procedures. Our provisional results, which we expect to explain in detail in a future paper, point towards other factors, in addition to the value of the heterogeneity index, as features affecting the precision of the estimates. It seems that their accuracy also depends on the ratio between the number of equations in the model and the number of unknowns p_{jk} and on the degree of diversity of the results of election 1 in the I territorial units under consideration. These results seem logical and some of them have been pointed out by others authors (King, 1997). Since electoral results are generally available disaggregated for a large number of elementary units, a good knowledge of the factors influencing the quality of the estimates could indubitably help in the process of establishing a proper strategy for grouping elementary units into territorial units with the aim of reaching estimates as accurate as possible.

The main goal of LPHOM is the estimation of the overall voter transition matrix $P = [p_{jk}]$, and not the transfer matrices $P^i = [p^i_{jk}]$ in the different territorial units in which the whole territory has been partitioned. This does not mean that we have no local information. The residuals e_{ik} carry useful information about what has happened in each of these units. For example, a high positive e_{ik} would

indicate that voter transitions to party k in unit i have been less intense than that found in the average unit. A possible approach to estimate the P^i transition matrices, proposed by Corominas et al. (2015) and consistent with the electoral homogeneity hypothesis, could be to obtain de p_{jk}^i values that, matching perfectly electoral results in unit i, are more similar to the p_{jk} obtained for the whole territory. The adequacy of this approach and the properties of the outcomes it provides could be the object of a further research.

Lastly, the LPHOM approach could be generalized to a tri-electoral model to estimate the hypermatrix $[p_{jkm}]$ whose generic element would be the proportion of voters who having chosen option j in election 1 and option k in election 2 vote for option m in a later election. For example, in Spain there is much interest in knowing the proportion of voters who having swung from PP to C's or Abstention in a second election returned back to PP in the following election. This problem seems difficult to deal with from ecological inference approaches, but looks simpler from a mathematical linear framework. This could be addressed by means of a generalization of the PLHOM procedure.

Appendix: An R function to apply PLHOM procedure

This appendix describes the details of an *R*-function, called **lphom**, created by the authors to implement the PLHOM procedure described in the paper. The code of the function is available in the Online Supplemental Materials. The function estimates, given the results gained in a set of I spatial units by the J political options (parties or candidates) competing in election 1 and the K political options competing in election 2, the $J \times K$ matrix of vote transition probabilities between the two elections.

This function, which depends on **lpSolve** package (Berkelaar et al., 2014), has five arguments (***votes_election1***, ***votes_election2***, ***new_and_exit_voters***, ***structural_zeros*** and ***verbose***) and returns a list with four objects (***VTM***, ***OTM***, ***EHet*** and ***HTEe***).
The arguments of the function are:

- ***votes_election1***: data.frame (or matrix) of order $I \times J$ with the votes gained by the J political options competing on election 1 (or origin) in the I territorial units considered.
- ***votes_election2***: data.frame (or matrix) of order $I \times K$ with the votes gained by the K political options competing on election 2 (or destination) in the I territorial units considered.
- ***new_and_exit_voters***: A character argument indicating the level of information available regarding new entries and exits of the election censuses between the two elections. This argument captures the different options discussed on Section 3. This argument admits five values: `"regular"`, `"raw"`, `"simultaneous"`, `"full"` and `"gold"`.
    - `regular`: The default value. This argument accounts for the most plausible scenario. A scenario with two elections elapsed at least some months. In this scenario, (i) the column J of ***votes_election1*** corresponds to new young electors who have the right to vote for the first time, (ii) net exits (basically a consequence of mortality), and eventually net entries, are computed according equation (7), and (iii) we assume net exits affect equally all the first $J-1$ options of election 1, hence (8) and (9) constraints are imposed.
    - `raw`: This value accounts for a scenario with two elections where only the raw election data recorded in the I territorial units, in which the area under study is divided, are available. In this scenario, net exits (basically deaths) and net entries (basically new young voters) are estimated according to equation (7). Constraints defined by equations (8) and (9) are imposed. In this scenario, when net exits and/or net entries are negligible

(such as between the first- and second-round of French Presidential elections), they are omitted in the outputs.

