Improving estimates accuracy of voter transitions. Two new algorithms for ecological inference based on linear programming¹

Pavía, Jose M.

GIPEyOP, UMICCS, Universitat de Valencia, email: pavia@uv.es

Romero, Rafael

Universidad Politécnica de Valencia, email: rromero@eio.upv.es

Abstract. The estimation of RxC ecological inference contingency tables from aggregate data defines one of the most salient and challenging problems in the field of quantitative social sciences, with major solutions proposed from both the ecological regression and the mathematical programming frameworks. In recent decades, there has been a drive to find solutions stemming from the former, with the latter being less active. From the mathematical programming framework, this paper suggests a new direction for tackling this problem. For the first time in the literature, a procedure based on linear programming is proposed to attain estimates of local contingency tables. Based on this and the homogeneity hypothesis, we suggest two new ecological inference algorithms. These two new algorithms represent an important step forward in the ecological inference mathematical programming literature. In addition to generating estimates for local ecological inference contingency tables and amending the tendency to produce extreme transfer probability estimates previously observed in other mathematical programming procedures, they prove to be quite competitive and more accurate than the current linear programming baseline algorithm. The new algorithms place the linear programming approach once again in a prominent position in the ecological inference toolkit. We use a unique dataset with almost 500 elections, where the real transfer matrices are known, to assess their accuracy. Interested readers can easily use these new algorithms with the aid of the R package `lphom`.

Keywords. RxC contingency tables; Mathematical programming; Transfer probabilities; Split-ticket voting; `lphom`.

1. Introduction

Attempting to estimate vote transfers between elections using exclusively the aggregate results from voting units presents a challenge which dates back to the 1960s (Vangrevelinghe, 1961; Hawkes, 1969; Irwin and Meeter, 1969). This problem is just a specific case of a more general problem that came to light in the early part of the 20th century (e.g., Ogburn and Goltra, 1919; Ogburn and Talbot, 1929; Gosnell and Gill, 1935; Gosnell and Schmidt, 1936): how to ascertain voting outcomes for certain subgroups using data from precincts or counties. In general, the process of deducing individual behaviours from aggregated data is called ecological inference, which is exposed to what is known as the ecological fallacy (Robinson, 1950).

¹ The authors wish to thank Carolina Plescia for providing us with the electoral Scottish data handled in this paper. Thanks are also due to M. Hodkinson for revising the English of the paper. This 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 number ECO2017-87245-R]; and by Conselleria d'Innovació, Universitats, Ciència i Societat Digital, Generalitat Valenciana [grant number AICO/2019/053].

Within the ecological inference literature, the problem is usually stated as a two-way contingency table where the goal is to infer the unknown inner-cell values from the known margins. To infer how the collectives defined by the row-options (who are grouped according to some variable, such as race, religion, age, gender or previous electoral behaviour) split (vote) among the column-options. This is an ill-posed problem as many sets of substantively different inner-cell counts are consistent with a given marginal table, giving rise to concerns over identifiability and indeterminacy.

To estimate the internal cells, the marginal totals of I equivalent tables corresponding to the territorial units in which the whole population is divided out are used as data. This, however, does not solve the problem, but multiplies it by a factor of I . Now, instead of one table, we have I tables, each with their own interior cells. In order to overcome this issue, a basic hypothesis of homogeneity is routinely introduced to learn from the margin cross-unit statistical covariations. Whatever the approach, it is considered that row fractions or transition probabilities of (subgroups of) contingency tables of the different territorial units are in some way similar/related (Imai et al. 2008; Greiner and Quinn, 2009; Forcina and Pellegrino, 2019).

Based on this hypothesis many algorithms, grounded on different philosophical foundations and/or employing different mathematical approaches, can be found in the literature for estimating row fractions or row-conditional (underlying) probabilities. These include, among others, procedures from frameworks as diverse as Bayesian and frequentist statistics, mathematical programming or information theory.

Following the seminal papers of Goodman (1953, 1959) and Duncan and Davis (1953), the statistic framework has been the one most prolifically used, mainly after King (1997), who masterfully combined Goodman's regression and Duncan and Davies's method of bounds. Indeed, after the publication of King's book "A solution to the ecological inference problem", we have assisted to a resurgence of proposals within the so-called ecological regression approach, many of the earlier ones designed for dealing with 2x2 tables and later generalized to solving problems of RxC tables (e.g., King et al., 1999; Rosen et al., 2001). Within this framework, there are methods that explicitly model the spatial dimension of the data (e.g., Haneuse and Wakefield, 2004; Puig and Ginebra, 2015), that combine precinct aggregated data and exit polls (e.g., Greiner and Quinn, 2010; Klima et al., 2019) or that even mix both sources of information (Imai and Khanna, 2016). Interested readers in this approach can consult King et al. (2004) and Wakefield (2004), who offer some overviews, and Klima et al. (2016) and Plescia and De Sio (2018), who carry out a broad assessment of procedures.

The other major route followed by researches to deal with ecological inference has been to use mathematical programming. In this setting, deterministic bounds are incorporated in a natural way via exact and inequality constraints. The proposals within this framework, which have been almost exclusively focused on inferring voter transitions, can be traced back to Irwin and Meeter (1969) and McCarthy and Ryan (1977), who consider quadratic programming algorithms. Later, Tziafetas (1986) shows linear approaches being more efficient and Corominas et al. (2015) extend the number of possible discrepancy functions. This literature has been less prolific, with significantly less papers published, so many issues linked to the mathematical programming solutions are still to be resolved.

Romero et al. (2020) tackle two of these issues in a recent paper. They extend linear programming to explicitly deal with new entries and exits in the election censuses without assuming unrealistic hypotheses, and, as a main contribution, they develop a procedure to

measure the uncertainty of the estimates. They call their algorithm *lphom* after "Linear Programming based on HOMogeneity". We continue in the investigative direction taken by Romero et al. (2020) and, in this paper, we come up with solutions to two other, more important and as yet unresolved, issues within the mathematical programming framework: the estimation of local transition matrices and the excess of extreme estimated probabilities.

One of the limitations of current mathematical programming algorithms is that they only generate estimates for the joint cross-table distribution of the area under investigation as a whole. They do not provide inferences about the cross-tabulations for the tables of the different voting units in which the whole population is split out. Likewise, mathematical programming algorithms have been rightly criticized (e.g., Upton, 1978; Johnston and Hay, 1983) as tending to produce many extreme probabilities or fractions: zeros and ones. In this research we propose solutions to both these questions. It should be noted that these questions are not a current concern in the ecological regression literature. The first limitation is solved by the most developed ecological regression approaches which, moreover, do not suffer from the second weakness.

First, we suggest a novel procedure, based on linear programming and grounded on the homogeneity hypothesis, to estimate the inner-cells values at the local level. We call this procedure *lphom_local*. We then build on this two new algorithms (which we call *tslphom* and *nslphom*) to produce estimates at both local and global levels. These new algorithms overcome the problem of extreme values and systematically outperform *lphom*. As we show later in this paper, *tslphom* and *nslphom* produce estimates significantly more accurate than the ones generated by *lphom*. Using real data from almost 500 elections for which the actual cross-table corresponding to the whole territory is known, we see that *tslphom* systematically beats *lphom* and that, likewise, *nslphom* consistently beats *tslphom*. Moreover, in an independent study (Pavía and Romero, 2021b), we also show that *nslphom* produces, with less computational cost and in a simpler way, estimates at least as accurate as the ones attained by the statistical approach currently identified as the best in the literature (see, Klima et al. 2016 and Plescia and De Sio, 2018). In our view, these results, in addition to the capacity of the new algorithms to produce local solutions, place the linear programming approach once again in a prominent position in the ecological inference toolkit.

The fact that the new algorithms also equal the most developed ecological regression approaches in their capacity for generating local (precinct or polling station) transition matrices is relevant. It has multiple implications for historical analysis and for future elections. For example, in the latter case, local estimates could be used for micro-targeting and for defining marketing campaign strategies. Based on the analysis of polling station estimates of voting transfers between two previous elections (for instance, the last national and regional elections), party committees could decide where and which voters to target (for instance, during the next local or national elections) and, in knowing their past behaviour, which arguments to use to persuade them.

