Journal of the Royal Statistical Society Series A: Statistics in Society, 2024, **187**, 919–943
https://doi.org/10.1093/jrsssa/qnae013
Advance access publication 29 February 2024
Original Article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Symmetry estimating R × C vote transfer matrices from aggregate data

Jose M. Pavía[1] **and Rafael Romero**[2]

[1]GIPEyOP, Departamento de Economía Aplicada, Universitat de Valencia, Valencia, Spain
[2]Departamento de Estadística, Investigación Operativa y Calidad, Universidad Politécnica de Valencia, Valencia, Spain

Address for correspondence: Jose M. Pavía, GIPEyOP, Facultat d'Economía, Universitat de Valencia, 46022-Valencia, Spain. Email: pavia@uv.es

Abstract

Ecological inference methods are devised to estimate unknown inner-cells of 2-way contingency tables by inferring conditional distribution probabilities. This outlines one of the more long-standing social science problems, chiefly frequent in political science and sociology. To solve the problem, ecological inference algorithms consider an asymmetric relationship, with a main characteristic (e.g. race or social class) mapped to rows impacting on a dependent variable, usually the vote, mapped to columns. The problem arises because different solutions are reached depending on how variables are assigned to rows and columns. The models are asymmetric. In this paper, we propose 2 new sets of ecological inference algorithms and explore if accuracy could be improved by handling the problem in a symmetric way. We assess the accuracy of the proposed methods using real data from more than 550 concurrent elections where the true district-level cross-classifications of votes (straight- and split-tickets) are known. Our empirical assessment clearly identifies the symmetric solutions as more accurate. They outperform asymmetric methods 90% of the time and reduce error, on average, by 11%. Our results are based on data from simultaneous elections, so further research is required to see whether our conclusions can be maintained in other ecological inference contexts. Interested readers can easily use the proposed methods as they are implemented in the R package lphom.

Keywords: contingency tables, ecological inference, linear programming, lphom, simultaneous elections, voter transitions

1. Introduction

In elections, aggregate-level data are abundant and reliable whereas individual-level data are largely unavailable due to the secret ballot, and when they are available they tend to be inexact and scarce to estimate voter transitions (changes in voters' choices between elections) as they usually originate from polls (Romero et al., 2020). Despite this, many agents, including political parties, the media, and US legal practitioners of voting rights, are still interested in knowing the voting behaviour of different subgroups of people. Ecological inference has been devised to solve this issue by exploiting the aggregate-level data relationships (King, 1997).

Ecological inference refers to the process of inferring individual-level relationships from aggregate ('ecological') data when individual-level data are not available. It is routinely employed in many disciplines, from economics and epidemiology to sociology and political science (Pavía, 2022), despite being exposed to the so-called ecological fallacy (Robinson, 1950). For instance, ecological inference algorithms are used to estimate vote transfer matrices between elections, infer split-ticket voting behaviours or reveal social and racial voting patterns (Barreto et al., 2022; O'Loughlin, 2000; Park et al., 2014; Romero & Pavía, 2021).

Using a classical contingency table representation in which individuals are classified by rows according to the groups to which they belong (e.g. social class, previous vote or race) and by columns by, for instance, their votes, the unknown cross-distributions are estimated using as data the

Received: March 24, 2022. **Revised:** January 8, 2024. **Accepted:** February 4, 2024

observed row and column marginal distributions in a set of non-overlapping (geographical) units (e.g. precincts). The difficulty arises because substantially different inner cell counts can give rise to the same aggregated row and column totals, with this indeterminacy cannot being completely removed. It is intrinsic to the problem (see, e.g. Forcina & Pellegrino, 2019; Greiner, 2007; Manski, 2007). The solution depends on the assumptions made, which in a large extent are unverifiable with the available information (Gelman et al., 2001; Glynn & Wakefield, 2010; Wakefield, 2004), an issue that has led to festering disputes (see, e.g. Freedman et al., 1998, 1999; Greiner, 2007; Tam Cho, 1998; Tam Cho & Gaines, 2004). Practitioners customarily hypothesize similar/related conditional row (underlying) probability/fraction distributions across tables (sometimes with the help of covariates) relying on the common observation that people belonging to the same group tend to follow, probabilistically, similar behaviour patterns within the same political context (Pavía & Romero, 2022).

Although an extensive list of different ecological inference procedures has been suggested over time from fields as diverse as frequentist and Bayesian statistics (e.g. Brown & Payne, 1986; Goodman, 1953, 1959; Greiner & Quinn, 2009; King, 1997; King et al., 2004; Klima et al., 2019; Rosen et al., 2001), mathematical optimization (e.g. Hawkes, 1969; Pavía & Romero, 2022; Tziafetas, 1986), or information theory (e.g. Bernardini-Papalia & Fernández-Vázquez, 2020; Judge et al., 2004), they all consider analogous underlying assumptions and share a similar framework. This framework is based on a hypothesis of similarity/relationship in (electoral) behaviour by group across units and a scheme with an explanatory and a response variable.

Ecological inference may, therefore, be observed as an inverse problem where the goal is estimating (inferring) the conditional row fraction (underlying probability) distributions using the observed count marginal distributions as data (Jiang et al., 2020). Once the estimates of probability distributions are attained, cross-classification estimates of counts are obtained by multiplying the observed row margins and the estimated probabilities. In some models, such as in Greiner and Quinn (2009) and in its extension (Greiner & Quinn, 2010), counts are directly inferred.

The above general scheme means that different estimates are reached depending on which variable is assigned to rows and which to columns. In other words, even using the same method and the same data, a different solution is obtained if rows and columns are flipped. Thomsen's model (1987), which assumes that voters' choices are driven by individual latent factors, is the exception. The models proposed by Johnston and colleagues (e.g. Johnston et al., 1983; Johnston & Pattie, 2003), based on entropy maximization, while treating tables symmetrically, cannot be classified as genuine ecological inference approaches since they require prior information about the target cross-distributions to be applied.

In many applications, deciding which classification should be assigned to rows and which to columns is straightforward. For example, in a study of polarized voting in which we want to know how different collectives support different candidates, characteristics of the electorate (such as race or gender) are naturally assigned to rows while the categories of the columns are defined by the candidates. Equally, if we want to know the levels of voters' loyalty (and switching) by party between two consecutive elections, the natural way is to assign the electoral options of the first and second elections to rows and columns, respectively.

In the above examples, the implicit assumption is that somehow there is a causality relationship or, at least, a natural origin-destination temporal arrangement. But what happens in simultaneous elections? In this case, the answer might not be so straightforward.

When two elections are held simultaneously, it is usually considered that there is a first-order election and a subordinate second-order election, so the relationship is studied in that order. It is implicitly accepted that the majority of voters make a sequential choice: they first decide their vote for the first-order election to subsequently, in a second step, choose their vote for the subordinate election (Pavía & Cantarino, 2017). Sometimes, the order of primacy is clear, such as when a general election and a referendum coincide. Other times, however, such as when electors vote simultaneously in a national and in a regional election or for a party list and for a candidate, this is not so clear even though we can argue over an order of primacy. Indeed, the argument itself, the existence of a hierarchy between elections, is being progressively challenged by the literature (see, e.g. Schakel & Romanova, 2020). Anyway, although we can reasonably argue the

presence of a first-order and a second-order election, the question is whether this is the proper way to proceed. In other words, can we improve the accuracy of the estimated counts by considering both elections at the same level? More generally, given that (almost) all the ecological inference models exploit correlations and not causal relationships, we should ask ourselves whether the average accuracy of ecological inference estimates may be improved using methods that deal with rows and columns symmetrically. The aim of this paper is to provide answers to these questions.

To answer these questions, we propose two new sets of methods whose solutions do not depend on how classifications are assigned to rows and columns. On the one hand, we propose a new set of algorithms that achieve their solutions handling rows and columns symmetrically from the outset, by definition. This type of models can be logically specified from a mathematical optimization framework. Both dual constraints as well as congruency constraints can be naturally introduced in a linear programme, within the same optimization problem. Hence, in this research we adopt a linear programming approach. It is important to note that, although these models could also be specified within a quadratic programming framework, we prefer to state our models within the linear framework because, as Tziafetas (1986) shows, linear approaches are more efficient than quadratic ones in this context. On the other hand, we also consider a group of algorithms based on asymmetric solutions. This involves methods that reach their estimates by combining the solutions attained after applying an asymmetric ecological inference model to the two possible ways of assigning classifications to rows and columns.

In a recent paper, Pavía and Romero (2022) propose two new models, tslphom and nslphom, that, according to the authors, 'place the linear programming approach once again in a prominent position in the ecological inference toolkit'. These new models, based on linear programming, generate estimates for each unit table and solve the tendency of lphom, the basic linear programming model, to estimate extreme probabilities (zeros and ones). They have also been shown to be at least as accurate and significantly simpler to use (Pavía & Romero, 2023) than the multimonial-Dirichlet $R \times C$ ecological inference statistical model, previously identified as the best in the literature (Katz, 2014; Klima et al., 2016; Plescia & De Sio, 2018). In this paper, we propose new algorithms that deal with rows and columns symmetrically by building on the Pavía and Romero (2022) models in two directions.

On the one hand, using either lphom, tslphom, or nslphom, we suggest three new methods, each of them generating three reasonable solutions, by solving the two underlying dual problems, swapping origin and destination. We identify these new methods by adding the suffix '_dual' to the corresponding base algorithm and their respective new solutions by adding the suffixes '_min', '_dual_a', and '_dual_w'. On the other hand, we directly modify lphom, tslphom, and nslphom, looking for the joint congruent optimal solution of both dual problems. We specify new linear programmes in which the optimal solutions, being congruent, simultaneously verify the constraints imposed by the two dual ecological inference specifications. We identify these new models and their corresponding solutions adding the suffix '_joint' to the respective base algorithm. The main innovation of this paper, therefore, lies in proposing, as far as we know for the first time in the literature, ecological inference models that explicitly deal with rows and columns in a symmetric fashion. Conceptually, this is a novelty that could also be explored from other frameworks.

The performance of all the proposed algorithms is assessed using the data available in the R package ei.Datasets (Pavía, 2022), which contains the global true party-candidate cross-classification tables corresponding to more than 500 mixed-member elections. After comparing actual tables and estimates from all the new algorithms and also considering, as baseline, the solutions obtained using the base asymmetric models, we can assert that, at least for these datasets, more accurate solutions are obtained exploiting the information both ways: parties to candidates and candidates to parties.

The rest of the paper is organized as follows. Section 2 states the problem and introduces the notation. Section 3 details the methods based on lphom, tslphom, and nslphom proposed to obtain the dual solutions. Section 4 revises the basic linear programme model (lphom), adapting it to the symmetric case. Section 5 goes further and reworks tslphom and nslphom after modifying the lphom_local algorithm of Pavía and Romero (2022) to make it symmetric. Section 6 presents the results and discusses the findings of the empirical comparisons. Section 7 deepens and explores

on the factors impacting on the accuracy of the estimates. Section 8 concludes, discusses, and suggests directions for further research.

2. Mathematical representation of the problem

For convenience and without loss of generality, from now on we adopt a framework of simultaneous elections where voters, grouped in a set of units that jointly define a partition of the electoral space, cast two votes; one for a party list and another for a candidate. We denote by I, J, and K the number of units, parties, and candidates, respectively.

The observed data are the votes x_{ij} recorded for each of the $j = 1, \ldots, J$ parties and the votes y_{ik} gained for each of the $k = 1, \ldots, K$ candidates in each of the $i = 1, \ldots, I$ voting units. The basic unknowns are the number of voters v_{jk} who simultaneously voted for party j and candidate k in the entire population. Of interest are also the equivalent numbers in each polling unit v^i_{jk}.

As intermediate $J \times K$ and $K \times J$ unknowns, we denote p_{jk} as the proportion of voters in the entire electoral space who voted for candidate k among those who voted for party j and, equivalently, q_{kj} as the proportion of voters in the entire electoral space who voted for party j among those who voted for party k. Equally, we define the corresponding proportions at the unit level as p^i_{jk} and q^i_{kj}. Note that if we consider a probabilistic approach, the intermediate unknowns could be expressed by: $p_{k|j} = P(Y = k|X = j)$ and $q_{j|k} = P(X = j|Y = k)$, with X and Y representing the party and candidate vote variables, respectively. Similarly, one can define probabilities at the unit level. Ecological inference algorithms are designed to estimate these intermediate unknowns. The p_{jk} (and p^i_{jk}) proportions are the usual outputs when the x_{ij} and y_{ik} votes are assigned, respectively, to rows and columns while the q_{kj} (and q^i_{kj}) proportions are the outputs in the dual case, transposing rows and columns; see Figure 1.

All the above quantities are closely related. Among other relationships, the following equalities hold: $x_{ij} = \sum_k v^i_{jk}$, $y_{ik} = \sum_j v^i_{jk}$, $v_{jk} = \sum_i v^i_{jk}$, $p_{jk} = v_{jk} / \sum_i x_{ij}$, $q_{kj} = v_{jk} / \sum_k y_{ik}$, $p^i_{jk} = v^i_{jk}/x_{ij}$, and $q^i_{kj} = v^i_{jk}/y_{ik}$. Given these relationships, once we have estimates for $\hat{p}_{jk}$, $\hat{q}_{kj}$, $\hat{p}^i_{jk}$, and $\hat{q}^i_{kj}$ estimates can be directly obtained for $\hat{v}_{jk}$ and $\hat{v}^i_{jk}$ employing the observed quantities x_{ij} and y_{ik}.

In general, different estimates for v_{jk} and v^i_{jk} are obtained depending on the assignment of parties and candidates to rows and columns. We define an algorithm as symmetric if it produces the same estimates when rows and columns are interchanged: $x_{ij}\hat{p}^i_{jk} = y_{ik}\hat{q}^i_{kj}$. These methods are characterized by the fact that their row-conditional probability estimates verify the Bayes theorem with observed proportions, i.e. $\hat{p}^i_{jk}\, p^i_{j\cdot} = \hat{q}^i_{kj}\, q^i_{\cdot k}$, where $p^i_{j\cdot} = x_{ij} / \sum_r x_{ir}$ and $q^i_{\cdot k} = y_{ik} / \sum_c y_{ic}$. We refer to this property as the *probabilistic symmetric condition*.