- `simultaneous`: This value accounts for either a scenario with two simultaneous elections or a classical ecological inference problem. In this scenario, the sum by rows of $\boldsymbol{votes\_election1}$ and $\boldsymbol{votes\_election2}$ must coincide. Constraints defined by equations (8) and (9) are not included in the model.
- `full`: This value accounts for a scenario with two elections elapsed at least some months, where: (i) the column $J-1$ of $\boldsymbol{votes\_election1}$ totals new young electors that have the right to vote for the first time; (ii) the column J of $\boldsymbol{votes\_election1}$ measures new immigrants that have the right to vote; and (iii) the column K of $\boldsymbol{votes\_election2}$ corresponds to total exits of the census lists (due to death or emigration). In this scenario, the sum by rows of $\boldsymbol{votes\_election1}$ and $\boldsymbol{votes\_election2}$ must agree and constraints (8) and (9) are imposed.
- `gold`: This value accounts for a scenario similar to `full`, where total exits are separated out between exits due to emigration (column $K-1$ of $\boldsymbol{votes\_election2}$) and death (column K of $\boldsymbol{votes\_election2}$). In this scenario, the sum by rows of $\boldsymbol{votes\_election1}$ and $\boldsymbol{votes\_election2}$ must agree. The same restrictions as in the above scenario apply but for both columns $K-1$ and K of the vote transition probability matrix.

- $\boldsymbol{structural\_zeros}$: Default NULL. A list of vectors of length two, indicating the election options for which no transfer of votes are allowed between election 1 and election 2. For instance, when $\boldsymbol{new\_and\_exit\_voters}$ is set to `"regular"`, **lphom** implicitly states $\boldsymbol{structural\_zeros}$ = `list(c(J, K))`.
- $\boldsymbol{verbose}$: A TRUE/FALSE argument that indicates if the main outputs of the function should be printed on the screen. Default TRUE.

The outputs of the function are:

- $\boldsymbol{VTM}$: A matrix of order $J \times K$ with the estimated percentages of vote transitions from election 1 to election 2. Tables 2 and 6 are examples of $\boldsymbol{VTM}$ matrices.
- $\boldsymbol{OTM}$: A matrix of order $K \times J$ with the estimated percentages of the origin of the votes obtained for the different options of election 2. Table 3 is an example of a $\boldsymbol{OTM}$ matrix.
- $\boldsymbol{EHet}$: A matrix of order $I \times K$ measuring in each spatial unit the distance to the homogeneity hypothesis, that is, the differences under the homogeneity hypothesis between the actual recorded results and the expected results in each territorial unit for each option of election two. The matrix $[e_{ik}]$.
- $\boldsymbol{HTEe}$: The estimated heterogeneity index defined in equation (11).

References

Abou-Chadi, T. and Stoetzer, L. 2020. "How Parties React to Voter Transitions." *American Political Science Review*, online available.

Baydoğan, U. 2019. V*ote Transitions Analysis and Comparison of Turkish Local Elections in 2014 and 2019*. PhD Dissertation. Mef University.

Berkelaar, M. and others. 2014. *lpSolve: Interface to Lp_solve v. 5.5 to solve linear/integer programs*. R package version 5.6.10. https://CRAN.R-project.org/package=lpSolve

Bosch, A., and I. M. Durán. 2017. "How does economic crisis impel emerging parties on the road to elections? The case of the Spanish Podemos and Ciudadanos." *Party Politics*, online available.

Brown, P. J., and C. D. Payne. 1986. "Aggregate data, ecological regression and voting transitions." *Journal of the American Statistical Association* 81:453–460.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. 2017. "Stan: A probabilistic programming language." *Journal of Statistical Software* 76(1):1–32.

Caughey, D., and M. Wang. 2019. "Dynamic Ecological Inference for Time-Varying Population Distributions Based on Sparse, Irregular, and Noisy Marginal Data," *Political Analysis* 27(3):338–396.

CIS. 2014. *Estudio 3041. Barómetro octubre 2014*. Madrid: Centro de Investigaciones Sociológicas.

Cho, W. K. T. 1998. "Iff the Assumption Fits…: A Comment on the King Ecological Inference Solution." *Political Analysis* 7: 143–163.

Corominas A., A. Lusa, and M. D. Valvet, 2015. "Computing Voter Transitions: The Elections for the Catalan Parliament, from 2010 to 2012." *Journal of Industrial Engineering and Management* 8(1):122–136.

Dassonneville, R., and M. Hooghe. 2017. "The Noise of the Vote Recall Question: The Validity of the Vote Recall Question in Panel Studies in Belgium, Germany, and the Netherlands." *International Journal of Public Opinion Research* 29(2):316–338.

Duncan, O., and B. Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18:665–66.

Fisher, L. H., and J. Wakefield. 2020. "Ecological inference for infectious disease data, with application to vaccination strategies." *Statistics in Medicine* 39(3):220–238.