The rest of the paper is structured as follows. Section 2 briefly describes the *lphom* algorithm. Section 3 states our solution to estimate local contingency tables. The *tslphom* algorithm is introduced in Section 4, while Section 5 deals with the *nslphom* algorithm. Section 6 presents the data and the results obtained after comparing *lphom*, *tslphom* and *nslphom* solutions. Section 7 discusses the findings and suggests directions for further research. Finally, Section 8 summarises and concludes.

2. The baseline model: lphom

Without loss of generality and for the sake of convenience, from here on in the paper, we follow the terminology used in Romero et al. (2020) and consider that we are dealing with the problem of estimating the matrix of transfer of votes between two election processes. In particular, in the model stated by Romero and colleagues, which they call lphom, it is assumed that the aggregated results of I territorial units in which the electoral space is broken down are known and that J and K are the number of voting options in the elections E1 and E2, respectively. In both cases, abstention is considered as a possible voting option.

The data of the model are, for each of the $i = 1, \dots, I$ voting units, the votes x_{ij} recorded for the $j = 1, \dots, J$ election options available in E1 and the votes y_{ik} ($k = 1, \dots, K$) harvested by the different competing options in E2. The basic variates of the model are the $J \times K$ unknowns p_{jk} , each one defined as the proportion of voters in the entire electoral space who, having chosen option j in E1, have chosen option k in E2. That is, $p_{jk} = \sum_{i=1}^I p_{jk}^i \omega_j^i$ is a (weighted) average of the transfer proportions in the voting units, where p_{jk}^i is the proportion of voters in unit i who, having chosen option j in E1, have chosen option k in E2 and $\omega_j^i = x_{ij} / \sum_{i'=1}^I x_{i'j}$. According to this definition, the p_{jk} must meet the following constraints:

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

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

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

The above system has more unknowns than data. Hence, to deal with the indeterminacy, lphom introduces the hypothesis of homogeneity of electoral behaviour in the I units. In particular, the homogeneity hypothesis establishes that the unit vote transfer probabilities, p_{jk}^i , are *similar* to the average probabilities, p_{jk} , of the entire territory and that, consequently, the observed values y_{ik} must differ *little* from those values that would be obtained by applying to x_{ij} the average probabilities. Naming e_{ik} as these discrepancies (see equation (4)), we have that the e_{ik} should be *small*.

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

The basic lphom algorithm is a linear program by means of which one obtains the p_{jk} values that satisfying the four previous sets of constraints minimize (5), the sum of the absolute values of the e_{ik} .

$$\text{minimize } \sum_{i,k} |e_{ik}| \quad (5)$$

For equations (1), (2) and (3) to be compatible, it is necessary that the sums of the rows of the matrices $I \times J$ and $I \times K$ defined, respectively, as the row vector matrices $[x_{ij}]_{i=1}^I$ and $[y_{ik}]_{i=1}^I$ match exactly. This forces the analyst to explicitly include the changes in the electoral censuses

between the two elections, when they exist. There are no changes when E1 and E2 are simultaneous elections with the same election censuses (for instance, when each voter casts two votes, one for a party list and another for a candidate) and they are irrelevant when the two electoral processes are very close in time. In this latter case, the entries and exits in the census lists tend to be negligible and could be added, for instance, to the abstention without impacting in practice on the proportion estimates.

In general, entries in each unit are the sum of two groups: young people who join the census because they have reached the minimum age to vote between the dates of the two elections and new residents (immigrants) who have the right to vote. On the other hand, exits are made up of two groups: voters registered in E1 who have died before E2 and people who have emigrated out of the unit in the inter-election period.

Depending on the information available for entries and exits, different constraints have to be added to the basic model. The `lphom` algorithm programmed in the R function available in `lphom` package (Pavía and Romero, 2021a), considers all the possible scenarios. In the less-demanding (and quite common) information scenario, aggregated entries are treated as a possible source of votes and denoted as option J in E1, while aggregated exits are considered as a possible destination of votes and denoted as option K in E2. In this case, `lphom` assumes that census exits impact the first $J - 1$ options of E1 in a similar (relative uniform) way, therefore, together with the obvious constraint (7), it adds the additional constraints defined by (6).

$$p_{jK} = \left(\sum_{i=1}^I y_{iK} \right) / \left(\sum_{j=1}^{J-1} \sum_{i=1}^I x_{ij} \right) \quad j = 1, \dots, J - 1 \quad (6)$$

$$p_{JK} = 0 \quad (7)$$

3. Estimating voter transitions at the local level

The `lphom` algorithm estimates the matrix $[p_{jk}]$ of voting transfer probabilities between the options of two elections E1 and E2 of the area under investigation as a whole. Often, however, the estimation of the matrices $[p_{jk}^i]$ of voting transfer probabilities in the different voting units is also of interest. To this end, in this section we propose a new procedure, which we call `lphom_local`. This new procedure is consistent with the hypothesis of homogeneity of electoral behaviour on which `lphom` rests.

The data that `lphom_local` requires are the row-vector matrices $[x_{ij}]_{i=1}^I$ and $[y_{ik}]_{i=1}^I$, introduced in the previous section, and a global matrix $[p_{jk}^G]$ of transfer probabilities for the whole territory. This matrix could be obtained using, for instance, `lphom`. The unknowns of this new model are the $[p_{jk}^i]$ ($i = 1, \dots, I$) matrices, whose generic (j, k, i) -element denotes, for each unit i , the proportion of voters in unit i who, having chosen option j in E1, choose option k in E2. According to this definition the proportions p_{jk}^i must fulfil the following constraints:

$$p_{jk}^i \geq 0 \quad \text{for } j = 1, \dots, J \quad k = 1, \dots, K \quad i = 1, \dots, I \quad (8)$$

$$\sum_{k=1}^K p_{jk}^i = 1 \quad \text{for } j = 1, \dots, J \quad i = 1, \dots, I \quad (9)$$

$$\sum_{j=1}^J x_{ij} p_{jk}^i = y_{ik} \quad \text{for } k = 1, \dots, K \quad i = 1, \dots, I \quad (10)$$

As in the lphom model, this system of equations sets up an indeterminate system (being more precise, I indeterminate systems, one per each unit), which calls for new constraints to be included in the model in order to solve it. The homogeneity hypothesis stated in the previous section postulates that the p_{jk}^i are *similar* to the corresponding global p_{jk}^G . Thus, under this hypothesis, the estimate of the volume of voters v_{jk}^i in unit i that pass from voting option j in E1 to option k in E2 should differ *little* when it is estimated applying either p_{jk}^i or p_{jk}^G . Thus, the quantities ε_{jk}^i defined by equation (11) should be *small*.

$$\varepsilon_{jk}^i = x_{ij} p_{jk}^i - x_{ij} p_{jk}^G \quad \text{for } j = 1, \dots, J \quad k = 1, \dots, K \quad i = 1, \dots, I \quad (11)$$

The first step of our lphom_local procedure solves I linear programs, one for each voting unit i ($i = 1, \dots, I$), and estimates the p_{jk}^i as the values that satisfying the sets of constraints (8), (9), (10) and (11) minimize the sum of the absolute values of $\sum_{j,k} |\varepsilon_{jk}^i|$.

$$\text{minimize } Z = \sum_{j,k} |\varepsilon_{jk}^i| \quad \text{for } i = 1, \dots, I \quad (12)$$

As with lphom, lphom_local must satisfy, regarding entries and exits, the restrictions imposed in each unit i by the current scenario. In particular, if the last columns J and K of the matrices $[x_{ij}]_{i=1}^I$ and $[y_{ik}]_{i=1}^I$ correspond, respectively, to entries and exists, lphom_local includes the additional constraints given by equations (13) and (14).

$$p_{jK}^i = y_{iK} / \left(\sum_{j=1}^{J-1} x_{ij} \right) \quad \text{for } j = 1, \dots, J-1 \quad i = 1, \dots, I \quad (13)$$

$$p_{jK}^i = 0 \quad \text{for } i = 1, \dots, I \quad (14)$$

Equation (13) constraints translate the hypothesis that, in each unit, exits impact on a similar relative way to the $J-1$ options of election E1, while equation (14) sets down that the transfer of votes between entries and exits is, obviously, null.