In matrix form, we denote by $\boldsymbol{X} = [x_{ij}]$ the $I \times J$ observed matrix of party votes, by $\boldsymbol{Y} = [y_{ik}]$ the $I \times K$ observed matrix of candidate votes, by $\boldsymbol{P} = [p_{jk}]$ the $J \times K$ unknown matrix of parties-to-candidates global proportions, by $\boldsymbol{P}^i = [p^i_{jk}]$ the $J \times K$ unknown matrix of parties-to-candidates matrix proportions of unit i, by $\boldsymbol{Q} = [q_{kj}]$ the $K \times J$ unknown matrix

Parties to candidates vote transfer matrix

	C_1	...	C_k	...	C_K	
P_1	p^i_{11}	...	p^i_{1k}	...	p^i_{1K}	x_{i1}
...	...	...	...	...	...	...
P_j	p^i_{j1}	...	p^i_{jk}	...	p^i_{jK}	x_{ij}
...	...	...	...	...	...	...
P_J	p^i_{J1}	...	p^i_{Jk}	...	p^i_{JK}	x_{iJ}
	y_{i1}	...	y_{ik}	...	y_{iK}	

Candidates to parties vote transfer matrix

	P_1	...	P_j	...	P_J	
C_1	q^i_{11}	...	q^i_{1j}	...	q^i_{1J}	y_{i1}
...	...	...	...	...	...	...
C_k	q^i_{k1}	...	q^i_{kj}	...	q^i_{kJ}	y_{ik}
...	...	...	...	...	...	...
C_K	q^i_{K1}	...	q^i_{Kj}	...	q^i_{KJ}	y_{iK}
	x_{i1}	...	x_{ij}	...	x_{iJ}	

Figure 1. The two dual ways of representing the basic problem in a typical i unit: parties to candidates (left panel) and candidates to parties (right panel). Inner quantities are the unobserved proportions, the (intermediate) targets of ecological inference. In general, different solutions are reached solving the problem from parties-to-candidates than from candidates-to-parties. Symmetric algorithms generate congruent $\boldsymbol{P}^i$, $J \times K$ and $\boldsymbol{Q}^i$, $K \times J$ matrices: $x_{ij}\hat{p}^i_{jk} = y_{ik}\hat{q}^i_{kj}$.

of candidates-to-parties matrix of global proportions, and by $\boldsymbol{Q}^i = [q^i_{kj}]$ the $K \times J$ unknown matrix of candidates-to-parties matrix proportions of unit i. We will use hat-symbols to represent estimates of the corresponding unknown proportions and matrices.

Likewise, we denote by $\boldsymbol{V} = [v_{jk}]$ and $\boldsymbol{V}^i = [v^i_{jk}]$, respectively, the global and unit i unknown $J \times K$ matrices (contingency tables) of votes, by $\widehat{\boldsymbol{V}} = \left[\hat{p}_{jk} \sum_i x_{ij}\right]$ and $\widehat{\boldsymbol{V}}^i = [\hat{p}_{jk} x_{ij}]$ the corresponding estimates obtained using parties-to-candidates proportion estimates, and by $\widetilde{\boldsymbol{V}} = \left[\hat{q}_{kj} \sum_k y_{ik}\right]^T$ and $\widetilde{\boldsymbol{V}}^i = [\hat{q}_{kj} y_{ik}]^T$ the ones attained using candidates-to-parties proportion estimates, where the superscript T identifies transpose.

Additionally, we use $\boldsymbol{I}_n$ for naming the identity matrix of order n, $\boldsymbol{0}_{n\times m}$ and $\boldsymbol{1}_{n\times m}$ for representing the $n \times m$ matrices of zeros and ones, respectively, and diag($\boldsymbol{a}$), row$_h$($\boldsymbol{A}$), and vec($\boldsymbol{A}$) for noting the diagonalization, row-extractor, and matrix-vectorization operators. They operate as follows. Given a vector $\boldsymbol{a}$ of size n and a matrix $\boldsymbol{A}$ of order $n \times m$, diag($\boldsymbol{a}$) builds a matrix of order n with $\boldsymbol{a}$ in its main diagonal, row$_h$($\boldsymbol{A}$) extracts the hth-row of $\boldsymbol{A}$ as a column vector of order m and vec($\boldsymbol{A}$) is the $nm \times 1$ column vector obtained by stacking the rows of the matrix $\boldsymbol{A}$ on top of one another in row-wise order. We use $\otimes$ and $\odot$ to denote, respectively, the Kronecker product and the Hadamard (or matrix element-wise multiplication) product operators.

3. Solutions based on asymmetric methods

Given that the symmetric models we propose build on the lphom, tslphom, and nslphom algorithms, we limit ourselves to using the same framework for defining our asymmetric-based dual methods. In our view, this makes comparisons fairer. This auto-constraint could of course be relaxed and asymmetric-based dual solutions be defined using as building blocks other R × C ecological inference procedures, such as the Rosen et al. (2001) Bayesian-based multinomial-Dirichlet model (Lau et al., 2020), the iterative version of the 2×2 model proposed by King (Collingwood et al., 2020; King, 1997), or the generalization of the Goodman regression method (Collingwood et al., 2016, 2020), to name just some possibilities.

Before defining our asymmetric-based dual methods in Section 3.2, we first detail the lphom, tslphom, and nslphom models in Section 3.1. To do this, we focus on the nslphom model given that, although lphom, tslphom, and nslphom were suggested as different models, as Pavía and Romero (2023) shows, the lphom and the tslphom solutions can be attained as by-products of the nslphom algorithm, these being intermediate outputs of its iterative process: lphom after solving its first linear system (iteration 0) and tslphom after solving its firsts $2I + 1$ linear systems (iteration 1).

In this section, we introduce the equations when x_{ij} and y_{ik} are assigned, respectively, to rows and columns (see left panel in Figure 1). The equations of the dual problem (see right panel of Figure 1) are detailed in Sections 4 and 5, during the process of specifying the genuine symmetric models. There, we will adopt a more compact matrix representation. Despite the density of some of the mathematical formulae, we will include all the equations of the models because, in our view, this is the more precise way of defining them. Nevertheless, for those readers that prefer to skip the more mathematical details, we also offer explanations and intuitions for all of the equations as well as insights about the statistical and rational basis of the proposed algorithms.

3.1 Asymmetric linear programming models

The nslphom procedure is an iterative algorithm that, for simultaneous elections and typical ecological inference problems, uses equations (1)–(15) to attain its solution. In its iteration zero, nslphom solves the basic lphom linear system (Romero et al., 2020) defined by equations (1)–(5)—whose unknowns are the p_{jk} unobserved global proportions as well as the (non-negative) auxiliary quantities e^+_{ik} and e^-_{ik}. The lphom model estimates the p_{jk} values by solving a linear programme in which the solutions, besides fulfilling the obvious constraints of sum 1 for each row of the district-level proportion matrix and non-negativity (equations (2) and (1)), must be perfectly congruent to the observed outcomes at the global level (equation (3)) and minimize the sum of L_1-discrepancies at the unit level (equations (4) and (5)). Solving of the system (1)–(5) generates an initial solution, $\hat{\boldsymbol{P}}_0 = [{}_0\hat{p}_{jk}]$, of the matrix of global proportions, $\boldsymbol{P}$, which is used to start up

the iterative process that characterizes nslphom. The estimate $\hat{\boldsymbol{P}}_0$ is the (only) matrix that in addition to verifying all the aggregate row and column constraints minimizes the sum across units of the expected count discrepancies, in L_1-norm, between the observed and expected unit column totals when $\hat{\boldsymbol{P}}_0$ is applied to the observed unit row totals.

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

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

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

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

$$\text{minimize} \quad \sum_{i=1}^{I} \sum_{k=1}^{K} (e_{ik}^{+} + e_{ik}^{-}). \tag{5}$$

In the next iterations, for $l = 1, \ldots, ns$ (where ns is a fixed exogenous parameter), the algorithm generates recursive estimates of the unit proportions, p_{jk}^{i}, and of the global proportions, p_{jk}, through a two-step process. In the first step, the current available estimates of the global proportions, ${}_{l-1}\hat{p}_{jk}$, are employed to obtain unit proportion estimates, ${}_{l}\hat{p}_{jk}^{i}$, by solving the $2I$ sequential linear systems defined, respectively, by equations (6)–(10) and (11)–(13)—where ε_{jk}^{i+}, ε_{jk}^{i-}, δ_{jk}^{i+}, and δ_{jk}^{i-} are auxiliary (non-negative) unknowns. In particular, the observed votes in each unit together with the estimated global proportions ${}_{l-1}\hat{p}_{jk}$ are used as inputs to solve in each unit another linear programme in which the unit level proportions, p_{jk}^{i}, are estimated imposing a perfect match among the observed outcomes at the unit level (equation (8)) by minimizing the sum of the L_1-distances between the values that would be obtained by applying to x_{ij} either p_{jk}^{i} or ${}_{l-1}\hat{p}_{jk}$ (equations (9) and (10)). As this programme is indeterminate, the final solutions for p_{jk}^{i} are reached through a new linear programme (equations (11)–(13)) in which the new linear programme chooses among all the solutions of the (6)–(10) linear programme (equation (11)) the set of p_{jk}^{i} that minimizes the sum of the L_1-distances between all the $(p_{jk}^{i}, {}_{l-1}\hat{p}_{jk})$ pairs (equations (12) and (13)).

$$p_{jk}^{i} \geq 0 \quad \text{for } j = 1, \ldots, J, \ k = 1, \ldots, K, \ i = 1, \ldots, I, \tag{6}$$

$$\sum_{k=1}^{K} p_{jk}^{i} = 1 \quad \text{for } j = 1, \ldots, J, \ i = 1, \ldots, I, \tag{7}$$

$$\sum_{j=1}^{J} x_{ij} \, p_{jk}^{i} = y_{ik} \quad \text{for } k = 1, \ldots, K, \ i = 1, \ldots, I, \tag{8}$$

$$({}_{l-1}\hat{p}_{jk} - p_{jk}^{i}) x_{ij} = \varepsilon_{jk}^{i+} - \varepsilon_{jk}^{i-} \quad \text{for } j = 1, \ldots, J, \ k = 1, \ldots, K, \ i = 1, \ldots, I, \tag{9}$$

$$\text{minimise } w = \sum_{j=1}^{J}\sum_{k=1}^{K}(\varepsilon_{jk}^{i+} + \varepsilon_{jk}^{i-}) \quad \text{for } i = 1, \ldots, I, \tag{10}$$

$$w = \sum_{j=1}^{J}\sum_{k=1}^{K}(\varepsilon_{jk}^{i+} + \varepsilon_{jk}^{i-}) \quad \text{for } i = 1, \ldots, I, \tag{11}$$

$$p_{jk}^{i} = {}_{l-1}\hat{p}_{jk} + \delta_{jk}^{i+} - \delta_{jk}^{i-} \quad \text{for } j = 1, \ldots, J, \ k = 1, \ldots, K, \ i = 1, \ldots, I, \tag{12}$$

$$\text{minimize } \sum_{j=1}^{J}\sum_{k=1}^{K}(\delta_{jk}^{i+} + \delta_{jk}^{i-}) \quad \text{for } i = 1, \ldots, I. \tag{13}$$

In the second step, the unit proportion estimates are in turn used to update the global proportion estimates, ${}_{l}\hat{p}_{jk}$, via equation (14), where the global proportions are expressed as a weighted average of unit proportions.

$${}_{l}\hat{p}_{jk} = \sum_{i=i}^{I} {}_{l}\hat{p}_{jk}^{i}\, x_{ij} \Big/ \sum_{i=i}^{I} x_{ij} \quad \text{for } j = 1, \ldots, J, \ k = 1, \ldots, K. \tag{14}$$

In each iteration, the statistic defined by equation (15), which measures the distance to homogeneity of the lth-solution, is also calculated. This measure is employed to select the nslphom solution (Pavía & Romero, 2022), which corresponds to the estimates associated with the iteration l^* that minimizes (15).

$$HETe_l = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} |{}_{l}\hat{p}_{jk}^{i} x_{ij} - {}_{l}\hat{p}_{jk}\, x_{ij}|}{2\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}}. \tag{15}$$

Once the iterative process is finished, nslphom provides the lphom, tslphom, and nlsphom solutions. The matrix $\hat{\boldsymbol{P}}_0$ corresponds to the lphom solution. The set of matrices $\hat{\boldsymbol{P}}_1$ and $\hat{\boldsymbol{P}}_1^{\boldsymbol{i}} = [{}_1\hat{p}_{jk}^{i}]$, for $i = 1, \ldots, I$, achieved at the end of iteration 1 defines the tslphom solution. And the nslphom solution is composed of the set of matrices $\hat{\boldsymbol{P}}_{l^*}$ and $\hat{\boldsymbol{P}}_{l^*}^{\boldsymbol{i}} = [{}_{l^*}\hat{p}_{jk}^{i}]$, for $i = 1, \ldots, I$, selected taking into account (15). Note that the lphom solution only contains estimates of the global proportions.

The above algorithm, programmed in the function nslphom of the R package lphom (Pavía & Romero, 2022), has one parameter to be chosen: the number of iterations, *ns*. The default option of the nslphom function considers only 10 iterations since the minimum of equation (15) is typically reached in very few iterations. At the end of the iterative process, the matrices $\hat{\boldsymbol{P}}_{l^*}^{\boldsymbol{i}}$ ($i = 1, \ldots, I$) comprise the set of simultaneously updated unit matrices closest to $\hat{\boldsymbol{P}}_0$ that, verifying all the aggregate unit row and column constraints, prompt the compound matrix $\hat{\boldsymbol{P}}_{l^*}$ that minimizes the sum across units of the expected count discrepancies, in L_1-norm, between the observed and expected unit column totals when $\hat{\boldsymbol{P}}_{l^*}$ is applied to the observed unit row totals.

3.2 Symmetric models based on dual solutions

In the same way that equations (1)–(15) are defined, a similar linear system can be specified to obtain estimates of the proportion matrices $\boldsymbol{Q} = [q_{kj}]$ and $\boldsymbol{Q}^{\boldsymbol{i}} = [q_{kj}^{i}]$, for $i = 1, \ldots, I$. The problem, however, lies in the fact that these two dual representations of the same underlying ecological inference problem generate non-congruent solutions. The set of estimates $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{P}}^{\boldsymbol{i}}$ obtained using the x_{ij} votes as predictor variables and the y_{ik} votes as response variables is not congruent with the set of estimates $\widehat{\boldsymbol{Q}}$ and $\widehat{\boldsymbol{Q}}^{\boldsymbol{i}}$ that would be attained switching $\boldsymbol{X}$ and $\boldsymbol{Y}$. In general, the $\boldsymbol{P}$-based and $\boldsymbol{Q}$-based estimates for the number of votes of any generic (j, k) entry will be different.