Forcina, A., and G. M. Marchetti. 1989. "Modelling transition probabilities in the analysis of aggregate data." In A. Decarli, B. J. Francis, R. Gilchrist, and G. U. H. Seber, Eds. *Statistical Modelling*. Springer-Verlag.

Forcina A., and G. M. Marchetti. 2011. "The Brown and Payne Model of Voter Transition Revisited." In S. Ingrassia, R. Rocci, and M. Vichi, Eds. *New Perspectives in Statistical Modeling and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin: Springer.

Forcina, A. and Pellegrino, D. 2019. "Estimation of Voter Transitions and the Ecological Fallacy." Quality & Quantity 53:1859–1874.

Freedman, D. A., S. P. Klein, M. Ostland, and M. R. Roberts. 1998. "Review of 'A Solution to the Ecological Inference Problem'." *Journal of the American Statistical Association* 93:1518–1522.

Füle, E. 1994. "Estimating Voter Transitions by Ecological Regression." *Electoral Studies* 13:313–330.

Galais, C. 2014. "Don't Vote for Them: The Effects of the Spanish Indignant Movement on Attitudes about Voting." *Journal of Elections, Public Opinion and Parties* 24:334–350.

Glynn, A.N, and J. Wakefield. 2010. "Ecological inference in the social sciences." *Statistical Methodology* 7(3):307–322.

Goodman, L. A. 1953. "Ecological Regressions and the Behaviour of Individuals." *American Sociological Review* 18:663–666.

Goodman, L. A. 1959. "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64(6):610–625.

Greiner, D. J. and Quinn, K. M. 2009. "R×C ecological inference: Bounds, correlations, flexibility, and transparency of assumptions." *Journal of the Royal Statistical Society. Series A* 172:67–81.

Greiner, D., and K. M. Quinn. 2010. "Exit Polling and Racial Bloc Voting: Combining Individual-level and RxC Ecological Data." *The Annals of Applied Statistics* 4:1774–1796.

Hawkes, A. G. 1969. "An Approach to the Analysis of Electoral Swing." *Journal of the Royal Statistical Society, Series A* 132: 68–79.

Henn M., and N. Foard. 2012. *Young People and Politics in Britain: How do Young People Participate in Politics and What Can Be Done to Strengthen their Political Connection?* London: Nottingham Trent/ESRC.

Johnston, R. J., and A. M. Hay. 1983. “Voter Transition Probability Estimates: An Entropy Maximizing Approach.” *European Journal of Political Research* 11:93–98.

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

King, G., O. Rosen, and M. A. Tanner. 1999. “Binomial-beta Hierarchical Models for Ecological Inference.” *Sociological Methods & Research* 28:61–90.

King, G., O. Rosen, and M. A. Tanner, Eds. 2004. *Ecological Inference. New Methodological Strategies*. New York: Cambridge University Press.

Klein, J.M. 2019. *Estimation of Voter Transitions in Multi-Party Systems. Quality of Credible Intervals in (hybrid) Multinomial-Dirichlet Models*. Master Thesis Dissertation. Ludwig-Maximilians-Universität München.

Klima, A., P. W. Thurner, C. Molnar, T. Schlesinger, and H. Küchenhoff. 2016. “Estimation of voter transitions based on ecological inference: an empirical assessment of different approaches.” *AStA - Advances in Statistical Analysis* 100:133–159.

Klima, A., T. Schlesinger, P. W. Thurner, and H. Küchenhoff. 2019. “Combining Aggregate Data and Exit Polls for the Estimation of Voter Transitions.” *Sociological Methods & Research* 48:296–325.

Olivia L., O. R. T. Moore and M. Kellermann. 2018. *eiPack: Ecological Inference and Higher-Dimension Data Management*. R package version 0.1-8. https://CRAN.R-project.org/package=eiPack

McCarthy, C., and M. R. Terence 1977. “Estimates of Voter Transition Probabilities from the British General Elections of 1974.” *Journal of the Royal Statistical Society. Series A* 140:78–85.

Miller, W. L. 1972. “Measures of Electoral Change using Aggregate Data.” *Journal of the Royal Statistical Society, Series A* 135:122–142.

Núñez, L. 2016. “Expressive and Strategic Behavior in Legislative Elections in Argentina.” *Political Behavior* 38(4):899–920.

Payne, C., P. Brown, and V. Hanna. 1986. “By-election Exit Polls.” *Electoral Studies* 5:277–287.

Park, W.-h. 2008. *Ecological Inference and Aggregate Analysis of Elections*. PhD Dissertation. The University of Michigan.