Regardless of whether equations (13) and (14) are or are not added to the linear program system defined by equations (8)–(12), if p_{jk}^G verifies (3), the above system is indeterminate in the sense that an infinite set of substantively different $[p_{jk}^i]$ matrices fulfil all the equations of constraints and minimize (12). We have indeed confirmed that, under these circumstances, different solutions for the linear programs can be found scoring exactly the same optimal values in (12). An example of the impact of this is shown in Section 2S of the Supplemental Material.

In order to overcome the indeterminacy, we turn to the hypothesis of homogeneity. For each i , we suggest selecting, among those matrices minimizing (12) and fulfilling all the restrictions, the matrix $[p_{jk}^i]$ closest to the global matrix $[p_{jk}^G]$. In particular, we propose adding to the above linear program two new equations, (15) and (16), for each i and to minimize, in a second step, equation (17) subject to the constraints defined by equations (8)–(12) and (15) and (16) and, depending on the scenario, also equations (13) and (14).

$$Z = \sum_{j,k} |\varepsilon_{jk}^i| \quad \text{for } i = 1, \dots, I \quad (15)$$

$$p_{jk}^i = p_{jk}^G + \delta_{jk}^i \quad \text{for } j = 1, \dots, J \quad k = 1, \dots, K \quad i = 1, \dots, I \quad (16)$$

$$\text{minimize } \sum_{j,k} |\delta_{jk}^i| \quad \text{for } i = 1, \dots, I \quad (17)$$

Our proposal, *lphom_local*, to estimate voting transfer matrices in each unit is therefore a two-step procedure where, in a first step, the set of potential solutions is delimited to subsequently, in a second step, choose the matrix closest to the reference global matrix as the final solution.

Note that when $x_{ij} = 0$ for a given (i, j) -pair, whatever set of proportions $\{p_{jk}^i\}_{k=1}^K$ will verify the constraints (11). This is also true for the j -row of the global proportions, which will be the solutions of the two linear systems. Once proportions are transformed into votes, this has no effect as they are multiplied by zero. Nevertheless, we recommend forcing these proportions to be zero in the final solution.

4. Improving *lphom*: *tslphom*

4.1. Introduction.

Some authors (e.g., Upton, 1978; Johnston and Hay, 1983; Corominas et al., 2015) have pointed out that *lphom* has an excessive tendency to include p_{jk} estimates equal to 1 in its solution, which obviously forces the remaining row proportions, p_{jk^*} , for $k^* \neq k$, to take null values. In our opinion, this phenomenon is a natural consequence of the methodology used, since the optimal solution of a linear program is always an extreme point of the convex hull of the region of feasible solutions defined by its constraints. In the *lphom* model, constraints (1) and (2) generate many vertices with one or more p_{jk} equal to 1, which results in a relatively high probability of one of these vertices being in the optimal solution.

The *tslphom* algorithm, presented in the next subsection, was initially viewed by the authors as a way of alleviating the problem that the *lphom* algorithm has of the excessive number of p_{jk} equal to 1 and also with the expectation that it could even improve *lphom* by constructing a global solution as an aggregation of local solutions. The first issue is quickly and easily observed (see Table 4) and, as we show later in this paper, we also confirm that *tslphom* provides solutions with lower error than *lphom*.

4.2 The *tslphom* algorithm.

The name *tslphom*, which we propose for the new algorithm, is an acronym for "**Two Steps lphom**" and refers to the fact that in the process of estimating the final global matrix, $\mathbf{P} = [p_{jk}]$, of vote transition probabilities, the matrix $\mathbf{P}$ is obtained twice. The *tslphom* algorithm works as follows:

1. In a first step, given the data $[x_{ij}]_{i=1}^I$ and $[y_{ik}]_{i=1}^I$, a solution matrix $\hat{\mathbf{P}}_o$ is obtained by applying the *lphom* procedure as stated in section 2.
2. Next, using $\hat{\mathbf{P}}_o$ as the reference matrix of global transition probabilities, the *lphom_local* procedure proposed in section 3 is applied to obtain estimates of the matrices $\mathbf{V}_i = [v_{jk}^i]$ of vote transition in the I territorial units.

3. Finally, the $\hat{\mathbf{V}}_i = [\hat{v}_{jk}^i]$ matrices estimated in the previous step are aggregated to obtain a global vote transition matrix. The tslphom global estimated matrix of transition probabilities, $\hat{\mathbf{P}}_1 = [\hat{p}_{jk}]$, is calculated from this.

This operative will clearly decrease the number of p_{jk} equal to 1 in the final solution, since these will only appear in the event that the corresponding p_{jk}^i in the I territorial units are all equal to 1.

4.3. A measure to quantify the homogeneity hypothesis.

Given that both lphom and tslphom are based on the hypothesis of homogeneity of the electoral behaviour of the I territorial units, it is important to measure in each specific study the degree of non-compliance of this hypothesis with the achieved solution. According to Romero et al. (2020) this degree of non-compliance is quantified using the HET heterogeneity index, defined by equation (18).

$$HET = 100 \cdot \frac{0.5 \sum_{ijk} |v_{jk}^i - x_{ij} p_{jk}|}{\sum_{ij} x_{ij}} \quad (18)$$

In equation (18), the v_{jk}^i are the elements of the vote transition matrices in the I territorial units and the p_{jk} are the global transition probabilities. Although lphom obtains estimates of the latter quantities, the v_{jk}^i values still remain unknown with this algorithm, so the plug-in principle cannot be applied to estimate the HET heterogeneity index when lphom is used. In Romero et al. (2020) an estimate of the heterogeneity index, called HETe, is proposed based on the e_{ik} residuals of the lphom model, which are clearly outputs of lphom.

Estimates of v_{jk}^i , however, are obtained when we work with the tslphom algorithm. In this case, therefore, it is possible to obtain an estimate of the index of heterogeneity, which we will also call HETe, applying the plug-in principle. HETe in this case is obtained by replacing in (18) v_{jk}^i by $\hat{v}_{jk}^i$ and p_{jk} by $\hat{p}_{jk}$. This estimated heterogeneity index HETe will play an important role when studying the stopping criteria of the nslphom algorithm that we propose in the next section.

5. Extending tslphom: nslphom

5.1. From two steps to n steps: nslphom.

The algorithm tslphom reaches its solution after obtaining two sequential estimates of the global probability transition matrix. Hence, it is obvious and a logical consequence to consider the idea of extending tslphom by iterating steps two and three of tslphom up until reaching convergence. The proposal would be to perform an iteration process of re-estimating the matrix of global transition probabilities through lphom_local using in each iteration as global matrix, $\mathbf{P}^G$, the last attained transition probability matrix, and to stop the process when the matrix $\mathbf{P}^G$ does not vary more than a given threshold in two consecutive iterations. The initial reasonableness of this algorithm is reinforced by the fact that, as mentioned in the previous section and shown in section 6 using cases where the actual probability transition matrices are known, the solutions attained by tslphom are as a rule more accurate than those achieved with lphom.

It would be reasonable to consider that after a sufficient number of iterations the results provided by `nsiphom` would tend to stabilize around a solution that, in a sense, would be the best possible solution. However, this is not what really happens as we have verified with hundreds of elections. As we show in the next subsection, the solutions attained with this tentative algorithm do not converge but, at best, tend to oscillate around some reasonable attraction point. We discover that the process improves the estimates during the first steps, up to a certain point, after which it has less effect, even worsening the results in some cases.

In this section, we define a new algorithm, which we call `nsiphom` (as acronym of "**N** Steps **I**phom"), where in order to attain a solution we iterate steps two and three of `tsiphom` for a limited number of times. Hence, the critical point to define `nsiphom` lies in determining an optimal number of iterations or a proper stopping rule. This is the topic of subsection 5.2. In subsection 5.3 we propose two basic versions of `nsiphom` based on what we learn in subsection 5.2.

5.2. How many steps? Defining a stopping rule.

To show how estimates do not converge as iterations grow, we analyse the sequence of estimates provided by `nsiphom` as a function of the number of iterations for a particular election. As a case study, we consider the estimation of the vote transfers between the first and second rounds of the 2017 French presidential election using as inputs (i) the outcomes recorded in the 107 territorial departments in which the territory of France is divided plus (ii) the results tallied for the French electors living abroad, grouped in an artificial department. In order to make the estimation process simpler, entries and exits between both rounds (which are negligible) have been added to abstainers.

We focus on analysing the behaviour of just one of the p_{jk} : $p_{M,M}$, which represents the proportion of voters who, having voted for Macron in the first round, continue to vote for him in the second round. The evolution of these proportions will be linked with the evolution of the HETe statistic.