That is, that $\left(\sum_{i=1}^{I} x_{ij}\right)\hat{p}_{jk} \neq \left(\sum_{i=1}^{I} y_{ik}\right)\hat{q}_{kj}$ and that $x_{ij}\hat{p}^i_{jk} \neq y_{ik}\hat{q}^i_{kj}$. Therefore, once the problem is solved both ways, $\boldsymbol{X} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{Y} \rightarrow \boldsymbol{X}$, we have to look for some (ad hoc) way of reconciliation if we want to exploit the information contained in both solutions.

We envisage two basic reconciliation approaches maintaining the accounting constraints that delimit the problem. On the one hand, just one of the two solutions could be selected. On the other hand, both solutions could be combined in a manner that preserves the constraints among votes and proportions. Below, we explore solutions for both types.

As an alternative to the classical approach of reasoning a logical (causal, temporal, or primacy) order between classifications, equation (15) could be employed to choose between solutions. Given the central role played by the heterogeneity index determining linear programming solutions and its previously reported close relationship with the accuracy of the (global) estimates (Pavía & Romero, 2022, 2023; Romero et al., 2020), we propose selecting between the sets $\{\hat{\boldsymbol{P}}, \hat{\boldsymbol{P}}^i\}$ and $\{\hat{\boldsymbol{Q}}, \boldsymbol{Q}^i\}$ the one with the smallest *HETe*. In the case of lphom solutions, the expression $\sum_{ik} |e^+_{ik}+e^-_{ik}| / \sum_{ij} x_{ij}$, suggested by Romero et al. (2020), could be used to approximate *HETe*. We identify these solutions with the suffix '_min'.

The above 'combination' of asymmetric solutions scarcely exploits the information contained in one of the two dual solutions, so we propose building combined solutions as averages, since averages compared to individual estimates tend to reduce expected errors. We consider either unweighted or weighted averages, of the two dual solutions, with the weighted solutions attained employing the inverses of the *HETe* statistics as weights. We denote with the suffix '_dual_a' the solutions associated with the sets of votes estimates $\{\overline{\boldsymbol{V}}_a,\ \overline{\boldsymbol{V}}^i_a\}$ given by equations (16) and (17) and with the suffix '_dual_w' the weighted solutions associated with the sets of votes estimates $\{\overline{\boldsymbol{V}}_w,\ \overline{\boldsymbol{V}}^i_w\}$ obtained using equations (18) and (19), where HET_p and HET_q represent the values of the *HETe* statistic of the respective solutions $\boldsymbol{X} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{Y} \rightarrow \boldsymbol{X}$. Note that for lphom_dual solutions, equations (17) and (19) do not apply.

$$\overline{\boldsymbol{V}}_a = \frac{\hat{\boldsymbol{V}} + \tilde{\boldsymbol{V}}}{2} = \left[\frac{\left(\sum_{i=1}^{I} x_{ij}\right)\hat{p}_{jk} + \left(\sum_{i=1}^{I} y_{ik}\right)\hat{q}_{kj}}{2}\right], \tag{16}$$

$$\overline{\boldsymbol{V}}^i_a = \frac{\hat{\boldsymbol{V}}^i + \tilde{\boldsymbol{V}}^i}{2} = \left[\frac{x_{ij}\hat{p}^i_{jk} + y_{ik}\hat{q}^i_{kj}}{2}\right], \tag{17}$$

$$\overline{\boldsymbol{V}}_w = \frac{\hat{\boldsymbol{V}} \cdot HET_p^{-1} + \tilde{\boldsymbol{V}} \cdot HET_q^{-1}}{HET_p^{-1} + HET_q^{-1}} = \left[\frac{HET_p^{-1}\left(\sum_{i=1}^{I} x_{ij}\right)\hat{p}_{jk} + HET_q^{-1}\left(\sum_{i=1}^{I} y_{ik}\right)\hat{q}_{kj}}{HET_p^{-1} + HET_q^{-1}}\right], \tag{18}$$

$$\overline{\boldsymbol{V}}^i_w = \frac{\hat{\boldsymbol{V}}^i \cdot HET_p^{-1} + \tilde{\boldsymbol{V}}^i \cdot HET_q^{-1}}{HET_p^{-1} + HET_q^{-1}} = \left[\frac{HET_p^{-1} x_{ij}\hat{p}^i_{jk} + HET_q^{-1} y_{ik}\hat{q}^i_{kj}}{HET_p^{-1} + HET_q^{-1}}\right]. \tag{19}$$

4. A symmetric linear model: lphom for simultaneous elections

The methods proposed in the previous section build ad hoc symmetric (congruent) solutions (i.e. solutions that do not depend on which classification is mapped to rows and which to columns) by selecting or averaging solutions from asymmetric methods. In this section, we modify the lphom model to directly produce global (optimal) symmetric solutions. In the next section, we extend the approach to also embrace the tslphom and nslphom models.

To make the notation more compact, we adopt a matrix representation. In particular, in matrix form, the linear programme, $\boldsymbol{X} \rightarrow \boldsymbol{Y}$, defined by equations (1)–(5) can be expressed by

$$\begin{aligned} &\text{minimize } \boldsymbol{f}_p^T \boldsymbol{z}_p, \\ &\text{subject to: } \boldsymbol{A}_p \boldsymbol{z}_p = \boldsymbol{b}_p, \\ &\qquad \boldsymbol{z}_p \geq 0, \end{aligned}$$

and, equally, the dual problem, $\boldsymbol{Y} \rightarrow \boldsymbol{X}$, can be expressed as

$$\begin{aligned} &\text{minimize } \boldsymbol{f}_q^T \boldsymbol{z}_q, \\ &\text{subject to: } \boldsymbol{A}_q \boldsymbol{z}_q = \boldsymbol{b}_p, \\ &\boldsymbol{z}_q \geq 0, \end{aligned}$$

where $\boldsymbol{f}_p^T = [0_{1\times JK},\ 1_{1\times 2IK}]$, $\boldsymbol{f}_q^T = [0_{1\times KJ}\ 1_{1\times 2IJ}]$, $\boldsymbol{A}_p$ is the $(J+K+IK)\times(JK+2IK)$ matrix given by equation (20), $\boldsymbol{A}_q$ is the $(J+K+IJ)\times(JK+2IJ)$ matrix stated in equation (21), and the rest of components are defined in (22); being $\boldsymbol{E}_p^+ = [{}_p e_{ik}^+]$ and $\boldsymbol{E}_p^- = [{}_p e_{ik}^-]$ and $\boldsymbol{E}_q^+ = [{}_q e_{ij}^+]$ and $\boldsymbol{E}_q^- = [{}_p e_{ij}^-]$, respectively, $I\times K$ and $I\times J$ matrices of auxiliary non-negative unknowns.

$$\boldsymbol{A}_p = \begin{bmatrix} \boldsymbol{I}_J \otimes \boldsymbol{1}_{1\times K} & 0_{J\times 2IK} \\ \boldsymbol{1}_{1\times I}\boldsymbol{X} \otimes \boldsymbol{I}_K & 0_{K\times 2IK} \\ \boldsymbol{X} \otimes \boldsymbol{I}_{IK} & [1,\ -1] \otimes \boldsymbol{I}_{IK} \end{bmatrix}, \tag{20}$$

$$\boldsymbol{A}_q = \begin{bmatrix} \boldsymbol{I}_K \otimes \boldsymbol{1}_{1\times J} & 0_{K\times 2IJ} \\ \boldsymbol{1}_{1\times I}\boldsymbol{Y} \otimes \boldsymbol{I}_J & 0_{J\times 2IJ} \\ \boldsymbol{Y} \otimes \boldsymbol{I}_{IJ} & [1,\ -1] \otimes \boldsymbol{I}_{IJ} \end{bmatrix}, \tag{21}$$

$$\boldsymbol{b}_p = \begin{bmatrix} \boldsymbol{1}_{\mathrm{J}\times 1} \\ \boldsymbol{Y}^T \boldsymbol{1}_{\mathrm{I}\times 1} \\ \mathrm{vec}(\boldsymbol{Y}) \end{bmatrix}, \quad \boldsymbol{b}_q = \begin{bmatrix} \boldsymbol{1}_{\mathrm{K}\times 1} \\ \boldsymbol{X}^T \boldsymbol{1}_{\mathrm{I}\times 1} \\ \mathrm{vec}(\boldsymbol{X}) \end{bmatrix}, \quad \boldsymbol{z}_p = \begin{bmatrix} \mathrm{vec}(\boldsymbol{P}) \\ \mathrm{vec}(\boldsymbol{E}_p^+) \\ \mathrm{vec}(\boldsymbol{E}_p^-) \end{bmatrix}, \quad \boldsymbol{z}_q = \begin{bmatrix} \mathrm{vec}(\boldsymbol{Q}) \\ \mathrm{vec}(\boldsymbol{E}_q^+) \\ \mathrm{vec}(\boldsymbol{E}_q^-) \end{bmatrix}. \tag{22}$$

Stacking both systems in the same programme and solving the resulting linear programme is not enough to guarantee congruency among the $\boldsymbol{P}$ and $\boldsymbol{Q}$ proportion estimates. To do this, restrictions detailed in equation (23) must be included as additional constraints. Equation (23) states that the global matrix of votes should be the same independent of what row-standardized proportion matrix is used to compute it, either $\boldsymbol{P}$ or $\boldsymbol{Q}$.

$$\left(\sum_{i=1}^{I} x_{ij}\right) p_{jk} = \left(\sum_{i=1}^{I} y_{ik}\right) q_{kj} \quad \text{for } j = 1,\ \ldots,\ J,\ \ k = 1,\ \ldots,\ K. \tag{23}$$

Obviously, these constraints can also be expressed in matrix form. In particular, $\begin{bmatrix} \boldsymbol{B}_p & \boldsymbol{B}_q \end{bmatrix}[\mathrm{vec}(\boldsymbol{P})^T,\ \mathrm{vec}(\boldsymbol{Q})^T]^T = 0_{JK\times 1}$, where $\boldsymbol{B}_p = \mathrm{diag}(\boldsymbol{1}_{1\times I}\boldsymbol{X}) \otimes \boldsymbol{I}_K$ and $\boldsymbol{B}_q = \sum_{k=1}^{K} \mathrm{row}_k(\boldsymbol{I}_k)^T \otimes (-\boldsymbol{I}_J \otimes \mathrm{row}_k(\boldsymbol{I}_k)(\boldsymbol{1}_{1\times K}\boldsymbol{Y})_k)$, with $(\boldsymbol{1}_{1\times K}\boldsymbol{Y})_k$ being the kth-component of the vector $\boldsymbol{1}_{1\times K}\boldsymbol{Y}$.

Combining all the equations, a new linear programme system emerges. The programme given by (24) describes the lphom_joint model, which corresponds to the symmetric version of the lphom model.

$$\begin{aligned} &\text{minimize } (\boldsymbol{f}_p^T \boldsymbol{z}_p + \boldsymbol{f}_q^T \boldsymbol{z}_q), \\ &\text{subject to:} \\ &\begin{bmatrix} \boldsymbol{A}_p & 0_{(J+K+IK)\times(JK+2IJ)} \\ 0_{(J+K+IJ)\times(JK+2IK)} & \boldsymbol{A}_q \\ \boldsymbol{B}_p\ 0_{JK\times(JK+2IK)} & \boldsymbol{B}_q\ 0_{JK\times(JK+2IJ)} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_p \\ \boldsymbol{z}_q \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_p \\ \boldsymbol{b}_q \\ 0_{JK\times 1} \end{bmatrix}, \\ &[\boldsymbol{z}_p^T,\ \boldsymbol{z}_q^T\ ]^T \geq 0. \end{aligned} \tag{24}$$

In this model, the unknowns p_{jk} and q_{kj} are jointly estimated using a single linear programme that includes the constraints of a lphom model in which the unknowns are the p_{jk} and those of another lphom model in which the unknowns are the q_{kj} as well as JK additional restrictions to guarantee congruence (i.e. that the probabilistic symmetric condition is fulfilled) between the p_{jk} and the q_{kj}

estimates for each (j, k) pair. These last additional constraints force the v_{jk} estimates from both sets of unknowns to coincide. That is, referring to solutions of both models in terms of matrices, $\boldsymbol{V}$, of votes, not in terms of matrices of proportions, it can be seen that lphom_joint verifies $\text{lphom\_joint}(\boldsymbol{X}, \boldsymbol{Y}) = \text{lphom\_joint}(\boldsymbol{Y}, \boldsymbol{X})^T$, in contrast to lphom, where $\text{lphom}(\boldsymbol{X}, \boldsymbol{Y}) \neq \text{lphom}(\boldsymbol{Y}, \boldsymbol{X})^T$.

The estimate $\hat{\boldsymbol{V}}_0$ of (24) is the (only) matrix that in addition to verifying all the aggregate/global row and column constraints minimizes, in L_1-norm, the sum across units of the accumulation of (i) the expected count discrepancies between the observed and expected unit column totals when $\text{diag}(\hat{\boldsymbol{V}}_0 1_{K\times 1})^{-1}\hat{\boldsymbol{V}}_0$ is applied to the observed unit row totals and (ii) the expected count discrepancies between the observed and expected unit row totals when $\text{diag}(1_{1\times J}\hat{\boldsymbol{V}}_0)^{-1}\hat{\boldsymbol{V}}_0^T$ is applied to the observed unit column totals.

5. Estimating local transfer matrices: nslphom for simultaneous elections

Although the lphom_joint model solves the inconsistency of the lphom procedure by generating estimates of matrices of votes that do not depend on how classifications are mapped to row and columns, it still reveals a crucial limitation of lphom. lphom_joint only estimates global (aggregate) matrices, yet having estimates of local (unit) matrices ($\boldsymbol{P}^i$, $\boldsymbol{Q}^i$, or $\boldsymbol{V}^i$) is also useful. Having local estimates is not only of interest in itself in many applications, but it also helps to obtain better aggregate estimates (King, 1997). The process of obtaining local estimates calls for a more intensive exploitation of all the unit observations, an approach that, as a rule and regardless of the methodological framework used, leads to achieving more accurate global solutions. Indeed, the tslphom and nslphom algorithms have also been revealed as significantly more accurate than the lphom procedure for a very wide range of scenarios (Pavía & Romero, 2023).

Continuing on from the previous sections, this section amends the tslphom and nslphom algorithms by proposing new versions of the procedures that, through their specification, force consistency between the estimates of all the $(\boldsymbol{P}^i, \boldsymbol{Q}^i)$ pairs. We anticipate that the new models (which we call tslphom_joint and nslphom_joint), in addition to generating local estimates, will also produce more accurate global solutions.

