Original Article

Sociological Methods & Research
1–43
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00491241221092725
journals.sagepub.com/home/smr
SAGE

Improving Estimates Accuracy of Voter Transitions. Two New Algorithms for Ecological Inference Based on Linear Programming

Jose M. Pavía[1]
and Rafael Romero[2]

Abstract

The estimation of RxC ecological inference contingency tables from aggregate data is 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

[1] GIPEyOP, UMICCS, Universitat de Valencia, Spain
[2] Universidad Politécnica de Valencia, Spain

Corresponding Author:
Jose M. Pavía, GIPEyOP, UMICCS, Universitat de Valencia.
Email: pavia@uv.es

probability estimates previously observed in other mathematical programming procedures, these two new algorithms prove to be quite competitive and more accurate than the current linear programming baseline algorithm. Their accuracy is assessed using a unique dataset with almost 500 elections, where the real transfer matrices are known, and their sensitivity to assumptions and limitations are gauged through an extensive simulation study. The new algorithms place the linear programming approach once again in a prominent position in the ecological inference toolkit. Interested readers can use these new algorithms easily 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 is a challenge that dates back to the 1960s (Vangrevelinghe, 1961; Hawkes, 1969; Irwin and Meeter, 1969). This problem is actually 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 behaviour from aggregated data is called ecological inference, which is exposed to what is known as the ecological fallacy (Robinson, 1950).

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. That is, 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 (Cho and Manski, 2008). Using observed data alone, one can identify, at best, a range for the feasible set of counts.

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. Instead of one table, we now 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. Typically, row fractions or transition probabilities of (subgroups of) contingency tables of the different territorial units are considered to be, to a certain extent, similar/related (Imai, et al. 2008; Greiner and Quinn, 2009; Forcina and Pellegrino, 2019). This is traditionally referred to as a credible assumption responding to the common observation that people belonging to the same group tend to follow similar behaviour patterns. Assuming this, however, does not presume anything about which of the usual mechanisms often argued to explain this phenomenon—endogenous effects, exogenous (contextual) effects and/or correlated effects (e.g., Manski, 2007)—is at play.

Based on this hypothesis, many algorithms for estimating row fractions or row-conditional (underlying) probabilities, grounded in different philosophical foundations and/or employing different mathematical approaches, can be found in the literature. 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 one most prolifically used has been the statistic framework, mainly after King (1997) who masterfully combined Goodman's regression and Duncan and Davis' method of bounds. King increased the credibility of the promised inferences after (mathematically) translating the homogeneity assumption in a significantly more flexible way than Goodman. Indeed, since the publication of King's book "A solution to the ecological inference problem", there has been a resurgence of proposals within the so-called ecological regression approach, many of the earlier ones being designed for dealing with 2 × 2 tables and later generalised for 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). Readers interested 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 research studies has been that of mathematical programming. In this setting, deterministic bounds are incorporated in a natural way via exact and inequality constraints. The proposals within this framework, which focus almost exclusively 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 to be 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 and many issues linked to the mathematical programming solutions 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”. In this paper, following the same investigative direction as Romero et al. (2020), we contribute solutions to two other more important but 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 criticised (e.g., Upton, 1978; Johnston and Hay, 1983; Romero and Pavía, 2021) as tending to produce many extreme probabilities or fractions: zeros and ones. In this research we propose solutions to both these questions.[1]

First, we suggest a novel procedure, based on linear programming and grounded in the homogeneity hypothesis, to estimate the inner-cells values at the local level. We call this procedure lphom_local. On this, we then build 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, as we show later with real data, systematically outperform lphom, with tslphom and nslphom producing 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 outperforms lphom and that, likewise, nslphom consistently outperforms tslphom. Furthermore, in an independent study (Pavía and Romero, 2022), 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 particularly 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, by knowing their past behaviour, which arguments to use to persuade them.

Despite the proven performance of the new proposals with real data, these assessments still leave some relevant questions unanswered, such as what happens when we have significant departures from the assumption on which the algorithms rest and what is the accuracy of estimates at the unit level. Because in our real datasets both the actual mechanisms generating the data and the cross-distributions of local units are unknown, we rely on simulated data to answer these questions. These extra analyses allow for a better understanding of the limitations of the proposed methods.

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 assessing lphom, tslphom and nslphom solutions with real data. Section 7 presents the outcomes of the simulations. Section 8 discusses the findings and suggests directions for further research. Finally, Section 9 summarises and concludes.

2. The Baseline Model: lphom

Without loss of generality, from here on in the paper, we follow the terminology used in Romero et al. (2020) and consider the problem of estimating the matrix of transfer of votes between two election processes. 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 $i = 1, \ldots, I$ voting units, the votes x_{ij} recorded for the $j = 1, \ldots, J$ election options available in E1 and the votes y_{ik} ($k = 1, \ldots, K$) harvested by the different competing options in E2. The basic variates of the model are the J × 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, $p_{jk} = \sum_{i=1}^{I} p^{i}_{jk}\ \omega^{i}_{j}$ is a (weighted) average of the transfer proportions in the voting units, where p^{i}_{jk} is the proportion of voters in unit i who, having chosen option j in E1, have chosen option k in E2 and $\omega^{i}_{j} = x_{ij} / \sum_{i'=1}^{I} x_{i'j}$. According to this definition, the p_{jk} must meet the following constraints:

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

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

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

The above system has more unknowns than data, being only partially identified. Hence, to reduce the indeterminacy, narrowing (under mild conditions) the region of feasible solutions to a point, lphom introduces the hypothesis of homogeneity/similarity of electoral behaviour in the I units. Specifically, the homogeneity hypothesis establishes that the unit vote transfer fractions/probabilities, p^{i}_{jk}, are *similar* to the average fractions, 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 the average fractions to x_{ij}. Naming e_{ik} as these discrepancies (see equation (4)), we have that the e_{ik} should be *small.*

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

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

$$\text{minimize} \sum_{i,k} |e_{ik}| \tag{5}$$

For equations (1), (2) and (3) to be compatible, it is necessary that the sums of the rows of the matrices *IxJ* and *IxK* 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, available on CRAN, 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).