Figure 1. Evolution of `nsiphom` solution for $p_{M,M}$ as a function of the number of iterations. In the left panel, the dashed-green and dotted-purple lines identify, respectively, the iterations in which HETe reaches its first minimum and its global minimum. In the right panel, the lines identify, respectively, the corresponding estimates for $p_{M,M}$.

Figure 1 shows the evolution of the estimates obtained by `nsiphom` for $p_{M,M}$ as a function of the number of iterations: in the left-panel from iteration 0 to iteration 100 and in the right-panel up to iteration 4000. It seems reasonable to assume that the true value of $p_{M,M}$ should be very high (close to one). In fact, the solution obtained by `lphom` resulted in $p_{M,M} = 1$. Figure 1 shows that, even after several thousand iterations, $p_{M,M}$ does not stabilize and, more importantly, that all the estimated values look reasonable and they show relatively small variations after the first iterations. Hence, given that when we build the model `nsiphom`, we rely on the homogeneity hypothesis, in our opinion, it seems reasonable that for defining a stopping rule we consider the evolution of the estimated heterogeneity index, `HETe`, presented in subsection 4.3.

Indeed, as Romero et al. (2020) already found for `lphom`, a clear positive correlation links the heterogeneity index associated with an electoral process and the error rate of the corresponding attached solution. The issue, therefore, is to decide how to translate this relationship to an operable rule. From the computation point of view, this will not pose any special difficulty as, in each iteration, together with the new solution, we can also calculate the `HETe` statistic. On the other hand, from the judgement point of view, we can exploit the pattern observed in Figure 2, which we have observed (with some variations) in all the elections we have analysed.

Figure 2. Evolution of `HETe` in `nsiphom` as a function of the number of iterations. In the left panel, the dashed-green and dotted-purple lines identify, respectively, the iterations in which `HETe` reaches its first minimum and its global minimum. In the right panel, the lines identify, respectively, the corresponding estimates for `HETe`.

Figure 2 shows the evolution of `HETe` when `nsiphom` is applied to the study of the 2017 French presidential elections (in the left-panel from iteration 1 to iteration 100 and in the right-panel up to iteration 4000). In the case of the example displayed in Figure 2, the `HETe` index decreases during the first eight iterations and reaches its global minimum in the twelfth iteration, to afterwards consistently begin to grow. Indeed, in almost half of the elections that we have analysed, we have found a pattern for the evolution of `HETe` equal to the one observed in Figure 2: the iteration corresponding to first local minimum does not match with the iteration corresponding to the global minimum. In the other half, the first local minimum, which is easily detected as the first iteration from which the `HETe` begins to grow, is also the global minimum (for any number, ns , of steps). Nevertheless, in all the cases, the first local and global minimums for `HETe` are found after very few iterations. As a rule, we have found that the `HETe` sequence consistently decreases in the first steps to subsequently (maybe after a period of some relative stabilization) start to grow.

In light of these results, we envisage two reasonable strategies for the `nslphom` algorithm to produce a solution. On the one hand, a reasonable stopping criterion for `nslphom` is to end the process at the first iteration in which HETe starts to grow and to take as solution the vote transfer matrix attained in the previous iteration. From now on, we will name `nslphom` with this criterion `ns_first`. This is equivalent to the use of the `nslphom` R function of `lphom` with the argument `min.first = T`. On the other hand, an alternative solution is reached by choosing the matrix that corresponds to the minimum value obtained for HETe after running `nslphom` with ns iterations, where ns is a value set in advance. With this second strategy, which we will call `ns_number` (where *number* is equal to the value ns set in advance) the question turns to how to set ns . This specification is equivalent to the use of the `nslphom` R function with the arguments `min.first = F` and `max.iter = ns`.

It is obvious that the higher the value set for ns , the greater the probability that the minimum HETe obtained corresponds to the minimum possible value HETe for the election at hand. Taking a larger ns , however, has two important drawbacks. On the one hand, the computational burden grows with ns , for any given election. On the other hand, as ns grows sometimes the solutions slight deteriorate, even ending up with smaller HETe. Hence, as a compromise solution, a reasonable specification for `nslphom` with this second version is to run `nslphom` with a relatively small number of iterations.

In section 6, we capitalise on having a large number of electoral processes in which the real transfer matrices are known to assess the accuracy (and computational costs) of `lphom`, `tslphom` and `nslphom`; with `nslphom` parametrized with different specifications: `ns_first`, `ns_10`, `ns_25`, `ns_50` and `ns_100`.

5.3. The `nslphom` algorithm.

Having determined two reasonable criteria to obtain estimates using the `nslphom` algorithm, this subsection describes exactly how `nslphom` works. Table 1 details the pseudo codes associated with each one of the two specifications for the `nslphom` algorithm introduced in subsection 5.2.

Table 1. Pseudo codes with the proposed stopping rules for `nslphom` algorithm.

nslphom algorithm 1. Pseudo code with stopping rule at the observed HETe first minimum.
<ol style="list-style-type: none"> 0. Let $\mathbf{X} = [x_{ij}]_{i=1}^I$ and $\mathbf{Y} = [y_{ik}]_{i=1}^I$ be the row-vector matrices of votes recorded in, respectively, E1 and E2 in the I voting units. 1. Estimate $\hat{\mathbf{P}}_0$ by applying the <code>lphom</code> algorithm to $\mathbf{X}$ and $\mathbf{Y}$. Assign $\mathbf{P}^G \leftarrow \hat{\mathbf{P}}_0, l \leftarrow 1, HETe_0 \leftarrow \infty$. 2. Estimate $\hat{\mathbf{P}}_l$ and $HETe_l$ by applying the <code>lphom_local</code> procedure to $\mathbf{X}, \mathbf{Y}$ and $\mathbf{P}^G$. Assign $\mathbf{P}^G \leftarrow \hat{\mathbf{P}}_l, l \leftarrow l + 1$. 3. If $HETe_l > HETe_{l-1}$ stop; otherwise come back to 2. 4. Select as solution $\hat{\mathbf{P}}_{l-1}$.
nslphom algorithm 2. Pseudo code with the number of iterations set in advance.
<ol style="list-style-type: none"> 0. Let ns be the maximum number of iterations to be performed and let $\mathbf{X} = [x_{ij}]_{i=1}^I$ and $\mathbf{Y} = [y_{ik}]_{i=1}^I$ be the row-vector matrices of votes recorded in, respectively, E1 and E2 in the I voting units. 1. Estimate $\hat{\mathbf{P}}_0$ by applying the <code>lphom</code> algorithm to $\mathbf{X}$ and $\mathbf{Y}$. Assign $\mathbf{P}^G \leftarrow \hat{\mathbf{P}}_0, l \leftarrow 1$. 2. Estimate $\hat{\mathbf{P}}_l$ and $HETe_l$ by applying the <code>lphom_local</code> procedure to $\mathbf{X}, \mathbf{Y}$ and $\mathbf{P}^G$. Assign $\mathbf{P}^G \leftarrow \hat{\mathbf{P}}_l, l \leftarrow l + 1$. 3. If $l > ns$ stop; otherwise come back to 2. 4. Select as solution the $\hat{\mathbf{P}}_{l^*}$ whose $HETe_{l^*}$ is minimum for $1 \leq l^* \leq ns$.

6. Assessing the accuracy of *lphom*, *tslphom* and *nslphom*

6.1 Introduction.

In the previous sections, two new algorithms, *tslphom* and *nslphom*, have been introduced as alternatives to *lphom*. These two new procedures reduce the chances of producing matrix solutions with extreme transition probabilities, a tendency usually observed as a weakness of mathematical programming procedures. This section aims to assess whether, in addition to this advantage, these two new procedures also provide more accurate results, that is, outcomes closer to the actual transition matrices. In the case of *nslphom*, we also evaluate what configuration (stopping rule) is more convenient in terms of accuracy and computational burden.