Both tslphom and nslphom are based on the lphom_local procedure (Pavía & Romero, 2022), a procedure devised to estimate unit matrices that is at the core of both algorithms. Indeed, their equations, starting from an arbitrary initial global transfer matrix, are for a fixed i similar to equations (6)–(13). Therefore, we first adapt, in Section 5.1, the lphom_local algorithm to build the lphom_local_joint method. Subsequently, in Section 5.2, we build the _joint versions of tslphom and nslphom on this.

5.1 The lphom_local_joint algorithm

The lphom_local algorithm for simultaneous elections can be observed as a two-step linear programme procedure that takes as inputs a row-standardized proportion matrix $\boldsymbol{T}$ (or $\mathbf{U}$) of order $J \times K$ (or $K \times J$) and two non-negative vectors $\boldsymbol{r} = [r_1, r_2, \ldots, r_J]$ and $\boldsymbol{s} = [s_1, s_2, \ldots, s_K]$, verifying $\sum_j r_j = \sum_k s_k$, and delivers as output a new row-standardized proportion matrix, $\hat{\boldsymbol{T}}$ (or $\hat{\boldsymbol{U}}$), as close as possible to $\boldsymbol{T}$ (or $\boldsymbol{U}$), verifying $\boldsymbol{r}\hat{\boldsymbol{T}} = \boldsymbol{s}$ (or $\boldsymbol{s}\hat{\boldsymbol{U}} = \boldsymbol{r}$). The second linear programme is required because the solution of the first programme is often indeterminate.

For a fixed unit i, the first two dual linear programmes, which correspond to equations (6)–(10), can be expressed with our notation as

$$\left.\begin{array}{l}\min\ \omega_p^i = \boldsymbol{g}^T\boldsymbol{z}_p^i \\ \text{s.t.}\quad \boldsymbol{A}_p^i\boldsymbol{z}_p^i = \boldsymbol{b}_p^i \\ \qquad \boldsymbol{z}_p^i \geq 0\end{array}\right\} \qquad \left.\begin{array}{l}\min\ \omega_q^i = \boldsymbol{g}^T\boldsymbol{z}_q^i \\ \text{s.t.}\quad \boldsymbol{A}_q^i\boldsymbol{z}_q^i = \boldsymbol{b}_q^i \\ \qquad \boldsymbol{z}_q^i \geq 0\end{array}\right\},$$

where $\boldsymbol{A}_p^i$ and $\boldsymbol{A}_q^i$ are the $(J + K + JK) \times 3JK$ matrices given by (25), and the rest of the components are defined in (26) and (27); being $\boldsymbol{P}_G$ and $\boldsymbol{Q}_G$, respectively, a parties-to-candidates and a candidates-to-parties proportion transfer matrix, $\boldsymbol{x}_i = row_i(\boldsymbol{X})$ and $\boldsymbol{y}_i = row_i(\boldsymbol{Y})$ the observed column-vectors of votes to parties and candidates, and $\boldsymbol{E}_p^{i+} = [{}_p\varepsilon_{jk}^{i+}]$ and $\boldsymbol{E}_p^{i-} = [{}_p\varepsilon_{jk}^{i-}]$ and

$\boldsymbol{E}_q^+=[{}_q\varepsilon_{kj}^{i+}]$ and $\boldsymbol{E}_q^-=[{}_q\varepsilon_{kj}^{i-}]$, respectively, $J\times K$ and $K\times J$ matrices of auxiliary non-negative unknowns.

$$\boldsymbol{A}_p^i=\begin{bmatrix}\boldsymbol{I}_J\otimes\mathbf{1}_{1\times K} & \mathbf{0}_{J\times 2JK}\\ (\boldsymbol{x}_i\otimes\boldsymbol{I}_K)^T & \mathbf{0}_{K\times 2JK}\\ \mathrm{diag}(\boldsymbol{x}_i)\otimes\boldsymbol{I}_K & [1,\ -1]\otimes\boldsymbol{I}_{JK}\end{bmatrix},\qquad \boldsymbol{A}_q^i=\begin{bmatrix}\boldsymbol{I}_K\otimes\mathbf{1}_{1\times J} & \mathbf{0}_{K\times 2KJ}\\ (\boldsymbol{y}_i\otimes\boldsymbol{I}_J)^T & \mathbf{0}_{J\times 2KJ}\\ \mathrm{diag}(\boldsymbol{y}_i)\otimes\boldsymbol{I}_J & [1,\ -1]\otimes\boldsymbol{I}_{KJ}\end{bmatrix}, \tag{25}$$

$$\boldsymbol{b}_p^i=\begin{bmatrix}\mathbf{1}_{\mathrm{J}\times 1}\\ \boldsymbol{y}_i\\ \mathrm{vec}(\boldsymbol{P}_G\odot(\boldsymbol{x}_i\otimes\mathbf{1}_{1\times K}))\end{bmatrix},\qquad \boldsymbol{b}_q^i=\begin{bmatrix}\mathbf{1}_{\mathrm{K}\times 1}\\ \boldsymbol{x}_i\\ \mathrm{vec}(\boldsymbol{Q}_G\odot(\boldsymbol{y}_i\otimes\mathbf{1}_{1\times J}))\end{bmatrix}, \tag{26}$$

$$\boldsymbol{z}_p^i=\begin{bmatrix}\mathrm{vec}(\boldsymbol{P}^i)\\ \mathrm{vec}(\boldsymbol{E}_p^{i+})\\ \mathrm{vec}(\boldsymbol{E}_p^{i-})\end{bmatrix},\qquad \boldsymbol{z}_q^i=\begin{bmatrix}\mathrm{vec}(\boldsymbol{Q}^i)\\ \mathrm{vec}(\boldsymbol{E}_q^{i+})\\ \mathrm{vec}(\boldsymbol{E}_q^{i-})\end{bmatrix},\qquad \boldsymbol{g}=[\mathbf{0}_{JK\times 1},\ \mathbf{1}_{2JK\times 1}]. \tag{27}$$

As with lphom, independently solving the two dual problems does not assure the congruency between both dual solutions, even starting from a congruent $(\boldsymbol{P}_G,\ \boldsymbol{Q}_G)$ pair of proportion matrices. To guarantee congruency, we need to also include the constraints given by (28), which states that the value for the votes v_{jk}^i must be the same independent of whether we compute them following a parties-to-candidates route or the reverse, candidates-to-parties. Equation (28) can be expressed in matrix form as $\begin{bmatrix}\boldsymbol{B}_p^i & \boldsymbol{B}_q^i\end{bmatrix}[\mathrm{vec}(\boldsymbol{P}_G)^T,\ \mathrm{vec}(\boldsymbol{Q}_G)^T]^T=\mathbf{0}_{JK\times 1}$, where $\boldsymbol{B}_p^i=\mathrm{diag}(\boldsymbol{x}_i)\otimes\boldsymbol{I}_K$ and $\boldsymbol{B}_q^i=\sum_{k=1}^K\mathrm{row}_k(\boldsymbol{I}_k)^T\otimes(-\boldsymbol{I}_J\otimes\mathrm{row}_k(\boldsymbol{I}_k)y_{ik})$.

$$x_{ij}\ p_{jk}^i=y_{ik}\ q_{kj}^i\quad\text{for } j=1,\ \ldots,J,\ \ k=1,\ \ldots,K. \tag{28}$$

Similarly, the second two dual linear programmes, which correspond to equations (11)–(13), can be stated as

$$\left.\begin{array}{c}\min \boldsymbol{h}^T\boldsymbol{w}_p^i\\ \text{s.t.}\quad \boldsymbol{C}_p^i\boldsymbol{w}_p^i=\boldsymbol{d}_p^i\\ \boldsymbol{w}_p^i\geq 0\end{array}\right\}\qquad\left.\begin{array}{c}\min \boldsymbol{h}^T\boldsymbol{w}_q^i\\ \text{s.t.}\quad \boldsymbol{C}_q^i\boldsymbol{w}_q^i=\boldsymbol{d}_q^i\\ \boldsymbol{w}_q^i\geq 0\end{array}\right\},$$

where $\boldsymbol{h}^T=[\mathbf{0}_{1\times 3JK},\ \mathbf{1}_{1\times 2JK}]$, $\boldsymbol{C}_p^i$ and $\boldsymbol{C}_q^i$ are the $(J+K+1+2JK)\times 5JK$ matrices defined in equation (29), and the rest of the components are defined in (30); being $\boldsymbol{\Delta}_p^{i+}=[{}_p\delta_{jk}^{i+}]$ and $\boldsymbol{\Delta}_p^{i-}=[{}_p\delta_{jk}^{i-}]$ and $\boldsymbol{\Delta}_q^+=[{}_q\delta_{kj}^{i+}]$ and $\boldsymbol{\Delta}_q^-=[{}_q\delta_{kj}^{i-}]$, respectively, $J\times K$ and $K\times J$ matrices of auxiliary non-negative unknowns.

$$\boldsymbol{C}_p^i=\begin{bmatrix}\boldsymbol{A}_p^i & \mathbf{0}_{(\mathrm{J+K+JK})\times 2JK}\\ \boldsymbol{g}^T & \mathbf{0}_{1\times 2JK}\\ \boldsymbol{I}_{JK} & \mathbf{0}_{\mathrm{JK}\times 2JK}\ \ [1,\ -1]\otimes\boldsymbol{I}_{JK}\end{bmatrix},\qquad \boldsymbol{C}_q^i=\begin{bmatrix}\boldsymbol{A}_q^i & \mathbf{0}_{(\mathrm{J+K+JK})\times 2JK}\\ \boldsymbol{g}^T & \mathbf{0}_{1\times 2JK}\\ \boldsymbol{I}_{JK} & \mathbf{0}_{\mathrm{JK}\times 2JK}\ \ [1,\ -1]\otimes\boldsymbol{I}_{JK}\end{bmatrix}, \tag{29}$$

$$\boldsymbol{d}_p^i=\begin{bmatrix}\boldsymbol{b}_p^i\\ \omega_p^i\\ \mathrm{vec}(\boldsymbol{P}_G)\end{bmatrix},\quad \boldsymbol{d}_q^i=\begin{bmatrix}\boldsymbol{b}_q^i\\ \omega_q^i\\ \mathrm{vec}(\boldsymbol{Q}_G)\end{bmatrix},\quad \boldsymbol{w}_p^i=\begin{bmatrix}\boldsymbol{z}_p^i\\ \mathrm{vec}(\boldsymbol{\Delta}_p^{i+})\\ \mathrm{vec}(\boldsymbol{\Delta}_p^{i-})\end{bmatrix},\quad \boldsymbol{w}_q^i=\begin{bmatrix}\boldsymbol{z}_q^i\\ \mathrm{vec}(\boldsymbol{\Delta}_q^{i+})\\ \mathrm{vec}(\boldsymbol{\Delta}_q^{i-})\end{bmatrix}. \tag{30}$$

Putting together all the pieces, we propose as lphom_local_joint algorithm the two-step linear programme procedure that yields as outcome the result of sequentially solving the programmes defined by the systems (31) and (32).

$$\begin{gathered}\text{minimize } \omega^i = (\boldsymbol{g}^T\boldsymbol{z}_p^i + \boldsymbol{g}^T\boldsymbol{z}_q^i),\\ \text{subject to:}\\ \begin{bmatrix} \boldsymbol{A}_p^i & \boldsymbol{0}_{(J+K+JK)\times 3JK} \\ \boldsymbol{0}_{(J+K+JK)\times 3JK} & \boldsymbol{A}_q^i \\ \boldsymbol{B}_p^i \ \boldsymbol{0}_{JK\times 2JK} & \boldsymbol{B}_q^i \ \boldsymbol{0}_{JK\times 2JK} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_p^i \\ \boldsymbol{z}_q^i \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_p^i \\ \boldsymbol{b}_q^i \\ \boldsymbol{0}_{JK\times 1} \end{bmatrix},\\ [\boldsymbol{z}_p^{iT}\ ,\ \boldsymbol{z}_p^{iT}\ ]^T \geq \boldsymbol{0}. \end{gathered} \tag{31}$$

Once the minimum value for ω^i is achieved solving the system (31), which characterizes the set of potential congruent solutions also verifying the two first dual local linear systems, the lphom_local_joint solution is attained by solving the extended system given by equation (32).

$$\begin{gathered}\text{minimize } [\boldsymbol{0}_{1\times 6JK}, \boldsymbol{1}_{1\times 4JK}\ ]\boldsymbol{w}^i,\\ \text{subject to:}\\ \begin{bmatrix} \boldsymbol{A}_p^i & \boldsymbol{0}_{(J+K+JK)\times 3JK} & \\ \boldsymbol{0}_{(J+K+JK)\times 3JK} & \boldsymbol{A}_q^i & \boldsymbol{0}_{(2(J+K)+3JK)\times 4JK} \\ \boldsymbol{B}_p^i \ \boldsymbol{0}_{JK\times 2JK} & \boldsymbol{B}_q^i \ \boldsymbol{0}_{JK\times 2JK} & \\ \boldsymbol{g}^T & \boldsymbol{g}^T & \boldsymbol{0}_{1\times 4JK} \\ \boldsymbol{I}_{JK}\ \boldsymbol{0}_{JK\times 2JK} & \boldsymbol{0}_{JK\times 3JK} & [1,\ -1]\otimes \boldsymbol{I}_{JK}\ \ \boldsymbol{0}_{JK\times 2JK} \\ \boldsymbol{0}_{JK\times 3JK}\ \ \boldsymbol{I}_{JK} & \boldsymbol{0}_{JK\times 2JK} & \boldsymbol{0}_{JK\times 2JK}\ \ [1,\ -1]\otimes \boldsymbol{I}_{JK} \end{bmatrix} \boldsymbol{w}^i = \begin{bmatrix} \boldsymbol{b}_p^i \\ \boldsymbol{b}_q^i \\ \boldsymbol{0}_{JK\times 1} \\ \omega^i \\ vec(\boldsymbol{P}_G) \\ vec(\boldsymbol{Q}_G) \end{bmatrix},\\ \boldsymbol{w}^i \geq \boldsymbol{0}. \end{gathered} \tag{32}$$

where $\boldsymbol{w}^i = [\boldsymbol{z}_p^{iT}\ ,\ \boldsymbol{z}_p^{iT}, \text{vec}(\boldsymbol{\Delta}_p^{i+})^T, \text{vec}(\boldsymbol{\Delta}_p^{i-})^T, \text{vec}(\boldsymbol{\Delta}_q^{i+})^T, \text{vec}(\boldsymbol{\Delta}_q^{i-})^T\ ]^T$.

This model generates solutions of the estimated unit level proportions that, verifying all the ecological accounting equalities, including constraints (28), equivalent to the probabilistic symmetric condition, are closest (in L_1-norm) to the $\boldsymbol{P}_G$ and $\boldsymbol{Q}_G$ matrices.