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

$$p_{JK} = 0 \tag{7}$$

3. Estimating Voter Transitions at the Local Level

The lphom algorithm estimates the matrix $[p_{jk}]$ of voting transfer fractions/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 $[p^i_{jk}]$ 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 $[x_{ij}]^I_{i=1}$ and $[y_{ik}]^I_{i=1}$, introduced in the previous section, and a global matrix $[p^G_{jk}]$ of transfer probabilities for the whole territory. This matrix could be obtained using, for instance, lphom. The unknowns of this new model are the $[p^i_{jk}]$ $(i = 1, \ldots, I)$ 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^i_{jk} must fulfil the following constraints:

$$p^i_{jk} \geq 0 \quad \textit{for} \;\; j = 1, \ldots, J \quad k = 1, \ldots, K \quad i = 1, \ldots., I \tag{8}$$

$$\sum_{k=1}^{K} p^i_{jk} = 1 \quad \textit{for} \;\; j = 1, \ldots, J \quad i = 1, \ldots., I \tag{9}$$

$$\sum_{j=1}^{J} x_{ij} p^i_{jk} = y_{ik} \quad \textit{for} \;\; k = 1, \ldots, K \quad i = 1, \ldots., I \tag{10}$$

As in the lphom model, this system of equations sets up an indeterminate system (to be 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^i_{jk} are *similar* to the corresponding global p^G_{jk}. Thus, under this hypothesis, the estimate of the volume of voters v^i_{jk} 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^i_{jk} or p^G_{jk}. Thus, the quantities ε^i_{jk} defined by equation (11) should be *small.*

$$\varepsilon^i_{jk} = x_{ij} p^i_{jk} - x_{ij} p^G_{jk} \quad \textit{for} \;\; j = 1, \ldots, J \quad k = 1, \ldots, K \quad i = 1, \ldots., I \tag{11}$$

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

$$minimize\ \ Z = \sum_{j,k} |\varepsilon^i_{jk}| \qquad for \quad i = 1, \ldots, I \tag{12}$$

As with lphom, lphom_local must satisfy, regarding entries and exits, the restrictions imposed in each unit i by the current scenario. Specifically, 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 can include the additional constraints given by equations (13) and (14).

$$p^i_{jK} = y_{iK} \,/ \left(\sum_{j=1}^{J-1} x_{ij} \right) \qquad for \quad j = 1, \ldots, J-1 \qquad i = 1, \ldots, I \tag{13}$$

$$p^i_{JK} = 0 \qquad for \quad i = 1, \ldots, I \tag{14}$$

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^G_{jk} verifies (3), the above system remains still partially identified, although with a set of feasible solutions smaller than the one derived from the observed data (equations (8)–(10)). It is indeterminate in the sense that an infinite set of substantively different $[p^i_{JK}]$ 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 S3 of the Supplementary Material.

In order to overcome the indeterminacy and to narrow down further the set of feasible solutions, 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 $[p^i_{jk}]$ closest to the global matrix $[p^G_{jk}]$. Specifically, we propose adding to the above linear program two new equations, (15) and (16), for each i and to minimize, as 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).

$$Z = \sum_{j,k} |\varepsilon^i_{jk}| \qquad for \quad i = 1, \ldots, I \tag{15}$$

$$p^i_{jk} = p^G_{jk} + \delta^i_{jk} \quad \mathit{for} \quad j = 1, \ldots, J \quad k = 1, \ldots, K \quad i = 1, \ldots., I \tag{16}$$

$$\mathit{minimize} \sum_{j,k} |\delta^i_{jk}| \quad \mathit{for} \quad i = 1, \ldots, I \tag{17}$$

Our proposal, lphom_local, to estimate voting transfer matrices in each unit is therefore a two-step procedure where, in the first step, the set of potential solutions is delimited to subsequently, in the 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 $\{p^i_{jk}\}^K_{k=1}$ 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

Various authors (e.g., Upton, 1978; Johnston and Hay, 1983; Corominas et al., 2015; Romero and Pavía, 2021) have pointed out that mathematical programming procedures have an excessive tendency to include p_{jk} estimates equal to 1 in its solutions, 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 clearly demonstrated (see subsection 6.4) 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 "**T**wo **S**teps **lphom**" and refers to the fact that in the process of estimating the final global matrix, $\boldsymbol{P} = [p_{jk}]$, of vote transition probabilities, the matrix $\boldsymbol{P}$ is obtained twice. The tslphom algorithm works as follows:

1. As a first step, given the data $[x_{ij}]_{i=1}^{I}$ and $[y_{ik}]_{i=1}^{I}$, a solution matrix $\hat{\boldsymbol{P}}_o$ is obtained by applying the lphom procedure as stated in section 2.
2. Next, using $\hat{\boldsymbol{P}}_o$ as the reference matrix of global transition probabilities, the lphom_local procedure proposed in section 3 is applied to obtain estimates of the matrices $\boldsymbol{V}_i = [v^i_{jk}]$ of vote transition in the I territorial units.
3. Finally, the $\hat{\boldsymbol{V}}_i = [\hat{v}^i_{jk}]$ matrices estimated in the previous step are aggregated to obtain a global vote transition matrix. The tslphom global estimated matrix of transition fractions/probabilities, $\hat{\boldsymbol{P}}_1 = [{}_1\hat{p}_{jk}]$, 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^i_{jk} 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 in 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).

$$HET = 100 \cdot \frac{0.5 \sum_{ijk} |v^i_{jk} - x_{ij}p_{jk}|}{\sum_{ij} x_{ij}} \tag{18}$$

In equation (18), the v^i_{jk} 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^i_{jk} 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^i_{jk}, however, are obtained when we work with the tslphom algorithm. In this case, it is possible to obtain an estimate of the index of heterogeneity, which we also call HETe, applying the plug-in principle. HETe in this case is obtained by replacing in (18) v^i_{jk} by $\hat{v}^i_{jk}$ and p_{jk} by ${}_1\hat{p}_{jk}$. 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 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, $\boldsymbol{P}^G$, the last attained transition probability matrix, and to stop the process when the matrix $\boldsymbol{P}^G$ 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 real instances 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 nslphom would tend to stabilize in a point 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 tend to oscillate around some reasonable attraction point. This should not be a surprise given that in essence this problem is only partially identified. We discover that the process improves the estimates during the first steps, up to a certain point, after which it has less effect, even slightly worsening the step-point solutions in some cases.

In this section, we define a new algorithm, which we call nslphom (as acronym of "**N** **S**teps **lphom**"), where in order to attain a solution we iterate steps two and three of tslphom for a limited number of times. Hence, the critical issue to define nslphom 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 nslphom 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 nslphom 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 shows the evolution of the estimates obtained by nslphom 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 in a point and, more importantly, that all the estimated values look reasonable and they show relatively small variations after the first iterations. They fluctuate between 0.990 and 0.994. Hence, given that when we build the model nslphom, 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 into an operable rule. From the computation point of view, this will not pose any particular difficulty as, in each iteration, together with the new solution, we can also calculate the HETe statistic. From the judgement point of view, we can exploit the

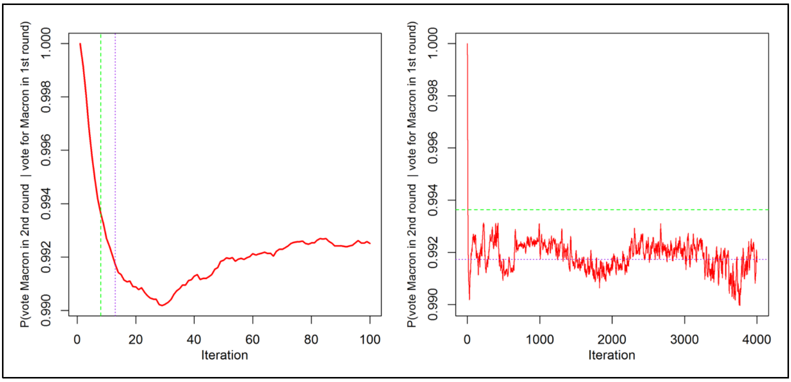

Figure 1. Evolution of nslphom solution for $p_{M,M}$ as a function of the number of iterations. In the left panel, the dashed-green and dotted-purple lines identify, respectively, the iterations in which HETe reaches its first minimum and its global minimum. In the right panel, the lines identify, respectively,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

pattern observed in Figure 2, which we have observed (with some variations) in all the elections we have analysed.

Figure 2 shows the evolution of HETe when nslphom 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; thereafter, it consistently begins 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 the 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 to finally stabilize again.

In light of these results, we envisage two reasonable strategies for the nslphom algorithm to produce a point solution. On the one hand, a reasonable stopping criterion for nslphom is to end the process at the first iteration

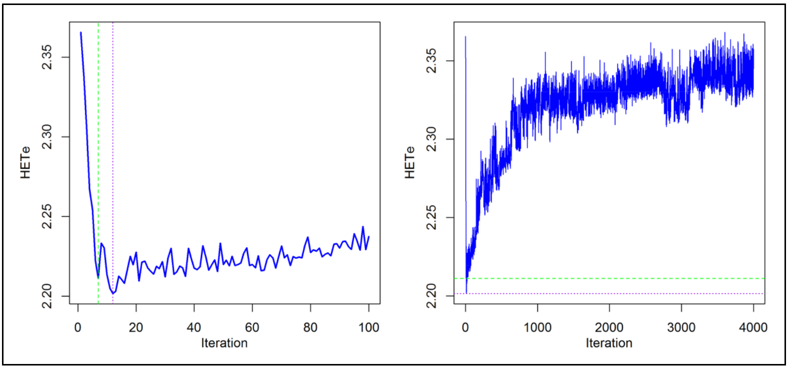

Figure 2. Evolution of HETe in nslphom as a function of the number of iterations. In the left panel, the dashed-green and dotted-purple lines identify, respectively, the iterations in which HETe reaches its first minimum and its global minimum. In the right panel, the lines identify, respectively, the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

in which HETe starts to grow and to take as solution the vote transfer matrix attained in the previous iteration. From here on, we name nslphom with this criterion ns_first. This is equivalent to the use of the nslphom R function of the package `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 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 iter.max = *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 point solutions slightly 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.

6. Assessing the Accuracy of lphom, tslphom and nslphom with Real Data

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. On the one hand, polls are exposed to significant sources of bias and generate estimates with large variances, with large doses of subjectivity pervading reasonableness of socio-political outcomes, mainly where there are no

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.

```
0. Let $\boldsymbol{X} = [x_{ij}]_{i=1}^{I}$ and $\boldsymbol{Y} = [y_{ik}]_{i=1}^{I}$ be the row-vector matrices of votes recorded in,
   respectively, E1 and E2 in the $I$ voting units.
1. Estimate $\hat{\boldsymbol{P}}_o$ by applying the lphom algorithm to $\boldsymbol{X}$ and $\boldsymbol{Y}$. Assign $\boldsymbol{P}^G \leftarrow \hat{\boldsymbol{P}}_o$, $t \leftarrow 1$,
   $HETe_0 \leftarrow \infty$.
2. Estimate $\hat{\boldsymbol{P}}_t$ and $HETe_t$ by applying the lphom_local procedure to $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{P}^G$.
   Assign $\boldsymbol{P}^G \leftarrow \hat{\boldsymbol{P}}_t$, $t \leftarrow t+1$.
3. If $HETe_t > HETe_{t-1}$ stop; otherwise go back to 2.
4. Select as solution $\hat{\boldsymbol{P}}_{t-1}$.
```

nslphom algorithm 2. Pseudo code with the number of iterations set in advance.

```
0. Let $ns$ be the maximum number of iterations to be performed and let $\boldsymbol{X} = [x_{ij}]_{i=1}^{I}$
   and $\boldsymbol{Y} = [y_{ik}]_{i=1}^{I}$ be the row-vector matrices of votes recorded in, respectively,
   E1 and E2 in the $I$ voting units.
1. Estimate $\hat{\boldsymbol{P}}_o$ by applying the lphom algorithm to $\boldsymbol{X}$ and Assign $\boldsymbol{P}^G \leftarrow \hat{\boldsymbol{P}}_o$, $t \leftarrow 1.\boldsymbol{Y}$.
2. Estimate $\hat{\boldsymbol{P}}_t$ and $HETe_t$ by applying the lphom_local procedure to $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{P}^G$.
   Assign $\boldsymbol{P}^G \leftarrow \hat{\boldsymbol{P}}_t$, $t \leftarrow t+1$.
3. If $t >$ ns stop; otherwise go back to 2.
4. Select as solution the $\hat{\boldsymbol{P}}_{t^*}$ whose $HETe_{t^*}$ is minimum for $t \leq t^* \leq ns$.
```

substantial differences between the solutions reached using different algorithms. On the other hand, the conditions defining the scenarios of the simulation exercises are sometimes set, even unconsciously, to favour one of the algorithms.

In certain circumstances, however, such as in mixed-member election systems in which voters cast two votes simultaneously in the same ballot and these 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 a one-off, 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. 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 and then, 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 2 offers some details about different characteristics of the datasets used to assess the performance of the algorithms, with significantly more details available in Pavía (2022). As can be seen, 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.

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, in 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 files with the raw data from 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 $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{P}$ matrices. 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 did 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, $\boldsymbol{V} = [v_{jk}]$ and $\hat{\boldsymbol{V}} = [\hat{v}_{jk}]$, and the real and estimated matrices of transition probabilities, $\boldsymbol{P} = [p_{jk}]$ and $\hat{\boldsymbol{P}} = [\hat{p}_{jk}]$. We use these 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 with these datasets as real values are not available at this level for the elections considered. We study local errors in the next section, with simulated data.

The error index (EI) statistic, defined in equation (19), quantifies the differences between $\boldsymbol{V}$ and $\hat{\boldsymbol{V}}$. This index, which was proposed by Romero et al. (2020) and is proportional to the AD statistic suggested by 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 $\boldsymbol{V}$ and $\hat{\boldsymbol{V}}$ 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.

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

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

Table 2. Summary of Some Features of the Datasets 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 NZ electoral commission and the Scotland Electoral Office.

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 $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ 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.

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

In the same way as the EI coefficient, the EPW coefficient varies between 0, when $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ 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 the 2007 Scottish Parliament election.

Figures 3 and 4 show the same information displayed in the two uppermost panels of Table 3, but graphically. Interested readers can also consult Tables S1 and S2 of the Supplementary Material as they show the pairwise differences between the average error measures, grouped by algorithm and blocks of elections. Observing pairwise differences could help some readers to more easily appreciate the differences in accuracy between the different algorithms.

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

Table 3. Summary of the Performance of the Algorithms and its Specifications with Real Data.

							NZ	
	NZ	NZ	NZ	NZ	NZ	NZ	02-17	SCO
	2002	2005	2008	2011	2014	2017	N =	2007
	N = 69	N = 69	N = 70	N = 70	N = 71	N = 71	420	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 of the R package `lphom`, attached as Supplementary Material to this paper, 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, iter.max = 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™ i7-4930 K (6 cores) 3.40 GHz and 32GB of RAM and, in the case of Scotland, on a laptop with a CPU processor Intel® Core™ i7-6820HK (4 cores) 2.70 GHz and 64GB of RAM.

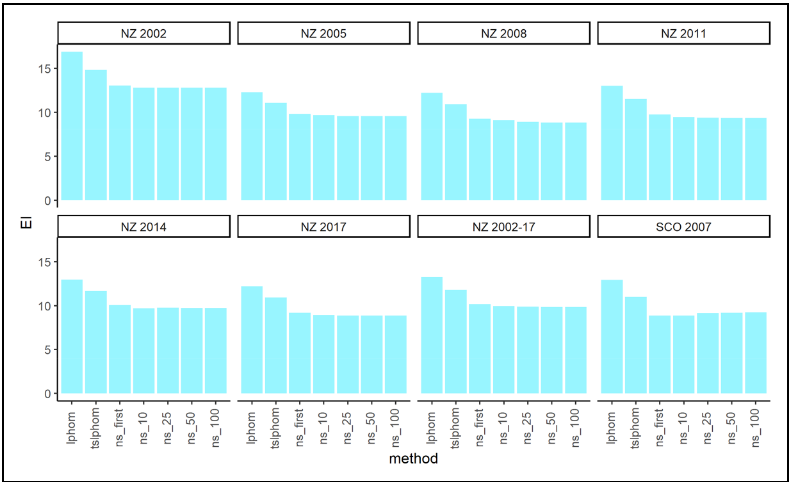

Figure 3. Graphical representation of average values of EI error measures grouped by election and algorithm. Individual solutions have been attained using the functions lphom, tslphom and nslphom of the R package `lphom`, available as Supplementary Material to this paper. The estimations labelled as ns_first have been obtained using nslphom with the argument min.first = T (this corresponds to 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, iter.max = 10, 25, 50 and 100 (these correspond to different versions of the nslphom algorithm 2 in Table 1).

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 outperforms tslphom for all the specifications considered (see Table 3 and Figures 3 and 4; and Tables S1 and S2 in the Supplementary Material). In all its versions, nslphom clearly outperforms lphom and tslphom, generating accurate results.

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

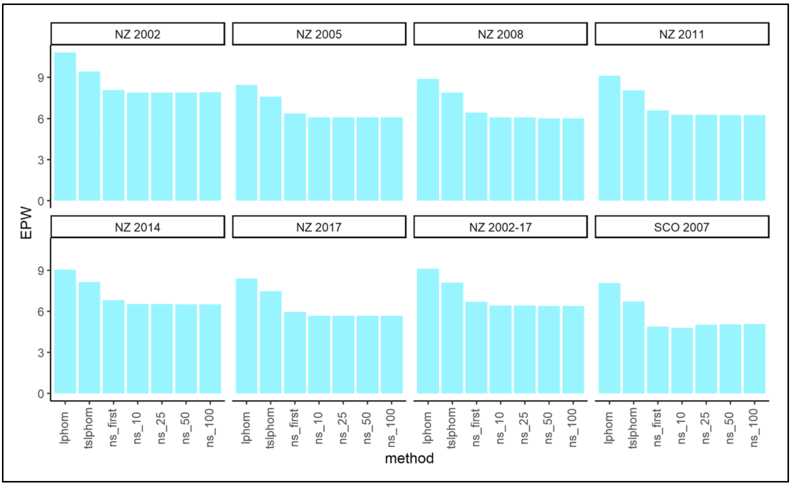

Figure 4. Graphical representation of average values of EPW error measures grouped by election and algorithm. Individual solutions have been attained using the functions lphom, tslphom and nslphom of the R package `lphom`, available as Supplementary Material to this paper. The estimations labelled as ns_first have been obtained using nslphom with the argument min.first = T (this corresponds to 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, iter.max = 10, 25, 50 and 100 (these correspond to different versions of the nslphom algorithm 2 in Table 1).

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 being computationally expensive, do not significantly improve the less computationally demanding specifications and, 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.

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 study, 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. 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 significantly higher number of territorial units than regular districts (see Table 2), with many of them holding 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.

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): 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. The results also highlight the enormous reduction in the frequency of extreme values that the nlsphom 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.

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	≥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 nslphom available in the R package `lphom` to the official data from the New Zealand electoral commission and the Scotland Electoral Office described in subsection 6.2.

7. Assessing Departure from Assumptions by Simulation

Although the lphom model could be placed, in the same way as other ecological inference methods, within the conceptual framework on partial identification stated by Manski (see, e.g., Manski 2003, 2007)[2], at first glance our proposals just look like heuristic algorithms: ad-hoc methods that solve the ecological inference problem by mathematically combining observed data and a credible hypothesis. The observed data—equations (1) to (3) and (8) to (10)—delimit the regions of feasible solutions and our mathematical translations of the homogeneity assumption—equations (4), (5), (11), (12) and (15) to (17)—narrow these regions to point estimates. When this happens, when a model yields point estimates, Cho and Manski (2008, p. 554) consider that the model "should be approached guardedly with attention to the impact of the assumptions." In this section we assess, through simulation, what happens when the data are generated under several levels of departure from the homogeneity hypothesis. Here, we also briefly gauge the accuracies of tslphom and nslphom estimating unit transition matrices.

As there are many issues with potential impact on the quality of an ecological inference estimate (Pavía and Romero, 2022), with many of them (e.g., the number of units, voters, rows and columns or how electors and votes are distributed across units) not being directly related to the question of interest, we have taken some real elections as reference. This is a common practice in ecological inference simulation analyses (see, e.g., Klima et al., 2016, 2019; Barreto et al., 2022) that allows the focus to be on the objective by keeping the non-relevant issues fixed. By taking as reference four elections—Bay of Plenty 2005 (50 units, 6 × 5 matrix), Rangitikei 2011 (101 units, 6 × 5 matrix), Te Tai Tokerau 2014 (299 units, 7 × 4 matrix) and Te Tai Tonga 2017 (705 units, 7 × 5 matrix)—we study how estimate accuracy depends on how voter transitions deviate across units from the (expected) global transition matrix. We build simulations by considering (i) four basic deviation schemes (constant, homogeneity, heterogeneity and several populations), (ii) the presence/absence of aggregation bias and (iii) how transfer rates are interpreted (either row fractions or row-conditional (underlying) probabilities). Their combinations determine a total of 12 different generating processes for unit-level voter/proportion transitions.

In constant schemes (hereafter identified with the acronym C), the fractions/probabilities (rates) are assumed to be constant across units. This corresponds to the maximum level of similarity and serves as a baseline. Homogeneity (H) transitions are designed as realistic rates. Here, the unit tables, which are simulated from common Dirichlet distributions, show levels of deviations (measured with the HET index) similar to the ones observed in real instances. Heterogeneity (T) transitions mimic scenarios where unit tables can deviate significantly between them. This is simulated by insufflating more variance in the Dirichlet distributions which generate the unit tables. In several populations schemes (S), three quite different district tables of rates are simulated for each simulation and one of them is randomly assigned to each unit as (initial) transition matrix. At this point, simulated unit matrices are either directly used without aggregation bias (W) or modified, as a function of the corresponding unit row totals, to induce aggregation bias (A). Finally, the transfer matrices of votes in each unit are simulated either (i) considering transition rates as underlying probabilities (P) and generating counts from multinomial processes or (ii) directly applying the rates (N) to the row totals and rounding them. Significantly more heterogeneity is induced in the P cases. In summary, we assume twelve different generating processes for voting transitions, which we identify with the acronyms: CWN, CWP, CAN, CAP, HWN, HWP, HAN, HAP, TWN, TWP, TAN, TAP, SWN, SWP, SAN, SAP. For instance, TWN refers to heterogeneous voter transitions without aggregation bias and no probabilistic approach. Technical details of the simulation design are given in the Supplementary Material (see Section S2).

For each of the four reference elections and each generating process, we have simulated thirty complete individual datasets[3] and used the margins from the unit voter transitions tables as input data to estimate unit and district tables. The estimates have been attained using methods lphom, tslphom, ns_first, ns_10 and ns_25. To reduce the computational burden, the estimates using ns_50 and ns_100 were not pursued here in light of the results attained in Section 6. Table 5 and Table S7 (in the Supplementary Material) show, respectively, averages of EI and EPW error measures, grouped by generating process and reference election, attained after comparing simulated and estimated global voter transition matrices. Figures S1 and S2 (in the Supplementary Material) display the corresponding box-plots of individual errors. Figure 5 and Figure S3 present the boxplots merging reference elections.

In terms of impact of the generating processes, the results are as expected. As a rule, the estimates worsen when we depart further from the homogeneity hypothesis: they tend to deteriorate with higher heterogeneity and aggregation bias. Nevertheless, the estimates can be considered reasonable even for

Table 5. Averages of EI Statistics Grouped by Generating Process in the Simulated Scenarios.

	Scenarios based on Bay of Plenty 2005					Scenarios based on Rangitikei 2011				
	lphom	tslphom	ns_first	ns_10	ns_25	lphom	tslphom	ns_first	ns_10	ns_25
Real	13.40	12.67	12.56	12.56	12.56	13.51	11.36	7.84	7.84	7.84
CWN	0.76	0.74	0.74	0.74	0.74	1.73	1.66	1.54	1.54	1.54
CWP	6.41	6.25	6.38	6.52	6.70	6.36	6.09	5.71	5.74	5.86
CAN	9.27	9.17	9.14	9.14	9.15	6.43	6.59	6.59	7.20	7.20
CAP	6.10	6.04	6.32	6.57	6.69	6.23	6.03	5.87	6.05	6.13
HWN	9.10	8.79	9.15	9.45	9.52	5.57	5.34	5.03	4.95	5.01
HWP	10.25	10.04	10.53	10.95	11.10	8.55	7.92	7.41	7.67	8.02
HAN	12.99	12.69	12.61	12.91	13.25	8.40	8.20	8.15	8.32	8.40
HAP	14.63	14.30	14.40	14.67	14.90	9.86	9.46	9.60	10.22	10.47
TWN	13.72	12.68	13.83	14.97	15.36	9.08	8.07	8.87	9.47	9.90
TWP	15.15	14.62	16.19	17.51	18.14	10.72	9.31	10.12	11.18	11.49
TAN	17.21	16.34	16.83	17.56	17.89	10.43	9.69	10.59	11.43	12.11
TAP	17.91	17.09	18.29	19.34	20.05	11.44	10.30	11.28	12.28	13.03
SWN	14.31	13.49	15.19	17.00	17.82	10.23	8.96	9.78	10.25	10.77
SWP	14.78	14.14	15.66	16.24	16.89	11.91	10.83	11.61	11.93	12.41
SAN	7.92	7.80	7.81	7.89	7.94	3.28	3.43	3.84	3.95	3.96
SAP	9.87	9.67	9.78	9.92	9.95	6.66	6.83	7.87	8.24	8.38
	Scenarios based on Te Tai Tokerau 2014					Scenarios based on Te Tai Tonga 2017				
	lphom	tslphom	ns_first	ns_10	ns_25	lphom	tslphom	ns_first	ns_10	ns_25
Real	14.03	11.70	8.75	8.75	8.75	12.68	11.09	10.47	10.25	10.25
CWN	3.78	3.43	3.43	3.43	3.43	4.75	4.32	4.13	4.13	4.13

(continued)

Table 5. Continued

	Scenarios based on Bay of Plenty 2005					Scenarios based on Rangitikei 2011				
	lphom	tslphom	ns_first	ns_10	ns_25	lphom	tslphom	ns_first	ns_10	ns_25
CWP	7.63	7.12	7.39	7.50	7.88	6.71	6.10	8.96	9.54	10.39
CAN	15.51	15.31	14.80	14.80	14.80	8.69	8.18	8.10	8.10	8.10
CAP	7.23	6.67	6.80	6.96	7.31	6.63	5.98	9.01	9.66	10.49
HWN	5.96	5.50	5.31	5.38	5.39	6.29	5.61	5.96	6.01	6.04
HWP	9.30	8.58	8.69	8.79	9.05	7.55	6.62	9.32	9.73	11.22
HAN	15.59	15.27	14.72	14.68	14.64	9.48	8.90	9.14	9.19	9.19
HAP	18.49	17.82	15.97	15.73	15.40	10.51	10.11	12.58	12.74	13.99
TWN	9.24	8.53	8.57	8.23	8.18	7.45	6.45	7.51	7.75	8.12
TWP	13.52	12.27	11.04	11.01	10.72	9.95	8.52	11.10	11.43	13.00
TAN	17.41	16.81	16.19	16.16	15.98	10.73	10.05	10.92	11.12	11.29
TAP	20.73	19.64	17.39	16.85	15.99	12.14	11.51	13.96	14.07	15.34
SWN	10.81	9.90	9.36	9.41	9.36	8.95	8.13	9.04	9.20	9.41
SWP	12.06	11.01	11.03	11.13	11.60	10.51	8.93	11.73	12.11	13.39
SAN	12.54	12.37	12.07	11.93	11.89	8.10	7.91	7.97	7.97	7.97
SAP	14.95	14.71	14.40	14.25	14.29	9.11	9.18	12.19	12.62	13.12

Source: Compiled by the authors after applying the functions lphom, tslphom and nslphom of the R package `lphom`, attached as Supplementary Material to this paper, to the simulated data. The estimations labelled as ns_first have been obtained using nslphom with the argument min.first = T (nslphom algorithm 1 in Table 2) and the estimations labelled as ns_10 and ns_25 with the arguments min.first = F and, respectively, iter.max = 10 and 25 (nslphom algorithm 1 in Table 2).

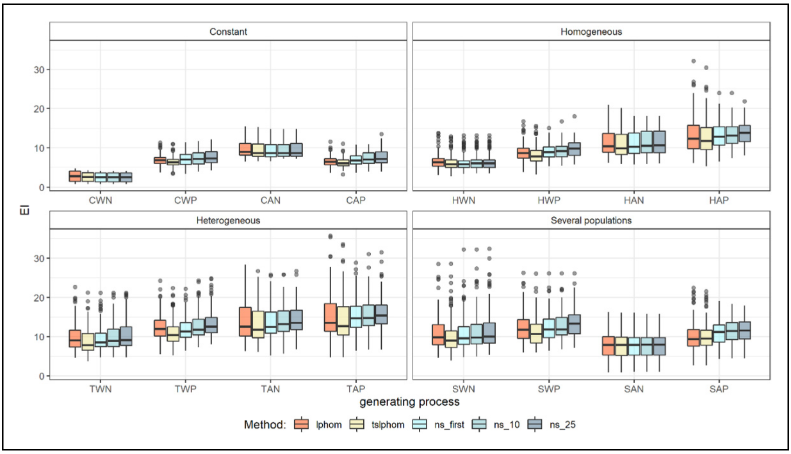

Figure 5. Box-plots of EI errors grouped by generating process and method, with 120 values in each set. Individual solutions have been attained using the functions lphom, tslphom and nslphom of the R package `lphom`, available as Supplementary Material to this paper. The estimations labelled as ns_first have been obtained using nslphom with the argument min.first = T (this corresponds to nslphom algorithm 1 in Table 1) and the estimations labelled as ns_10 and ns_25 with the arguments min.first = F and, respectively, iter.max = 10, and 25 (these correspond to different versions of the nslphom algorithm 2 in Table 1).

the worst scenarios. On average, EI (EPW) errors are under 15.5 (10.5) in the less accurate TAP scenarios. The simulations also show that it is worth including local data in the inference process when tslphom is used as prediction method: tslphom systematically outperforms lphom. However, we were unable to replicate with simulations the consistent results we attained with real data. Contrary to what was obtained with the real data, nslphom almost never yields the most accurate estimates in the fabricated scenarios. Clearly, the generating process considered could not capture the wild diversities that reign in the real world. Surprisingly, the presence of aggregation bias can sometimes improve accuracies, as happens with the generating processes based on several populations schemes. This paradoxical result is also reported in Klima et al. (2019).

The above analyses clearly show that the reliability of the estimates obtained with our methods tend to decrease with heterogeneity and aggregation bias. The point is that the (intensity of the) presence of these issues is unknown in real elections, so the question is whether there is a way to assess the accuracy of the estimates obtained without knowing the true values. Fortunately, the answer to this is yes. Given the high correlations (around 0.70) tying HETe and EI statistics, a high value for HETe could be interpreted as an indicator of inaccurate estimates. Indeed, the simulation-based approach suggested in Romero et al. (2020) could even be used to quantify it.

To finish the simulation analyses, we switch our focus to the assessment of the accuracies estimating unit transition matrices. To this end, we use the statistics EIw and EPWw which, defined as the natural extensions of EI and EPW, can be seen as weighted averages of EI and EPW unit-errors (see equations (S1) and (S2) in the Supplementary Material). Figures 6 and S4 to S6 and Tables S8 and S9 (in the Supplementary Material) summarise EIw and EPWw error measures. In general, a close relationship links global and unit errors: the greater the global error, the greater the unit errors. In the case of EI-type errors, the EIw errors tend to be higher than the EI errors due to the global compensating effects of under- and overestimates of counts at the local levels, which are more evident under P

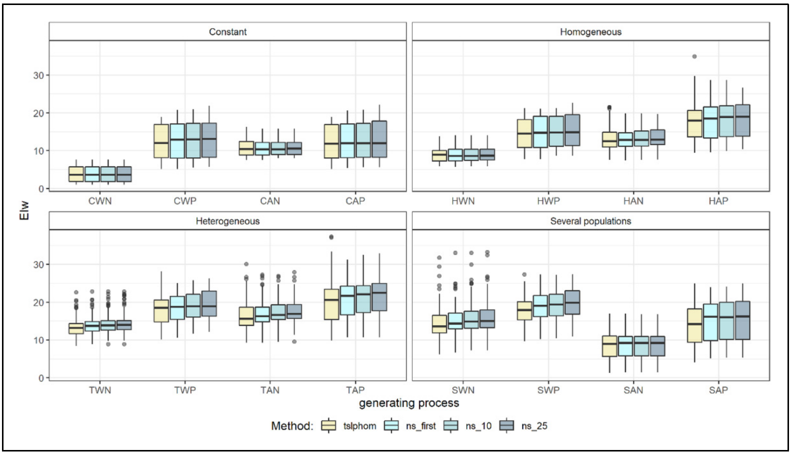

Figure 6. Box-plots of EIw error measures grouped by generating process and method, with 120 values in each set. Individual solutions have been attained using the functions tslphom and nslphom of the R package `lphom`, available as Supplementary Material to this paper. The estimations labelled as ns_first have been obtained using nslphom with the argument min.first = T (this corresponds to nslphom algorithm 1 in Table 1) and the estimations labelled as ns_10 and ns_25 with the arguments min.first = F and, respectively, iter.max = 10, and 25 (these correspond to different versions of the nslphom algorithm 2 in Table 1).

processing processes (where rates are contemplated as underlying probabilities). This is an expected result, given the greater heterogeneity that the Multinomial distribution insufflates at the local level. In the case of EPW-type errors, however, the above result does not hold. Finally, we also note that the number of units seems to have an impact on the ratios between local and global errors, EIw (EPWw) and EI (EPW). These ratios grow linearly with the log of the number of units.

8. Discussion and Further Research

In the previous sections, two new algorithms, tslphom and nslphom, have been developed and their global performance assessed with real and simulated data. These new algorithms, in addition to being satisfactorily accurate (even with relevant departures on the assumptions in which they rest), 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 a partially identified linear program. Although, fortunately, the second step of the lphom_local algorithm seems to resolve the indeterminacy, in our opinion, pursuing local adjustments could have been beneficial from a practical perspective, even with indeterminacies. As the computations with both real and simulated data show, introducing into the problem all the information available through local constraints is valuable. This increases the global accuracy of the reached point solution even under indeterminacy. We have verified this in the real datasets using different solvers after specifying some indeterminate algorithms as local adjusters, since 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.

Specifically, 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 (Berkelaar et al., 2020). Although the solvers programmed in linprog and lp consistently find different point solutions under indeterminacy (see Section S3 of the Supplementary Material), 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. This can be verified for lp with reference to Tables S10 and S11 of the Supplementary Material. 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 (see Tables S9 and S10 in the Supplementary Material).

Although the fact that the solution under indeterminacy is not unique could be seen as a drawback, the truth is that many of the current 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 to decide whether it could be used as a measure of uncertainty. Indeed, the fact that nslphom does not converge to a fixed point should not be perceived as a weakness by necessity: nslphom tends to quickly stabilize 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, after further consideration of the homogeneity hypothesis, 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 for the real datasets 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

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 the same 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 real transfer matrices using nslphom with 10 as the 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 needed 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.

Other ideas to improve the nslphom stopping rule that also deserve to be investigated include analysing and exploiting the properties related to the time series defined by the sequences $\{_t\hat{p}_{jk}\}_{t=1}^{ns}$ and/or $\{HETe_t\}_{t=1}^{ns}$. For example, we can borrow from the Bayesian approach and take for each (j, k)-pair the mean of the sequence $\{_t\hat{p}_{jk}\}_{t>t^*}^{ns}$ as solution; with t^* chosen large enough so as to guarantee that the series of estimates arrives at its stationary distribution. This strategy presumes the existence of a stationary distribution and that the algorithm is going to eventually reach it. Although in all the cases analysed we have observed the estimates quickly stabilizing around a distribution, we still lack formal proof of convergence. Therefore, this is an issue that should be addressed in the future. Despite this strategy presenting a more complex solution with higher computational cost than our proposals, 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 nlsphom estimates could be computed by mimicking the procedure proposed for lphom in Romero et al. (2020).

The above questions do not exhaust the possible future lines of research. In our approach, we reach estimates through an iterative procedure, an issue that can cast doubt on the optimality of the solutions. Ideally, as one reviewer suggests, one should aspire to have a model capable of estimating the global and local matrices jointly. In our opinion, the problem here lies in finding a full linear program specification that includes all equations and unknowns in a system with just one unique solution. At the moment, we cannot envisage such a specification, but perhaps other research studies could achieve this. We would focus the research on how to define the objective function. This issue could, perhaps, be solved if the two systems of the lphom_local model could be stacked in only a linear program with the solution reached in just one step.

9. Conclusions

The estimation of RxC ecological inference contingency tables from aggregate results is one of the most salient and challenging problems in the field of quantitative social sciences. During the past quarter-century, the ecological regression (statistical) 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 in the homogeneity hypothesis, enables estimates to be attained of the joint cross-distributions of each unit in which the whole population is split out. Two new ecological inference algorithms, tslphom and nslphom, are built grounded in this.

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. Furthermore, in an extensive simulation study, which helps to delineate the limitations of the approaches, we also find that despite the accuracy of the estimates tending to decrease with heterogeneity and aggregation bias, they still look satisfactory even in the worst scenarios considered.

In short, the new algorithms, being at least as accurate as their best competitors from the statistical framework, 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 different 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 available in CRAN (cran.r-project.org/web/packages/lphom/index.html).

Acknowledgments

The authors wish to thank the chief editor and three anonymous reviewers for their valuable comments and suggestions, the managing editor for her help, and 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. The usual disclaimers apply.

Authors' note about reproducibility

The data, programs, (ad-hoc) codes and outputs to reproduce (via the statistical software R) all the results (including tables, figures and statistical comments) of this research, presented in both the paper and the supplementary material, are available in <https://links.uv.es/72uQiop>, DOI: 10.17605/OSF.IO/DY2SE.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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 Consellería d'Innovació, Universitats, Ciència i Societat Digital, Generalitat Valenciana [grant numbers AICO/2019/053 and AICO/2021/257].

ORCID iD

Jose M. Pavía https://orcid.org/0000-0002-0129-726X

Supplemental material

Supplemental material for this article is available online.

Notes

1. 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 (statistical) approaches which, moreover, do not suffer from the second weakness.
2. Under Manski conceptual framework, the lphom algorithm could be observed as the method that, without assuming distributional statistical assumptions, solves the problem of finding the p_{jk}'s that minimize, within the region of feasible solutions, the expected value of the weighted absolute loss function (with weights proportional to the sizes of the units) between observed and expected counts. It should be noted, however, that to observe our approaches under this framework, we should acknowledge that the basic unknowns (the p_{jk}'s) are underlying probabilities and also that we are under a super-population scheme in which the observed election records represent one of the possible outcomes that could have been observed if the election were repeated a large number of times under similar conditions.
3. Note that in the case of CWN and CAN generating processes once the reference election is set there is no variability among datasets.

References

Barreto, M., L. Collingwood, S. Garcia-Rios, and K. A. R. Oskooii. 2022. "Estimating Candidate Support in Voting Rights Act Cases: Comparing Iterative EI and EI-R_C Methods." *Sociological Methods & Research* 51:271–304.

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.

Cho, W. K. T. and C. F. Manski. 2008 "Cross Level/Ecological Inference." pp. 547–69 in *The Oxford Handbook of Political Methodology*, edited by

J. M. Box-Steffensmeier, H. E. Brady, and D. Collier. New York: Oxford University Press.

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* 8:122–36.

Duncan, O. and B. Davis. 1953. “An Alternative to Ecological Correlation.” *American Sociological Review* 18:665–6.

Forcina, A. and D. Pellegrino. 2019. “Estimation of Voter Transitions and the Ecological Fallacy.” *Quality & Quantity* 53:1859–74.

Goodman, L. A. 1953. “Ecological Regressions and the Behaviour of Individuals.” *American Sociological Review* 18:663–4.

Goodman, L. A. 1959. “Some Alternatives to Ecological Correlation.” *American Journal of Sociology* 64(6):610–25.

Gosnell, H. F. and N. N. Gill. 1935. “An Analysis of the 1932 Presidential Vote in Chicago.” *The American Political Science Review* 29:967–84.

Gosnell, H. F. and M. J. Schmidt. 1936. “Factorial and Correlational Analysis of the 1934 Vote in Chicago.” *Journal of the American Statistical Association* 31:507–18.

Greiner, D. J. and K. M. Quinn. 2009. “R × C Ecological Inference: Bounds, Correlations, Flexibility, and Transparency of Assumptions.” *Journal of the Royal Statistical Society, A* 172:67–81.

Greiner, D. J. and K. M. Quinn. 2010. “Exit Polling and Racial Bloc Voting: Combining Individual-Level and RxC Ecological Data.” *The Annals of Applied Statistics* 4:1774–96.

Haneuse, S. and J. Wakefiled. 2004 “Ecological Inference Incorporating Spatial Dependence.” Pp. 266–301 in *Ecological Inference. New Methodological Strategies*, edited by G. King, O. Rosen, and M. Tanner. New York: Cambridge University Press.

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

Imai, K. and K. Khanna. 2016. “Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records.” *Political Analysis* 24:263–72.

Imai, K., Y. Lu, and A. Strauss. 2008. “Bayesian and Likelihood Inference for 2 x 2 Ecological Tables: An Incomplete Data Approach.” *Political Analysis* 16:41–69.

Irwin, G. and D. Meeter. 1969. “Building Voter Transition Models from Aggregate Data.” *Midwest Journal of Political Science* 13:545–66.

Johnston, R. J. and A. M. Hay. 1983. “Voter Transition Probability Estimates: An Entropy-Maximizing Approach.” *European Journal of Political Research* 11:93–8.

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. 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., 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.

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* 100:133–59.

Manski, C. F. 2003. *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

Manski, C. F. 2007. *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

McCarthy, C. and T. M. Ryan. 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. and I. Goltra. 1919. "How Women Vote." *Political Science Quarterly* 34:413–33.

Ogburn, W. F. and N. S. Talbot. 1929. "A Measurement of the Factors in the Presidential Election of 1928." *Social Forces* 8:175–83.

Park, W., M. J. Hanmer, and D. R. Biggers. 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. 2022. "ei.Datasets: Real Datasets for Assessing Ecological Inference Algorithms." *Social Science Computer Review* 40:247–60.

Pavía, J. M. and C. Aybar. 2020. "Electoral Mobility in the 2019 Elections in the Valencian region." *Debats. Journal on Culture, Power and Society* 134:27–51.

Pavía, J. M. and R. Romero. 2022. 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* 100:1113–22.

Plescia, C. and L. De Sio. 2018. "An Evaluation of the Performance and Suitability of RxC Methods for Ecological Inference with Known True Values." *Quality & Quantity* 52:669–83.

Puig, X. and J. Ginebra. 2015. “Ecological Inference and Spatial Variation of Individual Behavior: National Divide and Elections in Catalonia.” *Geographical Analysis* 47:262–83.

Robinson W. S. 1950. “Ecological correlations and the behavior of individuals.” *American Sociological Review*, 15:351–57.

Romero, R. and J. M. Pavía. 2021. “Estimating Vote Party Entries and Exits by Ecological Inference. Mathematical Programming Versus Bayesian Statistics.” *BEIO* 34:85–97.

Romero, R., J. M. Pavía, J. Martín, and G. Romero. 2020. “Assessing Uncertainty of Voter Transitions Estimated from Aggregated Data. Application to the 2017 French Presidential Election.” *Journal of Applied Statistics* 47:2711–36.

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.

Tziafetas, G. 1986. “Estimation of the Voter Transition Matrix.” *Optimization* 17:275–9.

Upton, C. J. G. 1978. “A Note on the Estimation of Voter Transition Probabilities.” *Journal of the Royal Statistical Society, A* 141:507–12.

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

Vangrevelinghe, G. 1961. “Étude Statistique Comparée des Résultats des Référendums de 1958 et 1961.” *Revue de Statistique Applique* 9:83–100.

Wakefield, J. 2004. “Ecological Inference for 2 x 2 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*. Baltimore County, Baltimore, MD: Department of Mathematics and Statistics, University of Maryland.

Author Biographies

Jose M. Pavía, MSc in Maths and PhD in Economics and Business Science, is Quantitative Methods Professor in the Economics Faculty of the Universitat de Valencia, Spain. Pavía develops and applies statistical and machine learning methods in many areas of social sciences. With broad and varied research interests, his work focuses on the search for innovations that bridge the gap between theory and practical applications and include issues related to electoral processes, prediction, statistical (machine) learning, ecological inference, behavioural economics, public policy evaluation, crime detection, survey research, sampling, public opinion, experiments, regional economy and inequality. He is director of the Elections and Public Opinion Research Group of the Universitat de Valencia and Chair of Catedra Deblanc, for the application of statistical, economic and machine learning to detect money laundering and financial crimes. More information on him and his research publications is available at <http://go.uv.es/ZfX6E7w>.

Rafael Romero, PhD in Agricultural Engineering, is a (retired) Professor of the Department of Statistics and Operations Research at the Polytechnic University of Valencia. Romero applies advanced statistical and operational research methods in many areas. He has acted as a Consultant for the World Bank (developing mathematical models for hydro-agricultural planning of large areas in Tunisia and Senegal), for IBM Spain (applying data processing for fruit and vegetable marketing), for Aerolineas Argentinas (building mathematical models for the establishment of activity programs and fleet renewal), as well as many other companies. His two current research areas are focused on quality control in the steel industry, where he collaborates with the main Spanish and Portuguese companies in the sector, and on the application of mathematical optimisation models for the analysis of the outcomes of electoral processes.