The main difficulty of performing these evaluations lies in the fact that actual transition matrices are as a rule unknown. Except in very special circumstances direct comparisons are impossible. Hence, in the literature, different strategies have been carried out to gauge ecological inference solutions. We can find studies where ecological inference transfer matrices are compared to transfer matrices obtained from polls (mainly exit-polls or panel surveys) with the focus on analysing the socio-political soundness of the ecological results attained. In other studies, evaluations are accomplished via simulation exercises when, after setting the *actual* transfer probabilities, some outcomes are simulated for the second election conditioned on the data from the first election. None of these strategies is free from criticism. Polls are exposed to significant sources of bias and generate estimates with large variances. Large doses of subjectivity pervade reasonableness of socio-political outcomes, mainly where there are no substantial differences between the solutions reached using different algorithms. And, sometimes, the conditions defining the scenarios of the simulation exercises are set, even unconsciously, to favour one of the algorithms.

In some circumstances, however, such as in mixed-member election systems in which voters cast two votes simultaneously in the same ballot and they are recorded and made public, it is possible to know the actual transfer matrices. This is the case for the New Zealand general elections since 2002 and, as an exceptional experience, the 2007 Scottish Parliament election. In these cases, the electoral authorities publish/published marginal results at polling station level and split-ticket cross-tables at district level. This offers the unique opportunity of comparing ecological contingency tables, estimated by exploiting marginal results at polling level, with true quantities of interest, available in the observed district cross-tables. In particular, to assess the algorithms we compare the estimated ecological contingency tables and district split-ticket tables corresponding to 493 elections: 420 tables come from the 2002, 2005, 2008, 2011, 2014 and 2017 New Zealand general elections and 73 tables from the 2007 Scottish Parliament election. We describe the data in the next subsection to subsequently, in subsection 6.3, introduce the statistics used to measure the distances between estimated and actual matrices/tables. The findings are presented in subsection 6.4.

6.2. The data

New Zealand elects its parliament members using a mixed-member proportional system and Scotland does so by applying an additional member voting system. Both systems are quite similar. Each voter casts a ballot with two votes, one for a local candidate, which is used to choose the person who will be the parliamentary representative for the local area where the

voter lives, and another one for a regional or national party list. Representatives are elected taking into account both votes. In each local area (called constituency in Scotland and electorate in New Zealand; hereafter, we call them districts), the candidate who receives most votes is automatically elected. The remaining seats are allocated applying a proportional rule to party votes. In New Zealand (NZ), these seats are allocated in a national compensatory fashion. To guarantee that nationwide the share of seats a party wins is about the same as its share of votes, the partisan affiliations of the winners in the electorates are taken into account. In Scotland (SCO), the 73 constituencies in which electors are divided are grouped into regions and regional party votes used to apportion regional seats to parties using a modified D'Hondt rule (Pavía-Miralles, 2005). The idea is also to make the overall result more proportional.

A unique characteristic of the NZ electoral system is that across the country there are a number of seats reserved for the Māori (or people of Māori descent) who choose to enrol on separate lists of electors. The electoral boundaries of the seven Māori districts are superimposed over the electoral boundaries used for regular electorates, covering the whole NZ territory. Every area of New Zealand simultaneously belongs to both a regular district and a Māori district. This means that great differences in terms of number of polling stations and density of voters per polling station exist between regular and Māori districts. Pooling the six NZ general elections considered in this study, we can see that Māori districts have a mean of 325 polling stations per district with an average density of 59 voters per polling station. These figures are quite different in NZ regular districts. Their corresponding averages are 60 polling stations per district and 573 voters per polling station. This introduces a conspicuous variability that significantly enriches our analyses, given that assessing the performance of ecological inference algorithms across different types of contexts adds robustness to the conclusions (Park et al., 2014). Table 1 offers more details about different characteristics of the dataset used to assess the performance of the algorithms. As can be observed, we not only have great variability in terms of the number of polling stations and voters by district but also in terms of the sizes (number of rows and columns) of the analysed contingency tables.

Table 2. Summary of some features of the dataset used to assess the algorithms.

Country	Year	No. of districts (elections)	No. of regular polling stations (min-max)	No. of Māori polling stations (min-max)	No. of parties (min-max)	No. of candidates (min-max)	Average number of cells per table
New Zealand	2002	69	30-118	168-651	5-8	5-8	39.5
	2005	69	29-116	172-698	4-7	3-6	23.8
	2008	70	32-112	174-686	4-6	3-6	23.4
	2011	70	32-113	171-644	4-7	4-6	26.2
	2014	71	31-111	168-620	5-7	3-6	27.9
	2017	71	41-136	188-705	4-7	3-6	24.8
Scotland	2007	73	22-103	-	5-8	5-8	35.2

Source: Compiled by the authors using official data from the New Zealand electoral commission and the Scotland Electoral Office.

The raw cross-distributions of votes at district level (with parties by rows and candidates by columns) of New Zealand as well as the corresponding marginal distributions at polling voting level by parties and candidates were collected, on January 2019, from the official web page of the electoral commission of New Zealand (www.electionresults.org.nz). In the case of Scotland, it was not possible to obtain the corresponding raw figures from the official Scottish electoral

commission. Instead, we are grateful to Carolina Plescia for downloading the raw files from the data on the Scotland Electoral Office website in 2011.

Before starting the process the data were checked for internal consistency and pre-processed in order to guarantee a proper correspondence among the $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{P}$ matrices. In particular, in the case of NZ the following steps were taken. First, the rows with all zero values or non-available were eliminated in the parties and candidates' files. Second, the row corresponding to the polling unit identified as "Votes Allowed for Party Only" was eliminated in the parties' files, given that this voting unit had no equivalent in the candidates' files. Third, two actions were performed in the cross-distribution files. On the one hand, the column labelled "Party Vote Only" was eliminated, because this corresponds to the row "Votes Allowed for Party Only" in the party files and these proportions cannot be estimated as they are not available by voting units (i.e., there is no information about how many ballots are without a vote for a local candidate in each polling unit). On the other hand, the cross-distributions were recomputed in order to guarantee row-standardized matrices, as this property is lost as a consequence of eliminating the "Party Vote Only" column. Finally, in addition to these pre-processing tasks, as is usual practice when dealing with real data (e.g., van der Ploeg, 2008; Klima et al., 2016; Klein, 2019; Plescia and De Sio, 2018; Pavía and Aybar, 2020), very small electoral options were grouped. In both New Zealand and Scottish files, those parties or candidates which individually do not reach at least a 3% of the district vote were grouped in the option 'Others'.

A number of other (almost manual) minor pre-processing tasks were also performed. The most relevant was the collapsing of the voting units "Voting places where less than 6 votes were taken" (row 100) and "Ordinary Votes BEFORE polling day" (row 101) corresponding to the party and candidate files of the 43rd district of the 2014 NZ election (Rangitikei). They were added as a consequence of a mismatch between both files. Their respective aggregations in the parties' and candidates' files are 3 and 2 for the 100th row and 8465 and 8466 for the 101st row.

6.3. Measures of error

After running each algorithm, we have two pairs of two matrices for each election: the real and estimated matrices of votes, $\mathbf{V} = [v_{jk}]$ and $\hat{\mathbf{V}} = [\hat{v}_{jk}]$, and the real and estimated matrices of transition probabilities, $\mathbf{P} = [p_{jk}]$ and $\hat{\mathbf{P}} = [\hat{p}_{jk}]$. We use this to define two discrepancy statistics, EI and EPW, equations (19) and (20), which capture the amount of error associated with the estimates attained with each algorithm. These measures always refer to the global estimates, the matrices for the whole area of study. Analysis of the errors at local level is not possible as real values are not available at this level for the elections considered.