Pavía, J.M. and Aybar, C. 2020. “La Movilidad Electoral en las Elecciones 2019 en la Comunitat Valenciana.” *Debats* 134(1):27–51.

Pavía J. M, B. Larraz and J. M. Montero. 2008. “Election Forecasts Using Spatiotemporal Models.” *Journal of the American Statistical Association* 103:1050–1059.

Pavía J. M. and A. López-Quilez. 2013. “Spatial Vote Redistribution in Redrawn Polling Units.” *Journal of the Royal Statistical Society, Series A* 176:655-678

Pavía, J. M., E. Badal, and B. García-Cárceles. 2016. “Spanish exit polls: Sampling error or nonresponse bias?” *Revista Internacional de Sociología* 74(3):e043.

Pavía, J. M., A. Bodoque, and J. Martín. 2016. “The birth of a new party: Podemos, a hurricane in the Spanish crisis of trust.” *Open Journal of Social Sciences* 4:67–86.

Pavía J. M. and E. Veres-Ferrer. 2016a. “Un nuevo estimador para disgregar totales poblacionales. El caso de los nuevos electores.” *Anales de Economía Aplicada* XXX:817–826.

Pavía J. M. and E. Veres-Ferrer. 2016b. “Desagregando Estadísticas de Población.” In Herrerías, JM and Callejón J (eds.), *Investigaciones en Métodos Cuantitativos para la Economía y la Empresa*, Editorial Universidad de Granada, pp. 543-555.

Pavía J. M. and I. Cantarino I. 2017. “Dasymetric distribution of votes in a dense city.” *Applied Geography* 86:22–31.

Pavía-Miralles, J. M. 2005. “Forecasts from Non-Random Samples: The Election Night Case.” *Journal of the American Statistical Association* 100:1113–1122.

Piedras de Papel. 2015. *Aragón es Nuestro Ohio. Así Votan los Españoles*. Madrid: El Hombre del Tr3s.

Pons, V. 2017. “Comment expliquer les transferts de voix du premier au second tour?” Le Figaro, mercredi 17 mai 2017, 13.

Plescia, C. and De Sio, L. 2018. “An evaluation of the performance and suitability of RxC methods for ecological inference with known true values.” *Quality and Quantity* 52:669–683.

Puig, X., and J. Ginebra. 2015. “Ecological Inference and Spatial Variation of Individual Behavior: National Divide and Elections in Catalonia.” *Geographical Analysis* 47(3):262–283.

Robinson, W. S. 1950. “Ecological Correlations and the Behavior of Individuals”. *American Sociological Review* 15(3):351–357.

Romero, R. 2014. “Un modelo matemático para estimar el trasvase de votos entre partidos.” *Revista Digital de la Real Academia de Cultura Valenciana*, 3–23.

Romero, R. 2015. “Trasvase de votos entre partidos en las elecciones autonómicas catalanas del 27 de septiembre de 2015.” *Revista Digital de la Real Academia de Cultura Valenciana*, 3–15.

Romero, R. 2016. “Movilidad electoral entre las elecciones del 20D y del 26J en las comunidades autónomas valenciana, madrileña y andaluza.” *Revista Digital de la Real Academia de Cultura Valenciana. Segunda Época* 1:1–25.

Rosen, O., W. Jiang, G. King, and M. A. Tanner. 2001. “Bayesian and Frequentist Inference for Ecological Inference: The RxC Case.” *Statistica Neerlandica* 55:134–56.

Snelling, C. J. 2016. “Young People and Electoral Registration in the UK: Examining Local Activities to Maximise Youth Registration.” *Parliamentary Affairs* 69(3):663–685.

Tziafetas, G. 1986. “Estimation of the Voter Transition Matrix.” *Optimization* 17: 275–279.

Upton, G. J. G. 1978. “A Note on the Estimation of Voter Transition Probabilities.” *Journal of the Royal Statistical Society. Series A* 141:507–512.

van der Ploeg, C. 2008. *A Comparison of Different Estimation Methods of Voting Transitions with an Application in the Dutch National Elections*. Centraal Bureau voor de Statistiek.

Vázquez, E., and R. Romero. 2001. “Modelos para el estudio del cambio electoral.” *Actas del XXVI Congreso Nacional de Estadística e Investigación Operativa*. Jaen, Spain.

Wakefield, J. 2004. “Ecological Inference for 2x2 Tables (with discussion).” *Journal of Royal Statistical Society, Series A* 167:385–445.