As with the lphom_joint model, the lphom_local_joint algorithm also verifies the symmetry property. That is, given a matrix of counts $\boldsymbol{V}$ of order $J\times K$ and a fixed unit i, the equality lphom_local_joint$(\boldsymbol{P}_G,\ \boldsymbol{x}_i, \boldsymbol{y}_i)$ = lphom_local_joint$(\boldsymbol{Q}_G,\ \boldsymbol{y}_i, \boldsymbol{x}_i)^T$ holds, whereas the relationship lphom_local$(\boldsymbol{P}_G,\ \boldsymbol{x}_i, \boldsymbol{y}_i) \neq$ lphom_local$(\boldsymbol{Q}_G,\ \boldsymbol{y}_i, \boldsymbol{x}_i)^T$ is not verified (in general); where $\boldsymbol{P}_G = \text{diag}(\boldsymbol{V}\boldsymbol{1}_{K\times 1})^{-1}\boldsymbol{V}$ and $\boldsymbol{Q}_G = \text{diag}(\boldsymbol{1}_{1\times J}\boldsymbol{V})^{-1}\boldsymbol{V}^T$ and the outputs of the algorithms refer to local matrices of votes, not of proportions.

5.2 The nslphom_joint model

Once lphom and lphom_local are adapted to make them symmetric, nslphom can easily be modified in the same way; that is, we can build the nslphom_joint model. In particular, for a fixed *ns*, the natural definition of the nslphom_joint model is given by the following iterative procedure:

- Iteration 0. Apply lphom_joint to $\boldsymbol{X}$ and $\boldsymbol{Y}$ to obtain initial congruent estimates $\ddot{\boldsymbol{P}}_0$ and $\ddot{\boldsymbol{Q}}_0$ of the global proportion matrices, from which the aggregate matrix of votes can be derived through: $\ddot{\boldsymbol{V}}_0 = \ddot{\boldsymbol{P}}_0 \odot [(\boldsymbol{1}_{1\times I}\boldsymbol{X})^T \otimes \boldsymbol{1}_{1\times K}] = [\ddot{\boldsymbol{Q}}_0 \odot [(\boldsymbol{1}_{1\times I}\boldsymbol{Y})^T \otimes \boldsymbol{1}_{1\times J}]]^T$.
- Iteration l, for $l = 1, \ldots, ns$. Compute $\ddot{\boldsymbol{P}}_{l-1} = \text{diag}(\ddot{\boldsymbol{V}}_{l-1}\boldsymbol{1}_{K\times 1})^{-1}\ddot{\boldsymbol{V}}_{l-1}$ and $\ddot{\boldsymbol{Q}}_{l-1} = \text{diag}(\boldsymbol{1}_{1\times J}\ddot{\boldsymbol{V}}_{l-1})^{-1}\ddot{\boldsymbol{V}}_{l-1}^T$ and apply lphom_joint_local using as inputs $\ddot{\boldsymbol{P}}_{l-1}$, $\ddot{\boldsymbol{Q}}_{l-1}$, $\boldsymbol{x}_i = \text{row}_i(\boldsymbol{X})$, and $\boldsymbol{y}_i = \text{row}_i(\boldsymbol{Y})$, for $i = 1, \ldots, I$. This produces congruent estimates of proportion matrices $\ddot{\boldsymbol{P}}_l^i$ and $\ddot{\boldsymbol{Q}}_l^i$ for each unit i; from which an estimate of the global matrix of votes is built by aggregating the estimates of the corresponding unit matrices: $\ddot{\boldsymbol{V}}_l = \sum_{i=1}^I \ddot{\boldsymbol{V}}_l^i$; where $\ddot{\boldsymbol{V}}_l^i = \ddot{\boldsymbol{P}}_l^i \odot (\boldsymbol{x}_i \otimes \boldsymbol{1}_{1\times K}) = [\ddot{\boldsymbol{Q}}_l^i \odot (\boldsymbol{y}_i \otimes \boldsymbol{1}_{1\times J})]^T$.

During each iteration, the statistic given by equation (33), where ${}_l\ddot{v}^i_{jk}$, ${}_l\ddot{p}_{jk}$, and ${}_l\ddot{q}_{kj}$ are the generic (j, k) entries of the matrices $\ddot{\boldsymbol{V}}^i_l$, $\ddot{\boldsymbol{P}}_l$, and $\ddot{\boldsymbol{Q}}_l$, is also calculated.

$$HET\ddot{e}_l = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}|{}_l\ddot{v}^i_{jk} - {}_l\ddot{p}_{jk}\,x_{ij}| + \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J}|{}_l\ddot{v}^i_{kj} - {}_l\ddot{q}_{kj}\,y_{ik}|}{4\sum_{i=1}^{I}\mathbf{1}_{1\times J}\ddot{\boldsymbol{V}}_l\mathbf{1}_{K\times 1}}. \tag{33}$$

In a similar vein to nslphom, the lphom_joint, tslphom_joint and nlsphom_joint solutions are also attained once the above iterative process is finished. The set of matrices $\{\ddot{\boldsymbol{P}}_0,\ \ddot{\boldsymbol{Q}}_0, \ddot{\boldsymbol{V}}_0\}$ corresponds to the lphom_joint solution. The sets of matrices $\{\ddot{\boldsymbol{P}}_1,\ \ddot{\boldsymbol{Q}}_1, \ddot{\boldsymbol{V}}_1\}$ and $\{\ddot{\boldsymbol{P}}^i_1, \ddot{\boldsymbol{Q}}^i_1, \ddot{\boldsymbol{V}}^i_1\}_{i=1}^I$ achieved at the end of iteration 1 encapsulate the tslphom_joint solution. Finally, the nslphom_joint solution is made up of the sets of matrices $\{\ddot{\boldsymbol{P}}_{l^*},\ \ddot{\boldsymbol{Q}}_{l^*}, \ddot{\boldsymbol{V}}_{l^*}\}$ and $\{\ddot{\boldsymbol{P}}^i_{l^*}, \ddot{\boldsymbol{Q}}^i_{l^*}, \ddot{\boldsymbol{V}}^i_{l^*}\}_{i=1}^I$, with l^* corresponding to the iteration for which (33) reaches its minimum in the *ns* iterations.

At the end of the iterative process, the estimates $\ddot{\boldsymbol{P}}^i_{l^*}, \ddot{\boldsymbol{Q}}^i_{l^*}, \ddot{\boldsymbol{V}}^i_{l^*}$ $(i = 1, \ldots, I)$ comprise the set of simultaneously updated unit matrices closest to $\ddot{\boldsymbol{V}}_0$ that, verifying all the aggregate unit row and column constraints, generate the compound matrix $\ddot{\boldsymbol{V}}_{l^*}$ that minimizes, in L_1-norm, the sum across units of the accumulation of (i) the expected count discrepancies between the observed and expected unit column totals when $\ddot{\boldsymbol{P}}^i_{l^*}$ is applied to the observed unit row totals and (ii) the expected count discrepancies between the observed and expected unit row totals when $\ddot{\boldsymbol{Q}}^i_{l^*}$ is applied to the observed unit column totals.

All the new algorithms introduced in Sections 3–5 are available in functions with suffixes '_dual' and '_joint' in the R package lphom.

6. Assessment of procedures

In Sections 3–5, two new sets of symmetric algorithms based on linear programming have been suggested to solve the ecological inference problem. These procedures represent an alternative to the asymmetric methods, where rows and columns of the ecological matrices are not interchangeable. In this section, we assess, focused on accuracy, the performance of the proposed procedures. In addition to comparing the different solutions they provide, we also assess them in comparison to the equivalent asymmetric solutions, which act as baseline solutions.

Although it is possible to also include in the assessments other asymmetric models, we have restricted the comparisons to the asymmetric algorithms from which the symmetric methods stem. Under our view, this makes comparisons fairer and does not entail a limitation, as lphom-family algorithms are among the most accurate methods to estimate vote transition matrices. First, Klima et al (2016) state the ei.MD.bayes algorithm (Lau et al., 2020; Rosen et al., 2001) as the most accurate after comparing it to (i) a modification of classical ecological regression (Goodman, 1953), (ii) Thomsen's approach (Thomsen, 1987), and (iii) two recursive 2×2-approaches—the ones proposed by Andreadis and Chadjipadelis (2009) and by Kellermann (2011). Second, Plescia and De Sio (2018) also identify ei.MD.bayes as the best option after comparing it to (i) the RxCEcolInf method (Greiner et al., 2021; Greiner & Quinn, 2009) and (ii) classical ecological regression. Third, Barreto et al. (2022) conclude that King's iterative EI:RxC (Collingwood et al., 2020; King, 1997) and ei.MD.bayes can be used interchangeably when assessing precinct-level voting patterns in Voting Rights Act cases, with Ferree (2004) and Katz (2014) considering iterative EI:RxC inferior. Finally, Pavía and Romero (2023) extensively compare the lphom-family algorithms with ei.MD.bayes-solutions and find the former to be as least as accurate than ei.MD.bayes, being moreover more accurate when the information is scarce or convergence cannot be guarantee for ei.MD.bayes. In summary, if the new symmetric methods improve the asymmetric lphom-based solutions, by the transitivity property they should also improve the rest of the asymmetric methods.

6.1 Data

Due to the very nature of the problem, datasets with known true vote transfer matrices are scarce, so practical research typically relies on artificial, simulated data. In this paper, however, we gauge the different approaches using real data. The results of this section are based on comparing

estimated and true matrices of votes belonging to more than 500 elections available in the R package ei.Datasets (Pavía, 2022). The ei.Datasets package contains the matrices of party votes, $\boldsymbol{X}$, candidate votes, $\boldsymbol{Y}$, and party and candidate cross-distribution tables of votes at district/election-level, $\boldsymbol{V}$, corresponding to 565 mixed-member elections held between 2002 and 2020 in New Zealand (NZ) and in 2007 in Scotland (SCO). True cross-classification of votes at polling station (unit) level is not available. At unit level only the marginal distributions of votes (matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$) are known. New Zealand employs a mixed-member proportional system for electing parliament members, while Scotland uses an additional member voting system. Both systems involve voters casting two votes—one for a regional or national party list and another for a local candidate (who need not come from the chosen party). The candidate with the most votes in each district is automatically elected, and the remaining seats are allocated based on proportional party votes. In New Zealand, national compensatory seats ensure that each party's share of seats aligns closely with its share of votes across the country.

Although, as far as we know, this is the first time in the literature that all the datasets included in ei.Datasets are employed for ecological inference assessment, some of these elections have already been utilized for evaluating ecological inference procedures (Pavía & Romero, 2022, 2023; Plescia & De Sio, 2018). In using these data, we follow in the footsteps of the above authors and, as is usual practice when handling real data (e.g. Barreto et al. 2022; Klein, 2019; Klima et al., 2016), we have only considered sizeable populations and grouped very small election options in 'Others'. We have merged in 'Other parties' and 'Other candidates' those parties and candidates, respectively, who individually do not attain at least a 3% of the district vote. Despite this operation notably reducing the sizes of the ecological tables (from 147.6 to 28.2 in average number of cells), we still retain a large diversity of size tables, with tables of 23 different sizes ranging from a minimum of 15 cells (5×3) to a maximum of 56 cells (8×7); 8 and 3 being the maximum and minimum number for both rows and columns. Before merging, ei.Datasets tables range from a minimum of 51 cells (17×3) to a maximum of 300 cells (20×15), 27 being the maximum number of rows for a single table and 15 the maximum number of columns. In terms of sizes of the districts, measured in number of units (polling stations), we have districts ranging from a minimum of 22 to a maximum of 833, with an average of 90.5 units per district. More details about the datasets can be found in Pavía (2022).

6.2 Error measure

Regarding the metric utilized to assess the estimated matrices of votes, $\check{\boldsymbol{V}} = [\check{v}_{jk}]$, we rely on the error index, *EI*. This statistic, formulated in equation (34), accounts for the percentage of votes incorrectly allocated in the estimated matrix. This is the minimum number of votes that should be relocated among entries of the matrix to achieve a perfect fit. Although this statistic had been defined by other authors omitting the 0.5 coefficient (e.g. Klima et al., 2016), we prefer to follow the specification proposed by Romero et al. (2020) because it avoids counting every wrongly assigned vote twice.

$$EI = 100 \cdot \frac{0.5 \sum_{j=1}^{J} \sum_{k=1}^{K} |v_{jk} - \check{v}_{jk}|}{\sum_{j=1}^{J} \sum_{k=1}^{K} v_{jk}}. \tag{34}$$

6.3 Results

Tables 1 and 2 offer a summary of the accuracy of the different algorithms. The summaries are presented after grouping the elections following two different categorizations. In Table 1, the elections have been grouped by country and by year in which they were held. In our view, this is the most natural way of grouping these elections since all the datasets belonging to the same year and country share the same political environment. They all correspond to district elections held in the context of the same national general election. Nevertheless, according to Pavía (2022), another natural way of grouping the NZ elections is by type of district, either Māori or regular.

Table 2 presents the summaries grouping the NZ elections as alternatively suggested by Pavía (2022), but focusing only on the most accurate algorithms: those based on the nslphom model.

Table 1. Averages of EI errors by group of elections

Country year	NZ 2002	NZ 2005	SCO 2007	NZ 2008	NZ 2011	NZ 2014	NZ 2017	NZ 2020
# of elections	$N = 69$	$N = 69$	$N = 73$	$N = 70$	$N = 70$	$N = 71$	$N = 71$	$N = 72$
Avg. # of units	$\bar{I} = 83.2$	$\bar{I} = 81.8$	$\bar{I} = 70.2$	$\bar{I} = 84.1$	$\bar{I} = 85.7$	$\bar{I} = 81.2$	$\bar{I} = 101.9$	$\bar{I} = 134.9$
Avg. # of cells	$\overline{JK} = 39.5$	$\overline{JK} = 23.8$	$\overline{JK} = 35.2$	$\overline{JK} = 23.4$	$\overline{JK} = 26.2$	$\overline{JK} = 27.9$	$\overline{JK} = 24.8$	$\overline{JK} = 24.5$
lphom-based solutions								
lphom(X, Y)	16.88	12.14	12.92	12.22	12.99	12.95	12.20	14.02
lphom(Y, X)	16.14	12.74	10.78	11.87	13.40	13.70	12.08	12.18
lphom_min	16.30	11.86	11.73	11.58	12.65	12.80	11.61	12.59
lphom_dual_a	14.87	11.47	10.50	11.03	12.14	12.03	11.19	12.16
lphom_dual_w	14.89	11.37	10.41	10.97	12.05	12.01	11.10	12.11
lphom_joint	15.36	11.65	10.52	11.31	12.51	12.35	11.48	12.60
tslphom-based solutions								
tslphom(X, Y)	14.80	10.91	11.00	10.89	11.50	11.66	10.91	12.59
tslphom(Y, X)	14.52	11.50	9.46	10.77	12.22	12.50	10.95	11.07
tslphom_min	14.06	10.65	9.64	10.30	11.20	11.61	10.36	11.18
tslphom_dual_a	13.18	10.27	8.96	9.87	10.83	10.83	10.03	10.92
tslphom_dual_w	13.15	10.17	8.87	9.79	10.74	10.78	9.95	10.85
tslphom_joint	14.07	10.73	9.42	10.40	11.46	11.45	10.49	11.54
nslphom-based solutions								
nslphom(X, Y)	12.79	9.47	8.85	9.11	9.46	9.69	8.91	9.82
nslphom(Y, X)	12.71	9.89	8.11	9.17	10.86	11.15	9.29	9.39
nslphom_min	12.55	9.22	8.42	8.63	9.41	9.67	8.48	9.07
nslphom_dual_a	**11.10**	8.63	**7.11**	8.08	9.01	9.08	8.04	8.52
nslphom_dual_w	11.14	**8.58**	**7.11**	**8.04**	**8.91**	**9.01**	**7.96**	**8.48**
nslphom_joint	11.54	8.86	7.53	8.44	9.30	9.39	8.23	8.99

Source: Compiled by the authors after applying, with default options, the functions lphom, tslphom, and nslphom of the R package lphom (Pavía & Romero, 2022) and the new algorithms, described in Sections 3–5, to the data from the New Zealand electoral commission and the Scotland Electoral Office available in the R package ei.Datasets (Pavía, 2022). In the case of nslphom-based models, *ns* is always fixed at 10. The smaller the number, the better the accuracy. Values in bold correspond to the most average accurate solutions in each set of elections.

To offer more context and make comparisons easier, we also include in Table 2 the results attained pooling all the elections and combining all the NZ elections in a unique group and, again, the Scottish results. The upper panels of both tables contain some summary statistics of the corresponding group of elections. Here, one of the differential characteristics of the Māori districts stands out: the relative higher number of units in which their electorates are split out. Figure 2 shows graphically the average errors attained by the nslphom-family algorithms in all the groupings.

In light of the empirical assessments, two clear findings emerge. On the one hand, our results confirm the order of preference for the symmetric models already stated by Pavía and Romero (2022, 2023) for the asymmetric procedures. As with asymmetric models, tslphom-based methods systematically improve lphom-based ones and, equally, nslphom-based models consistently outperform tslphom-based ones (see Table 1). More importantly, given our main research question, by comparing asymmetric and symmetric models we can assert that symmetric algorithms produce, conditioned on the underlying based algorithm, consistently more accurate estimates than asymmetric procedures (at least for the data at hand). Overall, symmetric solutions beat asymmetric solutions almost 90% of the time.

The family of the algorithm (i.e. the underlying procedure, either lphom, tslphom, or nslphom), however, has a higher impact on the accuracy of the estimates than the character, either symmetric

Table 2. Averages of EI errors of nslphom-based solutions by alternative groupings of elections

Group of elections	NZ—regular	NZ—Māori	NZ—all	SCO 2007	NZ + SCO
# of elections	$N = 443$	$N = 49$	$N = 492$	$N = 73$	$N = 565$
Avg. # of units	$\bar{I} = 65.9$	$\bar{I} = 343.1$	$\bar{I} = 93.5$	$\bar{I} = 70.2$	$\bar{I} = 90.5$
Avg. # of cells	$\overline{JK} = 26.8$	$\overline{JK} = 29.9$	$\overline{JK} = 27.1$	$\overline{JK} = 35.2$	$\overline{JK} = 28.2$
nslphom(X, Y)	9.85	10.20	9.89	8.85	9.75
nslphom(Y, X)	10.47	9.14	10.34	8.11	10.05
nslphom_min	9.54	9.83	9.57	8.42	9.42
nslphom_dual_a	9.02	7.99	8.92	**7.11**	8.68
nslphom_dual_w	**8.96**	8.04	**8.87**	**7.11**	**8.64**
nslphom_joint	9.40	**7.85**	9.24	7.53	9.02

Source: Compiled by the authors after applying, with default options, the function nslphom of the R package lphom (Pavía & Romero, 2022) and the new algorithms based on nslphom, described in Sections 3–5, with $ns = 10$ to the data from the New Zealand electoral commission and the Scotland Electoral Office available in the R package ei.Datasets (Pavía, 2022). The smaller the number, the better the accuracy. Values in bold correspond to the most average accurate solutions in each set of elections.

or asymmetric, of the model. tslphom-based models, including asymmetric ones, are on average more accurate than lphom-based models, and the same relationship occurs with nslphom-based models with regard to tslphom-based ones. In short, the nslphom-based models are clearly the most accurate, so henceforth we focus only on these by concentrating our analysis on the results available in Table 2 and Figure 1 and in the last panel of Table 1. In particular, pooling all the elections and algorithms, nslphom symmetric solutions are seen to be, on average, almost 11% more accurate than nslphom asymmetric ones.

Focusing now on the relationships within the symmetric solutions, we observe some general, although not completely systematic, patterns. On the one hand, pooling all the results we see that, on average, the nslphom_dual_w solutions are the most accurate, followed by nslphom_dual_a, nslphom_joint, and nslphom_min solutions in that order (see Table 2 and Figure 2). For some of the groups, however, nslphom_dual_w solutions are not the best on average; in some groups they are beaten by nslphom_dual_a (in the 2002 elections group) and by nslphom_joint (in the of Māori districts group).

In any case, what we can affirm is that nslphom_dual_w generates the more robust solutions. This algorithm produces, on average, the most accurate solutions despite being the algorithm that out of the four generates the smallest *EI* error fewer times. Indeed, even nslphom_min, the algorithm whose error average is the greatest out of the four, produces better solutions more times than nslphom_dual_w. In fact, nslphom_min at the individual level is the best 25% of the time, the same figure recorded for nslphom_dual_a, while the figures for nslphom_joint and nslphom_dual_w are 30% and 20%, respectively. Looking in more depth at the analysis of the individual matrices, we also find that, as expected, the nslphom_dual_w and nslphom_dual_a solutions are quite close, being midway between the nslphom_min and the nslphom_joint solutions, although a little closer to the latter.

The results for the Māori electorates are particularly interesting (see Table 2 and Figure 2). In all the groups, except the Māori, the dual solutions built as means are the best on average, when for the Māori the best on average are the predictions based on the joint algorithm. At first glance, one could think that this is a consequence of some of their differential characteristics. According to Pavía (2022), the Māori districts are remarkably different in many issues; for instance, compared to the rest of the districts of the database, their electorates are split out in a greater number of units, their transfer matrices record weaker relationships between row and column options, the electoral behaviour of their voters shows a greater heterogeneity across units and their datasets show more across-unit variances for both parties and candidates. A closer look at the individual errors, however, reveals that almost all the relative advantage of nslphom_joint compared to nslphom_dual_w and nslphom_dual_a is grounded on the 2002 elections. Indeed, nslphom_joint is only

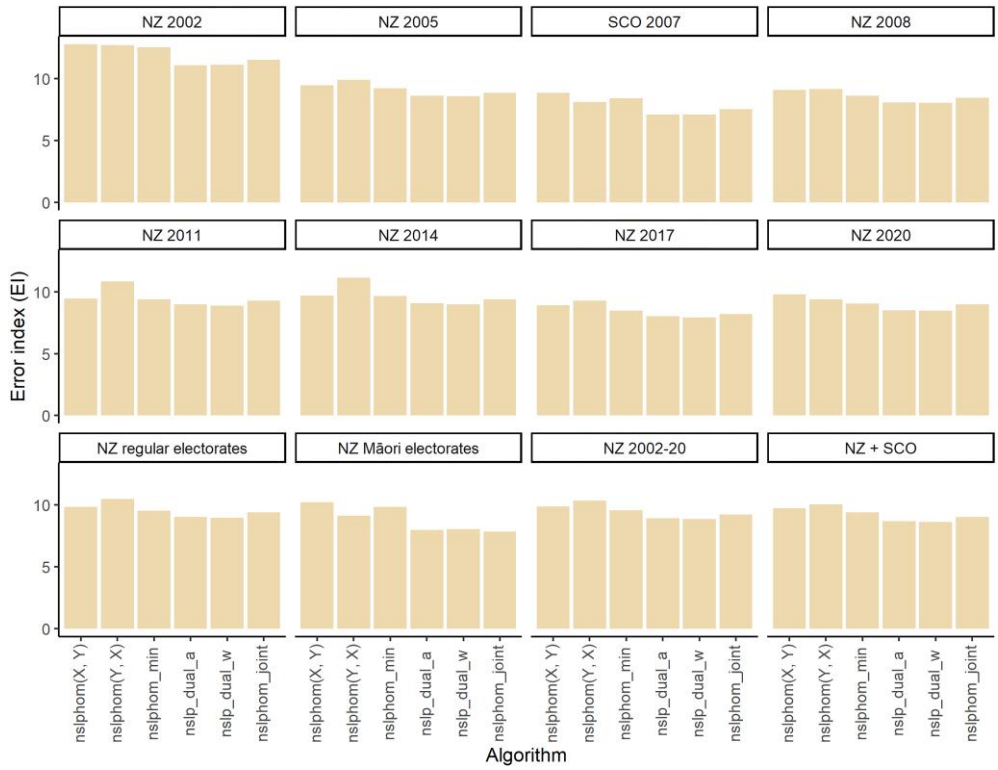

Figure 2. Graphical representation of average values of *EI* error measures grouped by election and nslphom-family algorithm. The smaller the number, the better the accuracy. Individual solutions are attained after applying, with default options, the function nslphom of the R package lphom (Pavía & Romero, 2022) and the new algorithms, described in Sections 3–5, to the data from the New Zealand electoral commission and the Scotland Electoral Office available in the R package ei.Datasets (Pavía, 2022). Detailed figures are available in Tables 1 and 2.

more accurate on average than nslphom_dual_w and nslphom_dual_a in the 2002 and 2005 elections. As a rule, therefore, we can say that, at least for datasets with similar characteristics to the ones analysed in this research, the nslphom_dual_w solutions are preferable. They are more accurate, on average, and quite robust.

Finally, focusing on computational times (see Table 3), we see that all the methods are quite fast. As expected, the nslphom solutions are slower but they only require a few seconds to reach their solutions. Obviously, the slowest algorithm is the nslphom_joint but, on average, it only takes in these datasets a few more than 20 s to reach its solution. In general, the computational burden grows with the number of units and the number of cells of the corresponding ecological matrix.

7. On the factors impacting on the accuracy of the estimates

In the previous section, we have compared the accuracy averages of the different solutions, omitting their variability. However, the accuracies of the estimates not only differ across methods but they obviously also differ across districts/elections. In this section, we consider several observed features, specific of each election, and study whether and to what extent they can explain the observed differences in accuracy across elections. To do that, we just focus on nslphom_dual_w and nlsphom_joint solutions. On the one hand, the former are revealed as the more accurate in the analysed database. On the other hand, the latter have been obtained with the theoretically finest method, as it solves the problem dealing with all the constraints simultaneously.

Figure 3, where the distributions obtained for *EI* are drawn in the 565 elections analysed, clearly shows the existence of important accuracy differences across elections in both groups of solutions. For instance, focusing on nslphom_dual_w solutions, we have that *EI* ranges from a minimum of

Table 3. Averages of computational burden (in seconds) by group of elections

Country year	NZ 2002	NZ 2005	SCO 2007	NZ 2008	NZ 2011	NZ 2014	NZ 2017	NZ 2020
lphom-based algorithms								
lphom(X, Y)	0.20	0.18	0.06	0.12	0.20	0.19	0.18	0.38
lphom(Y, X)	0.23	0.21	0.06	0.24	0.33	0.34	0.30	0.36
lphom_dual	0.42	0.40	0.12	0.31	0.48	0.42	0.44	0.74
lphom_joint	1.08	0.98	0.25	0.76	1.14	1.01	1.05	1.81
tslphom-based algorithms								
tslphom(X, Y)	0.69	0.60	0.46	0.51	0.62	0.63	0.60	1.01
tslphom(Y, X)	0.72	0.61	0.47	0.62	0.72	0.69	0.77	0.97
tslphom_dual	1.41	1.23	0.89	1.12	1.35	1.27	1.33	1.94
tslphom_joint	3.85	2.78	2.36	2.43	3.10	2.89	2.95	4.29
nslphom-based algorithms								
nslphom(X, Y)	5.11	4.35	4.07	4.11	4.73	4.55	4.61	6.43
nslphom(Y, X)	5.07	4.39	3.96	4.18	4.79	4.61	4.82	6.43
nslphom_dual	10.22	8.81	8.06	8.33	9.03	8.86	9.52	12.87
nslphom_joint	29.03	19.39	21.44	18.09	21.90	21.39	20.94	28.99

Source: Compiled by the authors after applying, with default options, the functions lphom, tslphom, and nslphom of the R package lphom (Pavía & Romero, 2022) and the new algorithms, described in Sections 3–5, to the data from the New Zealand electoral commission and the Scotland Electoral Office available in the R package ei.Datasets (Pavía, 2022). In the case of nslphom-based models, *ns* is always fixed at 10. The computations have been performed on a laptop with a CPU processor Intel Core i7-6820HK (4 cores) 2.70 GHz and 64GB of RAM.

2.45 to a maximum of 20.01; with the minimum being observed in an election where voters are distributed into 90 polling stations and the problem consists in estimating a transfer matrix of size 5×5 and the maximum being recorded in an election in which there are just data from 38 polling stations to estimate a 7×7 matrix. Indeed, this remark properly exemplifies a usual pattern/relationship typically shown for ecological inference estimators, which tend to be more accurate, the larger is the ratio between units and unknowns, $R = I/JK$. In particular, in our application, the *EI* average values for, respectively, nslphom_dual_w and nslphom_joint are 12.47 and 12.43 when $R \leq 1$; 9.46 and 9.97 when $1 < R < 2$; and 8.08 and 8.43 when $R \geq 2$.

The accuracy of the attained estimates, therefore, depends on structural characteristics of the particular election under study. In what follows, we analyse whether the observed variables previously identified in the literature as explainers of accuracy for asymmetric methods also work for the proposed symmetric models. In particular, the list of features recognized in the literature (e.g. King, 1997; Klima et al., 2016; Park et al., 2014; Pavía & Romero, 2023; Plescia & De Sio, 2018; Wakefield, 2004) as factors impacting on the accuracy of the estimates include: (i) the amount of available information, I; (ii) the complexity of the problem, JK; (iii) the level of heterogeneity among the local transfer matrices, *HET*; (iv) the strength of the relationship between the row and column categories, χ^2; (v) the degrees of diversity/variability within-units in both row and column marginal distributions, *DWR* and *DWC*, which account for how similar/different the sizes of different groups/categories are; and (vi) the extent of the variability across-units in both row and column marginal distributions, *VAR* and *VAC*, which measure how similar/different are the margin distributions across tables.

As a rule, ceteris paribus, the larger the number of polling stations and the smaller the contingency tables (the number of coefficients to be estimated), the more accurate the estimates are; with the impact of each of these two features being conditioned by the value taken by the other. Indeed, it is customarily stated that a ratio of at least two units (polling stations) per coefficient is necessary for a proper estimation of the vote transfer matrix (Plescia & De Sio, 2018). Thus, the impacts of I and JK are usually assessed together using as joint indicator the number of polling stations divided by number of columns times the number of rows: $R = I/JK$.

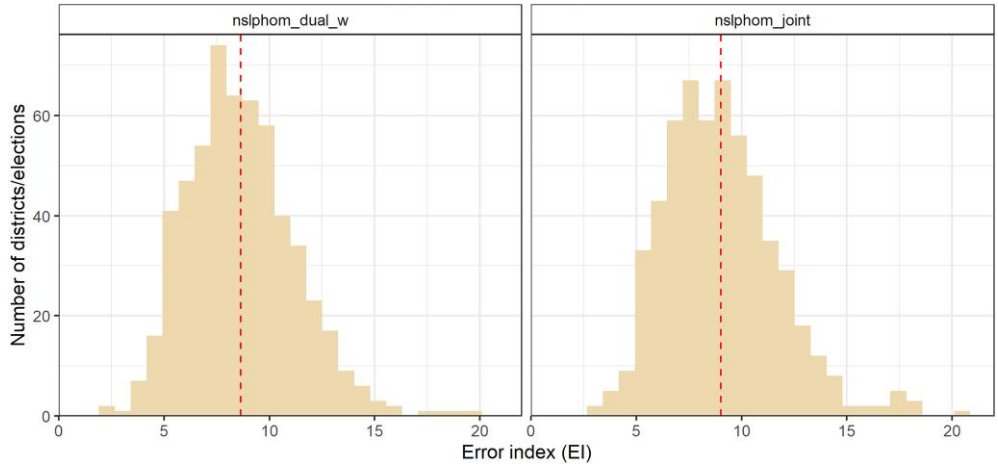

Figure 3. Histograms of the distributions of the error indexes (*EI*) corresponding to the nslphom_dual_w and nslphom_joint solutions attained in the 565 elections analysed. The discontinuous vertical lines place the means of the distributions.

The values HET and χ^2 are commonly unknown as they both depend on the actual district vote transfer matrix. Although in our case they could be exactly computed, we prefer to consider a typical situation and to study the impact of these features relying on their estimates, which are contingent to the specific solution reached. In particular, we approximate HET using the $HETe$ coefficient given by equation (33) and χ^2 through $\chi^2\mathrm{e} = \sum_{jk}\left[(\hat{v}_{jk} - \sum_j \hat{v}_{jk}\sum_k \hat{v}_{jk})^2/(J-1)(K-1)\sum_j \hat{v}_{jk}\sum_k \hat{v}_{jk}\right]$, where the divisor $(J-1)(K-1)$ accounts for the different number of cells that each matrix has.

For approximating district within-unit diversities, different statistics have been tried in the literature. They have been measured as (a) (weighted) averages of the standard deviations of the unit marginal distributions (using as weights the number of voters per unit), (b) directly using the standard deviations of the district marginal distributions, and (c) employing the formula $1 - \sum_h \pi_h^2$, with π_h representing the proportion of votes gained by option h in the total population. In practice, these three measures are (almost) equivalent. In our dataset, the correlations between (a) and (b), (a) and (c), and (b) and (c) are, respectively, 0.99, −0.99, and −0.99 for parties and 0.99, −0.99, and −0.98 for candidates. We approximate DWR and DWC using (b), the standard deviations of the district margin distributions of parties and candidates, respectively. Finally, to measure the extent of the diversity/similarity of party and candidate distributions across polling stations, we have used the compositional total variance (Pawlowsky-Glahn et al., 2015): $VAR = (2J)^{-1}\sum_{j,j'}^{J}\mathrm{Var}(\{log(x_{ij}/x_{ij'})\}_i)$ and $VAC = (2K)^{-1}\sum_{k,k'}^{K}\mathrm{Var}(\{log(y_{ik}/y_{ik'})\}_i)$, where the variances are computed across the I units. Note that in whatever application both VAR and VAC need to be different from zero for the ecological inference problem to have a solution, as all the methods learn from the statistical covariations between the polling stations margin distributions.

Once defined the features, we study using regression linear models their impact on the nslphom_dual_w and nslphom_joint solutions' accuracy. To do that, we adopt a two-step strategy. First, we analyse the marginal impact of each feature not excluding the possibility that its (increasing/decreasing) effect operates at a decreasing/increasing rate. That is, we consider the possibility of nonlinear effects in the one-feature models by also including as potential regressor the square of the feature. Second, we jointly estimate the impact on accuracy of all the regressors identified as statistically significant in the marginal (one-feature) models. To fit the models, we have excluded the Māori electorates from the analysis. Although the aggregate results are quite similar when these elections are also included in the analysis, we have opted to not consider them to fit the model because of its singular characteristics that would undermine the impact of the ratio I/JK. In fact, the Māori elections can be observed as non-standard as they are

Table 4. Statistical summary of the explicative variables, excluding Māori electorates ($N = 516$)

	I/JK	$HETe\_d$	$HETe\_j$	χ^2e_d	χ^2e_j	DWR	DWC	VAR	VAC
Mean	2.56	4.25	4.01	2932	2909	0.18	0.21	0.98	0.85
Standard deviation	1.19	1.25	0.95	966	960	0.04	0.04	0.36	0.35
Minimum	0.52	2.42	2.16	377	347	0.06	0.09	0.29	0.12
Maximum	7.73	9.44	8.45	6297	6613	0.34	0.33	2.77	2.12
Skewness	1.16	1.05	0.94	0.40	0.40	0.06	−0.13	1.10	0.74

Source: Compiled by the authors through $\chi^2\mathrm{e} = \sum_{jk}\left[(\hat{v}_{jk} - \sum_j \hat{v}_{jk}\sum_k \hat{v}_{jk})^2/(J-1)(K-1)\sum_j \hat{v}_{jk}\sum_k \hat{v}_{jk}\right]$, $VAR = (2J)^{-1}\sum_{j,j'}^{J}\mathrm{Var}(\{log(x_{ij}/x_{ij'})\}_i)$, and $VAC = (2K)^{-1}\sum_{k,k'}^{K}\mathrm{Var}(\{log(y_{ik}/y_{ik'})\}_i)$, and using equation (33) and the standard deviations of the district marginal distributions of parties and candidates to calculate $HETe$, DWR, and DWC, respectively. The suffixes _d and _j stand for dual and joint, respectively.

characterized for having their electorates distributed into a large number of polling stations, quite geographically spread, with the majority of them having very few voters.

Table 4 presents a statistical summary of the values corresponding to the abovementioned features in the 516 districts that remain in the database after excluding the Māori electorates. As can be observed, almost all variables show asymmetric positive, right-skewed, distributions. In terms of correlations, the largest ones are observed for the pairs of variables I/JK and DWR (0.74), I/JK and χ^2e (0.53), $HETe\_d$ and DWC (−0.64), and $HETe\_j$ and DWR (−0.55). Given the large sample size, we do not expect this to represent a problem in interpreting the models obtained.

In order to facilitate the interpretation of the estimated coefficients and the relative impact of their associated features on accuracy, all the explanatory variables have been standardized to zero mean and standard deviation 1. In this way, each estimated coefficient informs about the expected variation from the mean in the response variable due to one standard variation in the corresponding variable. From the one-feature models (not shown), one can infer that (i) I/JK is the feature with the largest impact on accuracy, followed by DWC and χ^2e, (ii) two variables (I/JK and χ^2e) show a significant curvature in its relationship with EI, i.e. accuracy improves but at decreasing rates as both variables grow, and (iii) accuracy also improves when either row or column within-units or across-units variances increase.

Some changes between the marginal and joint impacts of the features happen when we model the full multivariate specifications. Table 5 presents the coefficients of the fitted linear regression models. All variables, except χ^2e^2 and DWR, show a statistically significant impact. Together these variables explain around 25% of the observed variability of EI. As expected, larger I/JK ratios lead to more reliable estimates. This feature is moreover the one with the largest impact. The other two variables impacting positively by increasing accuracy (i.e. by reducing EI) are the column within-units and across-units variabilities. If in the one-feature models both row and column within-units and across-units variabilities improve accuracy in a similar fashion, when we consider all the features simultaneously, DWR lost its statistical significance and VAR reverses its sign. This result may seem puzzling in a first glance, given the symmetric nature of the algorithms, but their different impact in this database is consequence of the differences between the row and column distributions of variabilities/diversities. Finally, the models also show that when we control for the rest of variables, marginally the larger the heterogeneities or the relationships between row and column categories, the less accurate the solutions are, ceteris paribus.

8. Discussion, conclusions, and future research

Ecological inference methods are devised to infer conditional distribution probabilities from marginal distributions. In doing so, they consider a main characteristic variable (e.g. race or social class) impacting on a response variable, usually the vote. This scheme fits the majority of ecological inference problems. In some instances, however, such as in simultaneous elections, this general scheme can be questionable. The issue is that different solutions are achieved depending on which variable is considered the factor (origin, explanatory, cause) and which the response. The methods

Table 5. Impact of different features on solutions' accuracy

Variable	Response variable: error index (*EI*)			
	nslphom_dual_w		nslphom_joint	
	Estimate	*p*-Value	Estimate	*p*-Value
Intercept	8.23	<0.0001	8.72	<0.0001
I/JK	−1.60	<0.0001	−1.89	<0.0001
$(I/JK)^2$	0.32	<0.0001	0.37	<0.0001
$HETe$	0.46	0.0009	0.84	<0.0001
$\chi^2 e$	0.78	<0.0001	1.15	<0.0001
$\chi^2 e^2$	0.14	0.0583	0.05	0.5368
VAR	0.67	<0.0001	0.43	0.0027
VAC	−0.66	<0.0001	−0.51	0.0001
DWR	0.23	0.1904	0.34	0.0549
DWC	−0.44	0.0302	−0.52	0.0042
Adjusted R^2 (%)	24.90		25.73	
Residual Std. error	2.21		2.33	

Source: Compiled by the authors. All the predictor variables were standardized before fitting the models to make comparisons of coefficients easier.

are asymmetric. In this paper, we ask whether we can obtain some advantage in terms of accuracy by dealing with this inverse problem in a symmetric way. That is, by solving it as a purely mathematical puzzle that must verify some congruence properties, omitting the possible presence of a natural a priori relationship (i.e. a logical mapping of the variables to rows and columns). Afterwards, the researcher can recover the meaningful order when presenting the outcomes.

To answer the above research question, this paper builds within the linear programming framework two new families of algorithms whose solutions do not depend on how variables are mapped to rows and columns. These are symmetric methods. From a statistical standpoint, symmetric approaches offer the advantage of adhering to the probabilistic symmetry condition, meaning that their solutions conform to Bayes theorem. It should be noted, however, that while the symmetric condition could be considered a desirable property, it alone does not guarantee the attainment of accurate solutions. For example, an algorithm that assumes conditional independence between rows and columns given the unit (equivalent to the nonlinear neighbourhood model; Freedman et al., 1998) would multiply the average error by more than four times in the datasets analysed in this paper, even verifying the symmetric condition.

We evaluate the accuracy of the proposed methods using real data corresponding to 565 simultaneous elections where the true district-level cross-classifications of votes are known. Our empirical assessment indubitably points to the proposed symmetric solutions as being more accurate than the equivalent asymmetric approaches, with the ratio between available information and complexity/difficulty of the problem as the characteristic having more impact on accuracy. Overall, the optimal solutions reached with the joint specifications, those which solve linear programmes that simultaneously include all the constraints, are not the most accurate on average. They are outperformed by the methods that build their solutions as an average of asymmetric solutions. We conclude that for the datasets at hand, the nslphom_dual_w solutions are, on average, the most accurate. This method grounds its better accuracy on its greater robustness since it is the one generating the smallest error fewer times among the nslphom symmetric models. Indeed, there is no symmetric algorithm that systematically beats the rest of the symmetric methods, so a question to be answered is whether the circumstance itself could indicate which symmetric method to choose. In other words, can we find specific configurations derivable from the observed data that calls for the recommendation of a specific method for a particular dataset?

Given that mathematical programming offers the natural methodological setting for simultaneously handling all the constraints linked to a symmetric treatment of the problem, we have considered only algorithms within the linear programming framework where we have built the symmetric solutions from asymmetric methods. This has the advantage of maintaining all the proposals within the same framework, making comparisons fairer and producing methods easy to use and robust to claims of hacking. The user only needs to input a set of ecological data and a maximum number of iterations, and the procedures automatically return a sensible solution. Nevertheless, it would be worth extending the analysis to other methodological settings, i.e. to study whether similar results would be obtained if symmetric methods were built from scratch or using asymmetric algorithms from other frameworks. For example, the following question could be addressed: can inferences attained as an average of the two dual solutions of the Rosen et al. (2001) $R \times C$ model systematically improve the asymmetric solution achieved by choosing the most logical approach (of the two)? Indeed, in our view, the new angle for tackling the ecological inference problem offered by this research should also be systematically explored from other conceptual frameworks. This, however, demands new theoretical developments. More specifically, just as ecological inference linear programming requires new developments to introduce information from polls into its models (Pavía, 2023), an effort is also required to develop genuine symmetric models from other frameworks, such as the Bayesian and frequentist ecological inference ones. In this case, we consider that models based on conditional multivariate hypergeometric distributions or full-table multinomial distributions could prove successful.

Our empirical results are based on data from simultaneous elections so it would also be worth exploring whether our conclusions can be confirmed in other contexts, such as voting rights litigation or literacy studies. Despite electoral datasets with true answers in other target application areas being scarce and typically not available, it would be interesting to replicate the study of Barreto et al. (2022) on racially polarized voting and analyse the impact of using symmetric approaches on substantive conclusions and on the accuracy in estimating the inner-cells values of their simulated datasets. Likewise, the symmetric approach could also be tested estimating the tables of the datasets employed by Jiang et al. (2020). These datasets include data on US mortality rates by gender and race, or literacy rates and educational attendance by gender in India.

Finally, from a more methodological perspective, it would be interesting to study what impact initializing the iterative process with a different matrix of votes would have on the accuracy of the nslphom_joint optimal solutions; for instance, what the impact would be of initializing with the nslphom_dual_w solution instead of with the lphom_joint solution. Furthermore, given the lack of theoretical results about the (asymptotic) distributions of the estimators, we consider that the bootstrap approach (e.g. Efron & Tibshirani, 1994) could be used to measure the precision of estimates. Despite the fact that lphom-based specifications can be theoretically observed as a linear absolute fitting problem after applying Theorem 1 of chapter 6 in Bloomfield and Steiger (1983, pp. 164–165) and, therefore, their estimators 'usually' be considered as having 'a limiting normal distribution' (Bloomfield & Steiger, 1983, p. 44), in our view, this does not apply as a rule in the ecological inference problem and definitely cannot be used in the tslphom- and nslphom-based specifications. On the one hand, it is difficult to accept that the assumptions required by the different asymptotic theorems hold in the ecological inference problem, due to, among other issues, the discrete character of the margins and the cross structures of relationships that the row and column aggregations impose. On the other hand, and more importantly, the tslphom- and nslphom-specifications do not fulfil the hypothesis of the abovementioned Theorem 1, as it requires that the number of summands (auxiliary variables) in the objective function be greater than the number of equality constraints. The unit tables algorithms, lphom_local and lphom_local_joint, that are in the core of the tslphom- and nslphom-family models solve linear programmes where the number of auxiliary variables is smaller than the number of equality constraints.

The above issues do not imply that ecological inference linear programming approaches lack a statistical interpretation or that the estimated coefficients necessarily exhibit undesirable statistical properties. On the one hand, in addition to the established links between linear programming solutions and minimization of expected discrepancies, as previously mentioned when discussing the proposed models, it is important to note that tslphom- and nlsphom-based approaches, like many other ecological inference models, rely on the underlying assumption of independence across units of the two-way distributions of votes, given their aggregate two-way joint distribution. In other

words, the set of unit two-way fraction distributions can be viewed as a simple random sample of a common probability distribution. This entails that, without covariates, estimates can only be reliable when models are applied in homogeneous political regions. On the other hand, we believe that similar to how sampling properties of quadratic programming solutions can be obtained through the links between quadratic programming and inequality-constrained normal regression (Geweke, 1986; Judge & Takayama, 1966), the connection between linear programming and inequality-constrained linear absolute fitting with Laplace-distributed errors could be explored to derive statistical properties of basic ecological inference linear programming solutions. In addition, we also find remarkable that the iterative algorithm that characterize nslphom-based methods can be understood in terms of an expectation-maximization algorithm (Dempster et al., 1977). In this comparison, equations (6)–(13) would correspond to the expectation step, equation (14) to the maximization step, and equations (1)–(5) are required for generating an initial estimate of the expected transition probabilities. This connection becomes more evident when examining the variants introduced in the lclphom algorithm (Pavía, 2024).

Acknowledgements

The authors wish to thank the editors and four anonymous reviewers for their valuable comments and suggestions and M. Hodkinson for revising the English of the paper.

Conflicts of interest: none declared.

Funding

This research has been supported by Conselleria de Educación, Universidades y Empleo, Generalitat Valenciana [grant number AICO/2021/257] and by Ministerio de Economía e Innovación [grant number PID2021-128228NB-I00].

Data availability

The data used in this research is publicly available on the R package ei.Datasets (version 0.0.1-1) accessible on CRAN. The reproducible ad hoc R-code employed, based on functions included in the R package lphom (version 0.3.0-7), is available in the attached online supplementary material.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Andreadis, I., & Chadjipadelis, T. (2009). A method for the estimation of voter transition rates. *Journal of Elections, Public Opinion and Parties*, *19*(2), 203–218. https://doi.org/10.1080/17457280902799089

Barreto, M., Collingwood, L., Garcia-Rios, S., & Oskooii, K. A. R. (2022). Estimating candidate support in voting rights act cases: Comparing iterative EI and EI-R_C methods. *Sociological Methods & Research*, *51*(1), 271–304. https://doi.org/10.1177/0049124119852394

Bernardini-Papalia, R., & Fernández-Vázquez, E. (2020). Entropy-based solutions for ecological inference problems: A composite estimator. *Entropy*, *22*(7), 781. https://doi.org/10.3390/e22070781

Bloomfield, P., & Steiger, W. L. (1983). *Least absolute deviations: Theory, applications and algorithms*. Birkhäuser-Springer.

Brown, P. J., & Payne, C. D. (1986). Aggregate data, ecological regression and voting transitions. *Journal of the American Statistical Association*, *81*(394), 452–460. https://doi.org/10.1080/01621459.1986.10478290

Collingwood, L., Decter-Frain, A., Murayama, H., Sachdeva, P., & Burke, J. (2020). *eiCompare: Compares ecological inference, Goodman, rows by columns estimates*. R package version 3.0.0. https://CRAN.R-project.org/package=eiCompare

Collingwood, L., Oskooii, K., Garcia-Rios, S., & Barreto, M. (2016). eiCompare: Comparing ecological inference estimates across EI and EI:R×C. *The R Journal*, *8*(2), 92–101. https://doi.org/10.32614/RJ-2016-035

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Efron, B., & Tibshirani, R. J. (1994). *An introduction to bootstrap*. Chapman and Hall/CRC.

Ferree, K. E. (2004). Iterative approaches to RxC ecological inference problems: Where they can go wrong and one quick fix. *Political Analysis*, *12*(2), 143–159. https://doi.org/10.1093/pan/mph011
Forcina, A., & Pellegrino, D. (2019). Estimation of voter transitions and the ecological fallacy. *Quality & Quantity*, *53*(4), 1859–1874. https://doi.org/10.1007/s11135-019-00845-1
Freedman, D. A., Klein, S. P., Ostland, M., & Roberts, M. R. (1998). A solution to the ecological inference problem (book review). *Journal of the American Statistical Association*, *93*(444), 1518–1522. https://doi.org/10.2307/2670067
Freedman, D. A., Ostland, M., Roberts, M. R., & Klein, S. P. (1999). Reply to G. King. *Journal of the American Statistical Association*, *94*(445), 355–357. https://doi.org/10.2307/2669735
Gelman, A., Park, D. K., Ansolabehere, S., Price, L. C., & Minnite, L. C. (2001). Models, assumptions and model checking in ecological regression. *Journal of the Royal Statistical Society, Series A*, *164*(1), 101–118. https://doi.org/10.1111/1467-985X.00190
Geweke, J. (1986). Exact inference in the inequality constrained normal linear regression model. *Journal of Applied Econometrics*, *1*(2), 127–141. https://doi.org/10.1002/jae.3950010203
Glynn, A., & Wakefield, J. (2010). Ecological inference in the social sciences. *Statistical Methodology*, *7*(3), 307–322. https://doi.org/10.1016/j.stamet.2009.09.003
Goodman, L. A. (1953). Ecological regressions and the behavior of individuals. *American Sociological Review*, *18*(6), 663–664. https://doi.org/10.2307/2088121
Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, *64*(6), 610–625. https://doi.org/10.1086/222597
Greiner, D. J. (2007). Ecological inference in voting rights act disputes: Where are we now, and where do we want to be? *Jurimetrics*, *47*(2), 115–167. https://www.jstor.org/stable/29762964
Greiner, D. J., Baines, P., & Quinn, K.M. (2021). *RxCEcolInf: R x C ecological inference with optional incorporation of survey information*. R package version 0.1-5. https://CRAN.R-project.org/package=RxCEcolInf
Greiner, D. J., & Quinn, K. M. (2009). RxC ecological inference: Bounds, correlations, flexibility, and transparency of assumptions. *Journal of the Royal Statistical Society, Series A*, *172*(1), 67–81. https://doi.org/10.1111/j.1467-985X.2008.00551.x
Greiner, D. J., & Quinn, K. M. (2010). Exit polling and racial bloc voting: Combining individual level and RxC ecological data. *The Annals of Applied Statistics*, *4*(4), 1774–1796. https://doi.org/10.1214/10-AOAS353
Hawkes, A. (1969). An approach to the analysis of electoral swing. *Journal of the Royal Statistical Society, Series A*, *132*(1), 68–79. https://doi.org/10.2307/2343756
Jiang, W., King, G., Schmaltz, A., & Tanner, M. A. (2020). Ecological regression with partial identification. *Political Analysis*, *28*(1), 65–86. https://doi.org/10.1017/pan.2019.19
Johnston, R. J., Hay, A. M., & Rumley, D. (1983). Entropy-maximizing method for estimating voting data: A critical test. *Area*, *15*(1), 35–41. https://www.jstor.org/stable/20001867
Johnston, R. J., & Pattie, C. (2003). Evaluating an entropy-maximizing solution to the ecological inference problem: Split-ticket voting in New Zealand, 1999. *Geographical Analysis*, *35*(1), 1–23. https://doi.org/10.1111/j.1538-4632.2003.tb01098.x
Judge, G. G., Miller, D. J., & Tam Cho, W. K. (2004). An information theoretic approach to ecological estimation and inference. In G. King, O. Rosen, M. A, Tanner (Eds.), *Ecological inference. New methodological strategies* (pp. 162–187). Cambridge University Press.
Judge, G. G., & Takayama, T. (1966). Inequality restrictions in regression analysis. *Journal of the American Statistical Association*, *61*(313), 166–181. https://doi.org/10.1080/01621459.1966.10502016
Katz, J. N. (2014). *Expert report on voting in the city of Whittier*. Superior Court of the State of California.
Kellermann, T. (2011). Vom Wahlergebnis zur Wählerwanderung: Welche Wähler wechselten wie ihre Entscheidung. *Stadtforschung und Statistik*, *2011*(1), 34–40.
King, G. (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press.
King, G., Rosen, O., & Tanner, M. A. (Eds.) (2004). *Ecological inference. New methodological strategies*. 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., Schlesinger, T., Thurner, P. W., & Küchenhoff, H. (2019). Combining aggregate data and exit polls for the estimation of voter transitions. *Sociological Methods & Research*, *48*(2), 296–325. https://doi.org/10.1177/0049124117701477
Klima, A., Thurner, P. W., Molnar, C., Schlesinger, T., & Küchenhoff, H. (2016). Estimation of voter transitions based on ecological inference: An empirical assessment of different approaches. *AStA-Advances in Statistical Analysis*, *100*(2), 133–159. https://doi.org/10.1007/s10182-015-0254-8
Lau, O., Moore, O. R. T., & Kellermann, M. (2020). *eiPack: Ecological inference and higher-dimension data management*. R package version 0.2-1. https://CRAN.R-project.org/package=eiPack
Manski, C. F. (2007). *Identification for prediction and decision*. Harvard University Press.

O'Loughlin, J. (2000). Can King's ecological inference method answer a social scientific puzzle: Who voted for the Nazi party in Weimar Germany? *Annals of the Association of American Geographers*, *90*(3), 592–601. https://doi.org/10.1111/0004-5608.00213

Park, W.-H., Hanmer, M. J., & Biggers, D. R. (2014). Ecological inference under unfavorable conditions: Straight and split-ticket voting in diverse settings and small samples. *Electoral Studies*, *36*, 192–203. https://doi.org/10.1016/j.electstud.2014.08.006

Pavía, J. M. (2022). ei.Datasets: Real datasets for assessing ecological inference algorithms. *Social Science Computer Review*, *40*(1), 247–260. https://doi.org/10.1177/08944393211040808

Pavía, J. M. (2023). Adjustment of initial estimates of voter transition probabilities to guarantee consistency and completeness. *SN Social Sciences*, *3*(5), 75. https://doi.org/10.1007/s43545-023-00658-y

Pavía, J. M. (2024). *A local convergent ecological inference algorithm for RxC tables*.

Pavía, J. M., & Cantarino, I. (2017). Dasymetric distribution of votes in a dense city. *Applied Geography*, *86*, 22–31. https://doi.org/10.1016/j.apgeog.2017.06.021

Pavía, J. M., & Romero, R. (2022). Improving estimates accuracy of voter transitions. Two new algorithms for ecological inference based on linear programming. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221092725

Pavía, J. M., & Romero, R. (2023). Data wrangling, computational burden, automation, robustness and accuracy in ecological inference forecasting of RxC tables. *SORT—Statistics and Operations Research Transactions*, *47*(1), 151–186. https://doi.org/10.57645/20.8080.02.4

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons, Ltd.

Plescia, C., & De Sio, L. (2018). An evaluation of the performance and suitability of RxC methods for ecological inference with known true values. *Quality & Quantity*, *52*(2), 669–683. https://doi.org/10.1007/s11135-017-0481-z

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*(3), 351–357. https://doi.org/10.2307/2087176

Romero, R., & Pavía, J. M. (2021). Estimating vote party entries and exits by ecological inference. Mathematical programming versus Bayesian statistics. *Boletín de Estadística e Investigación Operativa*, *37*(2), 85–97.

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

Rosen, O., Jiang, W., King, G., & Tanner, M. A. (2001). Bayesian and frequentist inference for ecological inference: The RxC case. *Statistica Neerlandica*, *55*(2), 134–156. https://doi.org/10.1111/1467-9574.00162

Schakel, A. H., & Romanova, V. (2020). Vertical linkages between regional and national electoral arenas and their impact on multilevel democracy. *Regional and Federal Studies*, *30*(3), 323–342. https://doi.org/10.1080/13597566.2020.1774750

Tam Cho, W. K. (1998). Iff the assumption fits…: A comment on the King ecological inference solution. *Political Analysis*, *7*, 143–163. https://doi.org/10.1093/pan/7.1.143

Tam Cho, W. K., & Gaines, B. J. (2004). The limits of ecological inference: The case of split-ticket voting. *American Journal of Political Science*, *48*(1), 152–171. https://doi.org/10.1111/j.0092-5853.2004.00062.x

Thomsen, S. R. (1987). *Danish elections, 1920–79: A logit approach to ecological analysis and inference*. Politica.

Tziafetas, G. (1986). Estimation of the voter transition matrix. *Optimization*, *17*(2), 275–279. https://doi.org/10.1080/02331938608843128

Wakefield, J. (2004). Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society, Series A*, *167*(3), 385–445. https://doi.org/10.1111/j.1467-985x.2004.02046.x