The error index (EI) statistic, defined in equation (19), quantifies the differences between $\mathbf{V}$ and $\hat{\mathbf{V}}$. This index, which was proposed by Romero et al. (2020) and is proportional to the AD statistic suggested in Klima et al. (2016), accounts for the percentage of votes erroneously allocated, i.e., the minimum number of votes that should be moved among cells to reach a perfect fit. Multiplication by 0.5 in (19) is done to avoid counting every wrongly assigned vote twice. The EI coefficient varies between 0, when $\mathbf{V}$ and $\hat{\mathbf{V}}$ coincide, and 100, when not a single vote has been correctly allocated. Although different methods score differently in this statistic, Klima et al. (2016) record, in a broad simulation study where five different algorithms are compared, average values of EI around 14% for the most accurate algorithm.

$$EI = 100 \cdot \frac{0.5 \sum_{jk} |v_{jk} - \hat{v}_{jk}|}{\sum_{jk} v_{jk}} \quad (19)$$

The EPW index, defined in equation (20), quantifies the mean of the differences between the actual p_{jk} values and the estimated $\hat{p}_{jk}$ values after weighting each difference by the number of votes associated with the transfer between options j of E1 and k of E2. Given that the mean value of these differences will always be equal to 0, since the sum of each row of both matrices $\mathbf{P}$ and $\hat{\mathbf{P}}$ is always equal to 1, each of the differences is calculated in absolute value. In the computation of this value each difference is weighted proportionally to the effective number v_{jk} of votes it affects to give more weight to the errors corresponding to the most relevant proportions.

$$EPW = 100 \cdot \frac{\sum_{jk} v_{jk} |p_{jk} - \hat{p}_{jk}|}{\sum_{jk} v_{jk}} \quad (20)$$

In the same way as the EI coefficient, the EPW coefficient varies between 0, when $\mathbf{P}$ and $\hat{\mathbf{P}}$ coincide, and 100, when not a single vote has been correctly assigned. In our research we have verified that, as expected, the EI and EPW discrepancy measures are closely correlated.

6.4 Findings

Table 3 summarises the results attained after applying lphom and the two new algorithms tslphom and nslphom introduced in this paper to the data described in subsection 6.2. In the case of nslphom we test five different specifications. The table displays by group of elections mean values of EI (upper panel) and EPW (middle panel), as well as average computation times (lower panel). The groups of elections considered are those corresponding to the 2002, 2005, 2008, 2011, 2014 and 2017 New Zealand general elections, the set of 420 New Zealand elections and the set of 73 elections corresponding to 2007 Scottish Parliament election.

Figures 3 and 4 show the same information displayed in the two upper-most panels of Table 3, but graphically. Comparing lphom and tslphom we observe that for all groups of elections tslphom generates, on average, more accurate values than lphom, both from the point of view of the measure EI (see Figure 3) and of the measure EPW (see Figure 4). This average superiority of tslphom is also observed at the individual level. For instance, tslphom produces more accurate results than lphom in terms of the EI measure in all but one of the 493 elections analysed; the exception being one in which the lphom solution is slightly more accurate than the tslphom solution (10.61 versus 10.63). Indeed, using the EI measure, the tslphom solutions are on average 11.5% more accurate than the lphom solutions. This advantage even grows to 12.0% when we consider the EPW measure.

In light of the above results, we can conclude without doubt that tslphom solutions are preferable to lphom estimates. Our global preferences, however, change as soon as we include in the comparisons the nslphom algorithm. We observe that nslphom consistently beats tslphom for all the specifications considered (see Table 3 and Figures 3 and 4. In all its versions, nslphom clearly outperforms lphom and tslphom, generating accurate results.

Table 3. Summary of the performance of the algorithms and its specifications.

	NZ 2002 N = 69	NZ 2005 N = 69	NZ 2008 N = 70	NZ 2011 N = 70	NZ 2014 N = 71	NZ 2017 N = 71	NZ 02-17 N = 420	SCO 2007 N = 73
Averages of EI measures								
lphom	16.88	12.29	12.22	12.99	12.95	12.20	13.24	12.92
tslphom	14.80	11.09	10.89	11.50	11.66	10.91	11.80	11.00
ns_first	13.03	9.80	9.28	9.75	10.04	9.20	10.17	8.87
ns_10	12.79	9.68	9.11	9.46	9.69	8.91	9.93	8.86
ns_25	12.77	9.55	8.92	9.37	9.75	8.85	9.86	9.13
ns_50	12.77	9.55	8.84	9.36	9.72	8.85	9.84	9.19
ns_100	12.78	9.55	8.82	9.36	9.72	8.85	9.84	9.21
Averages of EPW measures								
lphom	10.82	8.46	8.89	9.13	9.04	8.39	9.12	8.07
tslphom	9.42	7.59	7.90	8.05	8.15	7.46	8.09	6.72
ns_first	8.07	6.35	6.43	6.59	6.82	5.96	6.70	4.89
ns_10	7.90	6.09	6.09	6.26	6.55	5.67	6.42	4.80
ns_25	7.90	6.09	6.09	6.26	6.55	5.67	6.42	5.02
ns_50	7.90	6.09	6.01	6.24	6.51	5.67	6.40	5.06
ns_100	7.91	6.09	6.01	6.24	6.51	5.67	6.40	5.08
Averages of computation burden (in secs)								
lphom	5.74	5.73	2.57	5.18	4.19	4.49	4.64	0.16
tslphom	6.73	6.39	3.10	5.82	4.81	5.24	5.34	0.55
ns_first	10.16	9.22	6.95	9.40	8.25	8.97	8.82	2.72
ns_10	12.87	11.43	8.15	11.39	10.03	11.57	10.90	3.98
ns_25	23.20	19.88	16.59	20.72	18.82	22.21	20.23	9.31
ns_50	40.44	33.81	30.66	36.45	33.49	39.96	35.80	18.45
ns_100	75.00	61.66	58.68	67.72	62.79	75.38	66.87	36.71

Source: Compiled by the authors after applying the functions lphom, tslphom and nslphom available in the R package lphom (Pavía and Romero, 2021a) to the official data from the New Zealand electoral commission and the Scotland Electoral Office described in subsection 6.2. The estimations labelled as ns_first have been obtained using nslphom with the argument min.first = T (nslphom algorithm 1 in Table 1) and the estimations labelled as ns_10, ns_25, ns_50 and ns_100 with the arguments min.first = F and, respectively, max.iter = 10, 25, 50 and 100 (nslphom algorithm 2 in Table 1). The computations have been performed, in the case of New Zealand, on a desktop computer with a CPU processor Intel® Core™ i7-4930K (6 cores) 3.40GHz and 32GB of RAM and, in the case of Scotland, on a laptop with a CPU processor Intel® Core™ i7-6820HK (4 cores) 2.70GHz and 64GB of RAM.

Focusing now on which of the nslphom analysed specifications is preferable, we find that, although in general there are no great differences between the different versions of nslphom, it seems that ns_10 (the version in which the solution is obtained as the one with smaller HETe after 10 iterations) is the one showing the best balance between accuracy and computational burden. Solutions ns_first and ns_25 are nevertheless also competitive. It should be noted that the ns_50 and ns_100 estimates, in addition to be computationally expensive, do not significantly improve the less computationally demanding specifications and that, moreover, they may even be slightly worse in some cases. We observe this behaviour more clearly in the case of the Scottish elections.

Comparing the lphom and tslphom estimates with the solutions reached with ns_10, we observe that the ns_10 estimated matrices are, on average, 26.0% and 16.3% more accurate than the corresponding estimates of lphom and tslphom when measured using the EI index, and that these figures even increase to 31.0% and 22.6% when we use the EPW error measure. In summary, in terms of accuracy, tslphom is better than lphom and furthermore nslphom systematically improves tslphom. Likewise, focusing on the absolute levels of error and not on the rankings, we also observe that the new algorithms are quite competitive. Pooling all elections, ns_10 has an average value for EI of 9.77, a level of error that could be catalogued as

quite satisfactory compared to the results obtained by Klima et al. (2016) in their simulation study.

Figure 3. Graphical representation of average values of EI error measures grouped by election and algorithm. Individual solutions have been attained using the functions `lphom`, `tslphom` and `ns_lphom` of the R package `lphom` (Pavía and Romero, 2021a). The estimations labelled as `ns_first` have been obtained using `ns_lphom` with the argument `min.first = T` (this corresponds to `ns_lphom` algorithm 1 in Table 1) and the estimations labelled as `ns_10`, `ns_25`, `ns_50` and `ns_100` with the arguments `min.first = F` and, respectively, `max.iter = 10`, `25`, `50` and `100` (these correspond to different versions of the `ns_lphom` algorithm 2 in Table 1).

Regarding the average computation times, which are shown in seconds in the lower panel of Table 3, we find that, as expected, these increase linearly with the number of iterations. The recorded computational times, however, should be considered small for this kind of studies, especially compared to the computation times required by the methods recommended in the ecological regression literature. This is probably the most striking result to note in this regard. As a curiosity, it outstands that the average computation times in the New Zealand elections are much higher than in the Scottish ones. This is due to the fact that the former includes the Māori districts, whose electors are distributed in a number of territorial units significantly higher than regular districts (see Table 2), with many of them including a very small number of voters. In fact, we have verified that if we do not consider the Māori districts, the average computation times of New Zealand and Scottish elections are quite similar.

Figure 4. Graphical representation of average values of EPW error measures grouped by election and algorithm. Individual solutions have been attained using the functions `lphom`, `tslphom` and `ns_lphom` of the R package `lphom` (Pavía and Romero, 2021a). The estimations labelled as `ns_first` have been obtained using `ns_lphom` with the argument `min.first = T` (this corresponds to `ns_lphom` algorithm 1 in Table 1) and the estimations labelled as `ns_10`, `ns_25`, `ns_50` and `ns_100` with the arguments `min.first = F` and, respectively, `max.iter = 10`, `25`, `50` and `100` (these correspond to different versions of the `ns_lphom` algorithm 2 in Table 1).

Table 4. Number of real and estimated proportions equal to zero and one.

	Number of actual proportions			lphom	tslphom	ns_first	ns_10	ns_25	ns_50	ns_100
	Total	$p_{jk} = 0$	$p_{jk} < .01$	Number of zeros estimated						
District tables	14,158	62	692	5,706	968	736	701	701	696	696
Voting unit tables	1,219,175	$\geq 172,734$	-	-	529,220	283,022	274,855	271,955	271,508	271,441
	Total	$p_{jk} = 1$	$p_{jk} > .95$	Number of ones estimated						
District tables	14,158	0	49	333	50	49	49	49	49	49
Voting unit tables	1,219,175	-	-	-	1,974	384	371	358	356	356

Source: Compiled by the authors after applying the functions `lphom`, `tslphom` and `ns_lphom` available in the R package `lphom` (Pavía and Romero, 2021a), to the official data from the New Zealand electoral commission and the Scotland Electoral Office described in subsection 6.2.

Finally, to end the empirical assessment, we focus on extreme values. Table 4 presents the number of zeros and ones estimated at district and voting unit level by the different algorithms. As can be seen, the number of extreme proportions attained in the district tables by `lphom` is hugely above the actual number. Our results for `lphom` are in line with previous literature (Upton, 1978; Johnston and Hay, 1983; Romero et al., 2020; Romero and Pavía, 2021): the classical linear programming algorithm has an excessive tendency to produce extreme values. The new algorithms, on the contrary, significantly reduce the number of estimated extreme proportions. Although they do not eliminate this tendency completely, they only estimate zeros and ones when the corresponding fraction is equal or really close to that number. It also

highlights the enormous reduction in the frequency of extreme values that the nslphom specifications record compared to tslphom in voting unit tables. Indeed, given that we can compute a lower bound for their total number of zeros using the fact that when $x_{ij} = 0$ or $y_{ik} = 0$ the corresponding row or column proportion estimates must be zero, we can also conclude that the number of extreme values estimated by the nslphom algorithm is, relatively, not so frequent. For example, the number of estimates equal to zero attained by ns_10 are only 59% above the minimum, whereas tslphom more than triples this minimum.

7. Discussion and further research

In the previous sections, two new algorithms, tslphom and nslphom, were developed and their global performance assessed with real data. These new algorithms, in addition to be satisfactorily accurate, are able to provide within the mathematical optimisation framework estimates of local transition tables. To the best of our knowledge, no model under this framework has been proposed in the literature to do this. As we outline in the introduction, this represents an important step forward.

At first glance, it is surprising that research to date has not considered extending ecological inference solutions from the mathematical programming framework by first locally adjusting global estimates. Likely, this gap in the literature is due to the fact that the logical specification of this problem, which under a linear programming approach reasonably corresponds to step one of our lphom_local procedure, states an indeterminate linear program. This has possibly misguided other authors, preventing them from pursuing this path. Although, fortunately, the second step of the lphom_local algorithm resolves the indeterminacy leading tslphom and nslphom algorithms to unique solutions, in our opinion, pursuing local adjustments could have been beneficial from a practical perspective, even with indeterminacies. It seems that introducing into the problem all the information available through (new) local constraints is valuable. This increases the global accuracy of the solutions without apparently showing inconsistencies given that, once the linear programming solver and the local adjuster is fixed, the family of algorithms stated in Table 1 provide a unique sequence of estimates for each election. We have verified this after specifying some indeterminate algorithms as local adjusters and different solvers.

In particular, we have assessed two local indeterminate adjusters—one in which only the first step of lphom_local is run and another in which the norm L_1 considered in equation (17) is replaced by the norm L_∞ —and have confirmed the value of the approach even with indeterminacies. We have observed this by using as linear programming solvers the linprog function of MATLAB (Zhang, 1995) and the lp function of the lpSolve package of R (Bekelaar et al., 2020). Although the solvers programmed in linprog and lp consistently find different solutions under indeterminacy, both functions (linprog and lp) guide tslphom and nslphom in the examples considered to more accurate solutions than the ones obtained just applying lphom, providing, moreover, local estimates. In any case, the solutions attained using lphom_local as internal local adjuster are always preferable. In addition to generating unique solutions, they are, for the elections analysed, more accurate than the solutions attained with the other two tested local adjusters.

Although the fact that, under indeterminacy, the solution is not unique could be seen as a drawback, the truth is that many of the currently most recommended algorithms for solving the

ecological inference problem, being based on Bayesian approaches, also share this characteristic. It would be interesting to study the magnitude of the range of solutions under indeterminacy and whether it could be used as a measure of uncertainty. The fact that `nslphom` does not converge to a fixed point should not be perceived as a weakness by necessity: `nslphom` tends to quickly stabilise within a range of values (see Figure 1) and this could be interpreted as it having arrived at its stationary distribution. This behaviour, of fluctuating in a stationary distribution, is also common in the Bayesian solutions of this problem, where the solution of each step of the chain is not the same, but fluctuates (when it converges) around a stationary distribution.

The fluctuating behaviour of the `nslphom` step-solutions led to the reasoning in Section 5 in deciding which of all these solutions (which vary little) to choose. In subsection 5.2, we have performed some analyses in order to argue reasonable stopping rules for the `nslphom` algorithm, to finally link the solution to be chosen to the observed value of the HETe statistic. Given that the actual contingency tables of the studied elections are known, we have extended our analyses and investigated whether more accurate solutions could have been obtained under the current framework. In particular, after running a hundred iterations of `nslphom` and computing the values of the EI and EPW statistics for the whole sequences of estimates, we have found that there is room for improvement. For instance, if we had selected in each election the estimate with the smallest EI, we would have obtained an average value of 8.22 for EI in the 493 elections. This result is a 19.5% smaller than the corresponding average value of 9.83 obtained under the criterion of the smaller HETe with the specification `ns_100`. Obviously that criterion cannot be used in practice as the actual contingency tables are unknown.

In this vein, in an attempt to improve the estimates, and partially inspired by Figure 1, we have tested the idea of including in the `nslphom` algorithm a burn-in parameter (i.e., an integer specifying the number of initial iterations to be discarded before determining the final solution) and we have achieved mixed results. For example, after estimating all the transfer matrices using `nslphom` with 10 as burn-in parameter and 25 as the total number of iterations, we have attained a slight improvement in the global average accuracy but some worsening for specific groups of elections. In terms of the EI and EPW statistics, and compared to the solutions attained employing the specification `ns_25`, we have obtained global reductions from 9.75 and 6.18 to 9.35 and 5.80 for EI and EPW, respectively. At the same time, however, the figures for the elections of Scotland worsened, from 9.13 to 9.39 for EI and from 5.02 to 5.25 for EPW. Given that we are still unclear as to when setting a burn-in could be beneficial, more research is still necessary on this issue. A future line of research could focus on studying what observed indicators, if any, (such as the number of cells to be estimated or a measure of the heterogeneity of the margins of the local tables) could guide us in the process of defining more suitable, election-specific stopping rules. Likewise, other ideas to improve the `nslphom` stopping rule that also deserve to be investigated include analysing and exploiting the properties related to the times series defined by the sequences $\{\hat{p}_{jk}\}_{l=1}^{ns}$ and/or $\{HETe_l\}_{l=1}^{ns}$. For example, we can borrow from the Bayesian approach and take for each (j, k) -pair the mean of the sequence $\{\hat{p}_{jk}\}_{l>l^*}^{ns}$ as solution; with l^* chosen large enough so as to guarantee that the series of estimates arrives at its stationary distribution. This is a more complex solution with higher computational cost than our proposals. Nevertheless, we still think it deserves further consideration because, as a by-product, it promises a straightforward way of measuring the estimates' uncertainty. With our stopping rules, the uncertainty of the `nslphom` estimates could be computed by mimicking the procedure proposed for `lphom` in Romero et al. (2020).

8. Conclusions

The estimation of RxC ecological inference contingency tables from aggregate results defines one of the most salient and challenging problems in the field of quantitative social sciences. During the past quarter-century, the ecological regression approach has been prolific in proposing procedures to solve this problem. The advances within the mathematical programming approach, however, have been less striking. This paper closes the gap between both approaches by providing new tools within the mathematical programming framework. In particular, we suggest an algorithm (`lphom_local`) based on linear programming that, grounded on the homogeneity hypothesis, enables estimates to be attained of the joint cross-distributions of each unit in which the whole population is split out to then build two new ecological inference algorithms: `tslphom` and `nslphom`.

These two new algorithms represent an important step forward compared to the mathematical programming algorithms available to date. In addition to generating estimates of local ecological inference contingency tables, they significantly reduce the tendency, previously shown by other mathematical programming solutions, to produce extreme transfer probabilities. Likewise, and more importantly, they reveal themselves as satisfactorily accurate and more accurate than the baseline algorithm, `lphom`. Using real data from almost 500 elections, we show that `tslphom` systematically produces more accurate outcomes than `lphom` and that, moreover, `nslphom` consistently improves `tslphom`, this being possible simply by slightly increasing the computation burden. In short, the new algorithms, being at least as accurate as their best competitors from the statistical framework (Pavía and Romero, 2021b), improve the current baseline linear programming procedure in three distinct ways. First, they estimate local transition tables. Second, they generate (global) transition matrix with fewer extreme probabilities. Third, they offer a good fit to actual data, better than the baseline algorithm. In our view, these results show the linear programming approaches to be a competitive option, placing them once again in a prominent position in the ecological inference toolkit.

Among the difference specifications tested for `nslphom`, we find that a proper balance between accuracy and computational cost is reached after applying the second version of the `nslphom` algorithm introduced in Table 1 with ten iterations (`ns_10`). Nevertheless, we also verify that both the first version of the `nslphom` algorithm introduced in Table 1 (`ns_first`) and the second version of the algorithm with twenty-five iterations (`ns_25`) are also valid. The interested reader can use these algorithms employing the functions, with the same names, of the R package `lphom` (Pavía and Romero, 2021a).

References

- Berkelaar, M. et al. (2020). *lpSolve: Interface to 'Lp_solve' v.5.5 to Solve Linear/Integer Programs*. R package version 5.6.15. <https://CRAN.R-project.org/package=lpSolve>
- Corominas A., Lusa, A. & Valvet, M. D. (2015). Computing voter transitions: the elections for the Catalan Parliament, from 2010 to 2012. *Journal of Industrial Engineering and Management*, 8, 122–136.
- Duncan, O. & Davis, B. (1953). An alternative to ecological correlation. *American Sociological Review*, 18, 665–666.

- Forcina, A. & Pellegrino, D. (2019). Estimation of voter transitions and the ecological fallacy. *Quality & Quantity*, 53, 1859–1874.
- Goodman, L. A. (1953). Ecological regressions and the behaviour of individuals. *American Sociological Review*, 18, 663–664.
- Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, 64(6), 610–625.
- Gosnell, H. F. & Gill, N. N. (1935). An Analysis of the 1932 Presidential Vote in Chicago. *The American Political Science Review*, 29, 967–984.
- Gosnell H. F. & Schmidt, M. J. (1936). Factorial and Correlational Analysis of the 1934 Vote in Chicago. *Journal of the American Statistical Association*, 31, 507–518.
- Greiner, D. J. & Quinn, K. M. (2009). RxC ecological inference: Bounds, correlations, flexibility, and transparency of assumptions. *Journal of the Royal Statistical Society, A*, 172, 67–81.
- Greiner, D. J. & Quinn, K. M. (2010). Exit polling and racial bloc voting: combining individual-level and RxC ecological data. *The Annals of Applied Statistics*, 4, 1774–1796.
- Haneuse, S. & Wakefield, J. (2004). Ecological inference incorporating spatial dependence. In King, Rosen and Tanner (eds.). *Ecological Inference. New Methodological Strategies*. New York: Cambridge University Press, pp. 266–301.
- Hawkes, A. (1969). An approach to the analysis of electoral swing. *Journal of the Royal Statistical Society, A*, 132, 68–79.
- Imai, K. & Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24, 263–272.
- Imai, K, Lu, Y. & Strauss, A. (2008). Bayesian and likelihood inference for 2x2 ecological tables: An incomplete data approach. *Political Analysis*, 16, 41–69.
- Irwin, G. & Meeter, D. (1969). Building voter transition models from aggregate data. *Midwest Journal of Political Science*, 13, 545–566.
- Johnston, R. J. & Hay, A. M. (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., Rosen, O. & Tanner, M. A. (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research*, 28, 61–90.
- King, G., Rosen, O. & Tanner, M. A. (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., Thurner, P. W., Molnar, C., Schlesinger, T. & Küchenhoff, H. (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., Schlesinger, T., Thurner, P. W. & Küchenhoff, H. (2019). Combining aggregate data and exit polls for the estimation of voter transitions. *Sociological Methods & Research*, 48, 296–325.
- McCarthy, C. & Terence, M. R. (1977). Estimates of voter transition probabilities from the British General Elections of 1974. *Journal of the Royal Statistical Society, A*, 140, 78–85.
- Ogburn, W. F. & Goltra, I. (1919) How Women Vote. *Political Science Quarterly*, 34, 413–433.
- Ogburn, W. F. & Talbot, N. S. (1929). A Measurement of the Factors in the Presidential Election of 1928. *Social Forces*, 8, 175–183.
- Park, W., Hanmer, M. J. & Biggers, D. R. (2014). Ecological inference under unfavorable conditions: straight and split-ticket voting in diverse settings and small samples. *Electoral Studies*, 36, 192–203.
- Pavía, J. M. & Aybar, C. (2020). La movilidad electoral en las elecciones de 2019 en la Comunitat Valenciana. *Debats. Revista de cultura, poder y sociedad*, 134, 27–51.
- Pavía, J. M. & Romero, R. (2021a) *lphom: Ecological Inference by Linear Programming under Homogeneity*. R package version 0.1.3. <https://CRAN.R-project.org/package=lphom>
- Pavía, J. M. & Romero, R. (2021b). Data wrangling, computational burden, automation, robustness and accuracy in ecological inference forecasting of RxC tables.
- Pavía-Miralles, J. M. (2005). Forecasts from non-random samples: The election night case. *Journal of the American Statistical Association*, 100, 1113–1122.
- Plescia, C. & De Sio, L. (2018). An evaluation of the performance and suitability of RxC methods for ecological inference with known true values. *Quality & Quantity*, 52, 669–683.
- Puig, X. & Ginebra J. (2015). Ecological inference and spatial variation of individual behavior: national divide and elections in Catalonia. *Geographical Analysis*, 47, 262–283.
- Romero, R. & Pavía, J.M. (2021). Estimating vote party entries and exits by ecological inference. Mathematical programming versus Bayesian statistics. *BEIO*, forthcoming.
- Romero, R., Pavía, J.M., Martín, 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, 2711–2736.
- Rosen, O., Jiang, W., King, G. & Tanner M. A. (2001). Bayesian and frequentist inference for ecological inference: The RxC Case. *Statistica Neerlandica*, 55, 134–56.
- Tziafetas, G. (1986). Estimation of the voter transition matrix. *Optimization*, 17, 275–279.
- Upton, C. J. G. (1978). A Note on the Estimation of Voter Transition Probabilities, *Journal of the Royal Statistical Society, A*, 141, 507–512.
- Vangrevelinghe, G. (1961). Étude statistique comparée des résultats des référendums de 1958 et 1961. *Revue de Statistique Appliquée*, 9, 83–100.
- 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.
- Wakefield, J. (2004). Ecological inference for 2x2 tables (with discussion). *Journal of the Royal Statistical Society, A*, 167, 385–445.

Zhang, Y. (1995). *Solving Large-Scale Linear Programs by Interior-Point Methods Under the MATLAB Environment*. Technical Report TR96-01, Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD.