Bulletin de Méthodologie Sociologique
2024, Vol. 164 97–121

DOI: 10.1177/07591063241277064
journals.sagepub.com/home/bms
Sage

Integer estimation of inner-cell values in RxC ecological tables

Jose M. Pavía
GIPEyOP, Universitat de Valencia, Spain

Résumé

Développement d'un algorithme pour l'estimation entière des entrées internes des tables de contingence RxC à partir des marges. L'estimation des entrées internes d'un ensemble de tables de contingence RxC, lorsque seules les marges sont connues, représente l'un des problèmes les plus complexes dans le domaine des sciences sociales. Ce problème est courant dans de nombreux domaines, allant du marketing à l'histoire quantitative, et est particulièrement prévalent en science politique et en sociologie. Bien que des dizaines de méthodes aient été proposées pour résoudre ce problème, presque toutes ont été conçues pour inférer des distributions de proportions conditionnelles, malgré le fait que les valeurs internes des tables de contingence soient des entiers. Ce document développe un nouveau algorithme dans le cadre de la programmation linéaire pour fournir des solutions entières, et l'évalue en utilisant des données réelles issues de plus de 500 élections où les distributions croisées réelles des votes sont connues. Bien que cette nouvelle approche soit proposée dans l'attente d'obtenir des solutions plus précises en réduisant l'espace de recherche d'un simplexe continu à un espace discret, les résultats obtenus suggèrent que l'utilisation d'une approche entièrement entière ne conduit pas à des solutions plus précises. Il est donc recommandé d'ajuster les solutions décimales en entiers lorsque l'accent est mis sur les comptages. Les praticiens intéressés peuvent facilement utiliser les nouveaux modèles car ils ont été programmés dans le package R lphom.

Abstract

The estimation of the inner entries of a set of related RxC contingency tables when only the margins are known poses one of the most difficult problems in the field of the social sciences, present in many areas from marketing to quantitative history; being particularly

Corresponding Author:
Jose M. Pavía, Facultat d'Economía, Universitat de Valencia, Av. Tarongers, s/n, 46022-Valencia, Spain.
Email: pavia@uv.es

prevalent in political science and sociology. With dozens of methods proposed to solve this, (almost) all of them have been devised to infer conditional proportion distributions, despite interior values of contingency tables being integers. This paper develops, within the linear programming framework, a new algorithm to output integer solutions, and assesses it using real data from more than 500 elections where actual cross-distributions of votes are known. Although the new approach is proposed with the expectation that more accurate solutions would be obtained by narrowing the search space from a continuous to a discrete simplex space, the results attained suggest that the use of a pure integer approach does not lead to more accurate solutions. The recommendation is therefore to integer-adjust decimal solutions when the focus is on counts. Interested practitioners can easily use the new models as they have been programmed in the R-package lphom.

Mots-Clés
droits de vote, inférence écologique, lphom, matrices de transition des votes, programmation linéaire en nombres entiers, tables de contingence RxC

Keywords
ecological inference, integer linear programming, lphom, RxC contingency tables, vote transition matrices, voting rights

Introduction

The estimation of the inner-cells values of a set of related RxC contingency tables when only the marginal aggregates are known poses one of the oldest and most difficult challenges in the field of the social sciences (King, 1997), present in many disciplines from marketing to quantitative history; being particularly prevalent in political science and sociology (see, e.g., King et al., 2004), where the problem is destined to persist due to the secret ballot. Ecological inference models offer solutions to this problem, especially valued in the absence of polls. They are particularly appreciated when conducting electoral analyses in historical elections and local politics, where surveys are impossible to gather or exceptional. Actually, in these contexts, analysts, faced with very few alternatives, frequently employ ecological inference models to infer individual electoral behaviours leveraging aggregate demographic data and electoral outcomes available in a set of geographical units (e.g., neighbourhoods, census sections or precincts). Audemard (2024), who uses ecological inference to estimate vote transfers between the first and second rounds of the municipal election held in Montpellier in 2014, and King et al. (2008), who employ the approach to discern the economic roots of the rise of the Nazi party to power, can be cited as recent examples.

It should be noted, however, that ecological inference models are also worthy to be employed when polls are available, as they also provide estimates at small area levels.[1] In general, having estimates of voter transitions and regarding the electoral behaviour of different population subgroups in small voting units or only for the entire

Table 1. Typical R × C unit in ecological inference

	col_1	...	col_k	...	col_C	
row_1	v_{i11}	...	v_{i1k}	...	v_{i1C}	x_{i1}
...	...	...	...	...	...	...
row_j	v_{ij1}	...	v_{ijk}	...	v_{ijC}	x_{ij}
...	...	...	...	...	...	...
row_R	v_{iR1}	...	v_{iRk}	...	v_{iRC}	x_{iR}
	y_{i1}	...	y_{ik}	...	y_{iC}	n_i

	col_1	...	col_k	...	col_C	
row_1	p_{i11}	...	p_{i1k}	...	p_{i1C}	$p_{i1\cdot}$
...	...	...	...	...	...	...
row_j	p_{ij1}	...	p_{ijk}	...	p_{ijC}	$p_{ij\cdot}$
...	...	...	...	...	...	...
row_R	p_{iR1}	...	p_{iRk}	...	p_{iRC}	$p_{iR\cdot}$
	$p_{i\cdot 1}$	...	$p_{i\cdot k}$	...	$p_{i\cdot C}$	1

Note: Red quantities correspond to unobserved counts/votes (left-panel) and unobserved row-standardized fractions/proportions (right panel). The observed election outcomes are in the margins (in black). Observed and unobserved quantities in both panels are deterministically related: $p_{ijk} = v_{ijk} / x_{ij}$, $x_{ij} = \sum_k v_{ijk}$, $y_{ik} = \sum_j v_{ijk}$, $p_{ij\cdot} = x_{ij\cdot} / n_i$ and $p_{i\cdot k} = y_{ik} / n_i$.

election space is relevant for many analysts, including sociologists, political parties, US voting rights litigants, quantitative historians and campaign strategists. Among other issues, it enables designing more tailored electoral campaigns, the identification of particular spatial groupings of voters and seeking for answers to many relevant sociological and political substantive questions. Indeed, ecological inference has been used for more than a century trying to ascertain issues such as differences of voting by gender (e.g., Hudson et al., 2010; Ogburn and Goltra, 1919) or social and economic conditions (e.g., Gosnell and Gill, 1935; Thomsen, 1987), and for disentangling historical puzzles (e.g., Brown, 1982; Penadés and Pavía, 2024).

Currently, the two most common uses of ecological inference in social sciences consist of approximating voting preferences by racial groups in the context of US voting rights disputes (Greiner, 2007), and estimating vote transfers, split-tickets and voter transitions (Pavía and Romero, 2024b). Certainly, inferring the cross-distribution of votes (see left-panel in Table 1) or transition probabilities/proportions (see right-panel in Table 1) between two elections from aggregate election outcomes has been one of the most frequently visited problems in the ecological inference literature for decades (e.g., Antweiler, 2007; Corominas et al., 2015; El Deeb, 2023; Füle, 1994; Irwin and Meeter, 1969; Tziafetas, 1986; Upton, 1978; Vangrevelinghe, 1961). We will focus on this problem for the remainder of the paper.

The difficulty of the problem arises because there are many combinations for the interior cell counts of a table with the same aggregated margins; an issue that allows room for the so-called ecological fallacy to occur (Robinson, 1950). This indeterminacy, which is intrinsic to the endeavour, cannot be completely ruled out and, under proper circumstances, can give rise to aggregation bias or spurious correlations (Cho and Manski, 2008; Forcina and Pellegrino, 2019; Manski, 2007). Indeed, according to Gnaldi et al. (2018), there are even scenarios where the estimation bias can never be eliminated. To solve the problem, practitioners usually assume (sometimes with the help of covariates) similar/related conditional row fraction/probability distributions across tables (Jiang et al., 2020). This assumption responds to the qualitative observation that people belonging to the same group tend to follow (at least probabilistically) similar behaviour patterns (Pavía and Romero, 2024a).

Quantitatively, different models emerge depending on how the hypothesis is operationalized. Dozens of methods from fields as varied as information theory (e.g., Bernardini-Papalia and Fernández-Vázquez, 2020; Johnston and Hay, 1983; Johnston and Pattie, 2000; Judge et al., 2003), mathematical programming (e.g., McCarthy and Terence, 1977; Pavía and Romero, 2024; Pavía, 2024a; Romero et al., 2020), latent structure theory (Pavía and Thomsen, 2024; Thomsen, 1987), or frequentist (e.g., Brown and Payne, 1986; Freedman et al., 1991; Forcina et al., 2012; Goodman, 1953, 1959; Hawkes, 1969) and Bayesian statistics (e.g., Greiner and Quinn, 2009; Imai et al., 2008; King, 1997; King et al., 1999; Klima et al., 2019; Puig and Ginebra, 2014; Rosen et al., 2001) have been proposed over time, with the ones suggested by Rosen et al. (2001) and Pavía and Romero (2024a)—programmed respectively in R packages eiPack (Lau et al., 2020) and lphom (Pavía and Romero, 2022)[2] – standing out as the main alternatives for RxC tables (Klima et al., 2016; Pavía and Romero, 2023; Plescia and De Sio, 2018; Romero and Pavía, 2021).

However, despite the (marginal and) interior cell numbers of contingency tables being integers, (almost) all the methods have been devised to infer row standardized distributions, with cross-tabulation counts derived after multiplying the observed row margins by the estimated proportions/fractions/probabilities.[3]

There are nuanced differences between estimating fractions and underlying probabilities. Fractions quantify the actual electoral behaviour of voters, while probabilities measure voters' propensities to behave in a particular manner given their characteristics, background, context, past behaviour, or latent preferences. The media, the public, and political parties are usually interested in knowing how electors have actually voted (and why), therefore their primary focus consists of imputing the missing cells within the context of a finite population, also measuring uncertainties. Conversely, some social scientists are more interested in estimating probabilities and their levels of credibility, as commonly seen in epidemiology and health sciences, therefore their primary focus centres on estimating conditional underlying row-probabilities under the assumption of an underlying generating statistical process. From this perspective, the primary objective when estimating fractions is prediction, whereas estimating causal effects is the customary objective when estimating underlying probabilities. Both objectives are not incompatible; they can be achieved using both approaches.

Among the two procedures identified as most accurate in the literature, the ei.MD.bayes function of the eiPack package estimates conditional row underlying probability distributions, while the nslphom function of the lphom package conditional row fractions. In the case of ei.MD.bayes, this entails that its estimations, which are attained within a Bayesian framework, are not completely congruent with the observed data. When the inferred row-probability distributions are multiplied by the known row margins, the results do not completely fulfill the sum-column constraints.[4] On the contrary, the nslphom estimates are in agreement with the observed data. The recently proposed lphom-family methods, devised within the linear programming framework, generate congruent solutions by design, although they also return decimal cross-classification counts.[5]

One possible way to ensure congruence with integer estimation of counts would be to adjust, subject to the row- and column-sum constraints, the final (decimal) estimates to whole numbers. This is an issue that could be straightforwardly done within a linear programming framework using integer linear programming. The nslphom estimates, nevertheless, are achieved using an iterative algorithm that in each step seeks to improve the local/unit solutions.[6] Therefore, given the actual integer nature of the real cross-distributions of votes, it seems reasonable to expect that more accurate solutions would be obtained if integer local estimates were used in each iteration.

The aim of this article is twofold. On the one hand, it proposes a modification of the lphom_local (LL) model, on which the lphom-family methods are grounded, to output integer solutions. We call this new model integer_lphom_local (ILL). On the other hand, it evaluates whether, as expected, the lphom-family procedures (lphom, tslphom, nslphom and lclphom) would give rise to more accurate cross-distribution estimates if the ILL model were employed instead of the LL model.

At first glance, the expectation of achieving more accurate estimates with the ILL model could be seen as contradictory to maximum likelihood (ML) and optimization principles given that adding more constraints to a model reduces its degrees of freedom, dismissing its ability to generate estimates/predictions close to the observed data. Nevertheless, this is not always the case. Despite ML estimates presenting good properties, it is well known that in some cases (as, for instance, happens in ecological regression) the unconstrained log-likelihood function may present spurious maxima and singularities (Wakefield, 2004), with more accurate results being attained when constraints are added to the model. Indeed, in specific ML settings, it has been proved, working with an iterative sequence of estimates attained using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), that sometimes considering constrained ML can produce better estimates than applying unconstrained ML (Greselin and Ingrassia, 2015).

Adding new constraints to a model only leads inevitably to an increase in the discrepancies between observations and model expected predictions/estimates when the constraints are alien to the problem. In the case in hand, the true inner-cell values are integers and the estimates/predictions that we obtain with both (constrained and unconstrained) models show exactly the same discrepancies with the observed data. The nslphom model, moreover, relies on an iterative algorithm, related to the EM algorithm (Pavía and Romero, 2024b), that in each step tries to improve the global solution by improving local (unit) solutions to, in the next step, try to improve local solutions using the global solution from the previous step, and so on. The expectation is, therefore, that better final solutions would eventually be reached by improving local solutions in each step. Although, at the global level, the estimated proportions are almost the same if we work with either decimal or whole numbers, rounding does have an impact on the estimates of the local transition probabilities. At the local level, significant differences in the estimated proportion transition matrix emerge depending on whether this matrix is directly obtained using decimal or integer counts. So, given that the true values are integers, the expectation is that the solutions can be improved working with integer local solutions in each step. Indeed, according to Greiner and Quinn (2009: 69–70): “a method that works with counts (as opposed to fractions) inside the precinct tables … provides a closer fit to the data-generating process.”

The assessment of the variants of lphom, tslphom, nslphom and lclphom based on ILL is made using the real data of 565 elections from New Zealand (NZ) and Scotland (SCO) available in the R package ei.Datasets (Pavía, 2022). In these datasets, in addition to the margins of the polling station tables, the actual cross-distributions of votes of the total electoral spaces are also available.

The evaluation is made by calculating the distances between the known and estimated global tables of votes, with the latter being attained by (i) using the LL model (decimal solutions), (ii) integer-adjusting the final decimal solutions (adjusted solutions) and (iii) utilizing the ILL model (integer solutions). In light of the results achieved, the use of integer local estimates in each step does not lead on average, at least for these datasets, to more accurate global solutions. Hence, if having integer solutions is worthy/required in a particular application, the results of this research point to better integer-adjustment of the final decimal solutions (adjusted solutions) than to directly derive them from scratch.

The rest of the paper is structured as follows. Section 2 deals with the methodological issues, providing a statistical interpretation of the approach. The mathematical details of the proposal, including the formulae for the new (ILL) local model, are presented in the Appendix. Section 3 introduces the data and outlines the measure of accuracy. Section 4 develops an example. Section 5 compares the different solutions and assesses the results. Finally, Section 6 discusses the findings and concludes.

Methodology

In ecological inference forecasting (Petropoulos et al., 2022), the units of analysis are I double-entry tables (matrices) of size $R \times C$ whose row and column margins are known, and the goal is to estimate the RC unobserved internal counts (or proportions/probabilities) of each unit/local table (see Table 1), and/or of the aggregation of all the tables (global table). Without loss of generality, we consider the case of estimating the matrix of transfer of votes between two elections, E1 and E2, and focus on the simpler case of simultaneous elections (split-ticket).[7]

The basic unknowns are the number of voters v_{ijk} who simultaneously voted options j and k in unit i in E1 and E2, as well as the equivalent figures $v_{jk} = \sum_i v_{ijk}$ in the entire population, and the observed data are the matrices $\boldsymbol{X} = [x_{ij}]$ and $\boldsymbol{Y} = [y_{ik}]$; where $x_{ij} = \sum_k v_{ijk}$ and $y_{ik} = \sum_j v_{ijk}$ account respectively, $\forall i, j, k$, for the votes recorded by the j ($j = 1, 2, \ldots, R$) and k ($k = 1, 2, \ldots, C$) election options of E1 and E2 in unit i ($i = 1, 2, \ldots, I$). These matrices are of respective orders $I \times R$ and $I \times C$.

The unknown counts, v_{ijk} and v_{jk}, are estimated using as intermediate unknowns the proportions p_{jk} and p_{ijk}; where p_{jk} is the proportion of voters in the entire electoral space who vote for candidate k among those who vote for party j and p_{ijk} is the corresponding proportion at the unit level.[8] The various ecological inference models differ in how they manage to infer the proportions (underlying probabilities) p_{jk} and p_{ijk} from the observed data ($\boldsymbol{X}$ and $\boldsymbol{Y}$). The lphom-family models (lphom, tslphom, nslphom and lclphom) attain their estimates by systematically exploiting the assumption of similarity of the p_{ijk} proportions (conditional row distributions) across tables under the constraint that the (global and local) estimates must be congruent with the observed data. Statistically,

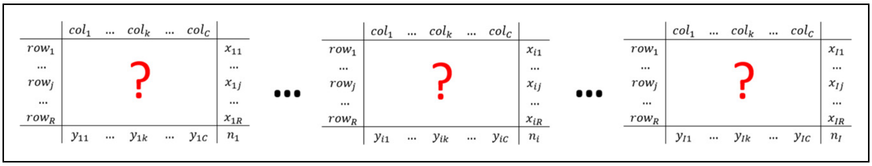

Figure 1. Schematic representation of the ecological inference problem. Ecological inference models infer the internal, unobserved cells of a set of related tables by learning from the joint covariations of the margins of the tables, under certain assumptions about the conditional row distributions across tables

this is equivalent to consider the set of unobserved unit tables a simple random sample of a common probability distribution (Pavía and Romero, 2024b). Figure 1 presents schematically the ecological inference problem.

In particular, these models rely on the (lphom_local) LL model (see equations A1-A8 in the Technical Appendix) to yield their estimates (Pavía and Romero, 2024a). The tslphom, nslphom, and lclphom lphom-models build on the lphom and lphom_local procedures to generate their estimates. Initially, global estimates are obtained using the lphom model (equations A14-A18 in the Technical Appendix). These estimates are then refined iteratively and spread across units using the LL model. Each model follows a different iterative process: tslphom achieves its estimates after a single iteration, nslphom determines the number of iterations needed by minimizing an aggregate unit-based heterogeneity index (equation A19 in the Technical Appendix), and lclphom accommodates more unit deviations by employing a variable number of iterations for each unit table. Further details, including the mathematical formulae, are available in the Technical Appendix.

According to Pavía and Romero (2024b), the lphom-family solutions may be interpreted as minimum least-absolute constrained estimates. For instance, the nslphom-estimate consists of the set of simultaneously-updated unit transfer matrices that, being closest to the lphom-estimate, verify all the aggregate unit row and column constraints and lead to the compound global transfer matrix that minimizes the sum across units of the expected count discrepancies, in L_1 -norm, between the observed and expected unit column margins when the compound matrix is applied to the observed unit row totals.[9]

In this paper, we enhance the LL model to provide integer estimates, resulting in the (integer_lphom_local) ILL model (equations A1 to A13 in the Technical Appendix). The estimates returned by the LL-versions of the models generate decimal counts. To produce integer counts, we have two alternatives: (i) adjusting the final attained solutions using the third step of ILL (equations A9-A13 of the Technical Appendix) or (ii) replacing the LL procedure with the ILL procedure in the iterative steps. Therefore, the application in a dataset of the three alternatives described above yields for each model (lphom, tslphom, nslphom and lclphom) three different estimates: a solution attained using the LL-version of the model, another resulting from integer-adjusting the LL-solution and a third one obtained with the ILL-version.[10] We call these three estimates decimal, adjusted and integer solutions, respectively. In the next sections we assess their relative accuracy with real data.

Data and accuracy measure

Data

The main difficulty in completing assessments of ecological inference models in election contexts comes from the fact that actual cross-distributions of votes are as a rule unknown due to the secret ballot. Two main strategies have been suggested in the literature to overcome this issue: the comparison with data derived from polls (mostly panel surveys and exit-polls) and the use of artificial, simulated data. Both approaches, however, are not free from criticism (see, e.g., Collingwood et al., 2016; Plescia and De Sio, 2018; Romero and Pavía, 2021; Russo, 2014; Sandoval and Ojeda, 2023). In some circumstances, however, such as when multiple offices are being decided by a single election and voters cast their votes in the same ballot, actual vote transfer matrices can be known. This is the case of the 565 datasets available in the ei.Datasets package (Pavía, 2022). This package contains the relevant data on the election results recorded in the 492 districts of NZ corresponding to the legislative (general) elections of 2002, 2005, 2008, 2011, 2014, 2017 and 2020 and in the 73 districts of the 2007 Scottish Parliament election. Quantitatively, this represents a large number of examples on which to test our models, while qualitatively involving "a broad diversity of electoral contexts" (Pavía, 2022, p. 253).

In both countries, voters cast two votes in the same ballot: one for a local candidate and another for a party list. Local candidates are different in each district so the outcomes of the different districts (electorates in NZ and constituencies in SCO) can be observed as different elections, although related when they took place in the same year. The data available for each district comprises (i) the votes recorded in each polling station (voting unit) for each party and candidate and (ii) the parties-to-candidates cross-distributions of votes of the whole district. The voting unit data correspond to the margins of the local tables. The crosstabs of votes are the actual global tables. The ecological inference algorithms use (i) to guess both (ii) and the local crosstabs.

Before using the data, as is usual practice (e.g., Barreto et al., 2022; Klima et al., 2016; Vizcaino and Pavía, 2022), very small (in number of votes) election options were merged. Those parties and candidates that individually did not reach at least a 3% of the district vote were grouped in 'Other parties' and 'Other candidates', respectively. Although the merging significantly reduces the average sizes of the tables (from 147.6 to 28.2 in number of cells), there is still a large diversity of size tables, with 23 different sizes. The sizes of the tables range from a minimum of 15 cells (5×3) to a maximum of 56 cells (8×7). In terms of number of units (polling stations) by district, the average is 90.5, with a minimum of 22 and a maximum of 833. This, together with the wide diversity in terms of many other variables (such as vote concentration, heterogeneity across units or skewness) reported in Pavía (2022), configures quite a varied database that enriches the analyses and adds robustness to the findings (Park et al., 2014).

Accuracy measure

Once the four lphom-family models are applied with the three variants (decimal, adjusted and integer) described in Section 2, twelve estimates for the global cross-distributions of votes are available for each election. Therefore, to assess the goodness-of-fit of the different methods, it is necessary to quantify and summarise the distances between the estimated matrices $[\tilde{v}_{jk}]$ and the actual observed matrices $[v_{jk}]$.

As accuracy statistic, we use the error index EI; previously employed among others by, for instance, Thomsen (1987), Klima et al. (2016) and Pavía and Romero (2024a). This measure, which in the specification given by Pavía and Romero (2024a) is detailed in equation (1), accounts for the percentage of votes incorrectly assigned in the forecasted matrix. In other words, the minimum number of votes that should be moved among cells of the global table to complete a perfect fit. The smaller this number, the closer the forecasted and actual matrices.

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

An application example

This section exemplifies, through an empirical application, how the models can be applied and provides as a sample two of the estimated transfer matrices that can be attained. In the next section, a systemic analysis is completed in each of the 565 elections focused on assessing the models with three variants considered.

For illustrative purposes, we have chosen the district of Wellington Central during the 2014 New Zealand General Election. This example is quite average and was selected solely because its merged transfer matrix is of manageable size, simplifying the presentation of results. In the Wellington Central 2014 election, the electors cast two votes in the same ballot, choosing a party among fifteen parties and a candidate among ten candidates. This results in a 16×11 transfer matrix, after adding the options of voting for an informal party or candidate.

A total of 38,353 voters, distributed across 64 polling stations, cast their votes in that election, in which a majority and parties and candidates gained less than 3% of the district vote. After merging parties and candidates that individually did not surpass this threshold into ‘Other parties’ or ‘Other candidates’, the resulting transfer matrix reduced to 5×4. This matrix includes four individual parties (Green Party, Labour, National, and New Zealand First) and three individual candidates (James Shaw of the Green Party, Paul Foster-Bell of the National Party, and Grant Robertson of the Labour Party). The actual outcomes recorded in that election are presented in Table 2, which displays the transfer matrix of votes in the left panel and the parties-to-candidates row-standardized transition matrix in the right panel. The National Party was the most voted party, with 14,451 individual votes (37.7%), whereas the candidate from the Labour Party (Grant Robertson) was the most voted candidate, receiving 19,807 individual votes (51.6%).

The 64×5 and 64×4 datasets of party ($\boldsymbol{X}$) and candidate ($\boldsymbol{Y}$) votes by polling units are utilized to estimate the transfer matrix of votes. Table 3 presents the integer-adjusted (left-panel) and integer (right-panel) and estimates obtained using the nslphom model, which produces the most accurate estimates with this dataset. The outcomes obtained in this example are quite representative of the results obtained in other elections/datasets. As can be observed, the procedure successfully captures the main trends of the transfers, with the integer-adjusted estimates proving to be more accurate than the integer estimates.

Table 2. Election outcomes recorded in the district of Wellington Central during the 2014 New Zealand General Election. The actual cross-distribution of votes and parties-to-candidates row-standardized transition matrix (in percentages) are in, respectively, the left- and right-panels

	Transfer matrix of votes					Row-standardized transition matrix				
	Shaw	**Foster**	**Robert.**	**Others**	Total	**Shaw**	**Foster**	**Robert.**	**Others**	Total
Green	3372	222	7400	280	**11274**	29.9	2.0	65.6	2.5	**29.4**
Labour	697	70	8172	196	**9135**	7.6	0.8	89.5	2.1	**23.8**
National	693	10516	2761	481	**14451**	4.8	72.8	19.1	3.3	**37.7**
NZ First	81	167	711	417	**1376**	5.9	12.1	51.7	30.3	**3.6**
Others	234	565	763	555	**2117**	11.1	26.7	36.0	26.2	**5.5**
Total	**5077**	**11540**	**19807**	**1929**	**38353**	**13.2**	**30.1**	**51.6**	**5.0**	**100**

Source: Own elaboration using data available in ei.Datasets (Pavía, 2022).

Table 3. Estimates of the integer-adjusted (left panel) and integer (right panel) transfer matrices attained using exclusively aggregate data with the nslphom model for the district of Wellington central during the 2014 New Zealand general election

	Integer-adjusted estimated transfer matrix					**Integer estimated transfer matrix**				
	Shaw	**Foster**	**Robert.**	**Others**	Error	**Shaw**	**Foster**	**Robert.**	**Others**	Error
Green	3795	387	7019	73	**1.5**	3846	357	7036	35	**1.6**
Labour	106	115	8843	71	**1.9**	83	82	8929	41	**2.0**
National	474	10510	3102	365	**0.9**	451	10635	3022	343	**1.0**
NZ First	27	27	536	786	**1.0**	5	6	520	845	**1.1**
Others	675	501	307	634	**1.4**	692	460	300	665	**1.5**
Error	**2.3**	**0.5**	**2.6**	**1.2**	**6.6**	**2.4**	**0.7**	**2.7**	**1.4**	**7.2**

Source: Own elaboration. The adjusted solutions have been obtained after integer-adjusting the decimal solutions, which have been obtained after applying, with default options, the function nslphom of the R package lphom, version 0.3.1–1 (Pavía and Romero, 2022), to the data from Wellington Central district of New Zealand corresponding to 2014 available in the R package ei.Datasets (Pavía, 2022). The integer solutions have been obtained using the ILL procedure described in Section 2. The columns and rows labeled as 'Error' display the contribution of the corresponding column/row to the total EI index defined in Section 3. Please see the reproducible code for the details.

Results

Despite the inner values of (social science) contingency tables being whole numbers, (almost) all the ecological inference methods proposed estimate decimal counts. Within the ecological inference linear programming framework, we have introduced the ILL procedure as an alternative to the LL method to generate integer solutions. This section aims to gauge whether, in addition to this advantage, the use of this procedure also leads, as could be reasonably expected, to more accurate results. That is, to solutions closer to the actual crosstabs. The analysis includes the study of whether more accurate results are attained by dealing with integer figures in all the steps or by integer-adjusting decimal solutions.

Assessing accuracy of integer and decimal solutions

Table 4 summarizes the accuracies, measured though the statistic *EI*, of the solutions obtained after applying the three alternatives detailed in Section 2 to the data described in subsection 3.1. The table displays the averages of *EI* measures grouping them in four different panels, each of them corresponding to a base algorithm. The average accuracies are presented after clustering the elections by country and year, as all the elections (datasets) belonging to the same year and country were held under the same political context. To make comparisons easier, the table also presents the *EI* averages after pooling together the 565 elections (NZ + SCO).

Several issues emerged when analysing the results in Table 4. First, there are almost no differences between the decimal and (adjusted) integer solutions when the lphom model is employed. This result was expected given that the values v_{jk} tend to be large and, therefore, the distances (in relative terms) between decimal and integer(-adjusted) estimates very small; in the vast majority of the cases just decimal places. Second, as expected, the integer-adjusted and integer solutions when lphom and tslphom are employed are (almost) equal. Indeed, they must be theoretically equal. In practice, however, some tiny differences could be recorded sporadically due to computational issues associated with computer rounding (Weihs et al., 2014).[11]

The above two results are fairly straightforward. The most striking findings appear when the decimal and (adjusted) integer solutions of lclphom and nslphom are compared. Third, as a rule (on average), integer-adjusting the decimal solutions tends to slightly deteriorate them, although they are, as expected, very close. This can be observed in the *EI* averages of adjusted and decimal solutions and can be confirmed looking in detail at their related p_{jk} and v_{jk} estimates. Fourth, contrary to what was expected, the integer solutions are on average (see Table 4) and in the majority of the cases (see Figure 2) clearly less accurate than the decimal ones. Pooling all the elections, their average errors are 3.3% larger in the case of lclphom solutions and even worse, 7.7%, in the case of the nslphom. Despite the internal cells being integers, and consequently the set of feasible solutions discrete, working in a continuous simplex space produces better solutions in the majority of the elections. In light of the results attained, it seems preferable to present the decimal solutions when the focus is on estimating the transfer probabilities/fractions $[p_{jk}]$ (and $[p_{ijk}]$) and to show the integer-adjusted solutions when the aim lies in forecasting the matrices of cross-

Table 4. Averages of ***EI***. errors by group of elections

Country-Year	NZ + SCO	NZ-2002	NZ-2005	SCO-2007	NZ-2008	NZ-2011	NZ-2014	NZ-2017	NZ-2020
# of elections	N = 565	N = 69	N = 69	N = 73	N = 70	N = 70	N = 71	N = 71	N = 72
				lphom-based solutions					
lphom-decimal	13.28	16.88	12.14	12.92	12.22	12.99	12.95	12.20	14.02
lphom-adjusted	13.28	16.88	12.14	12.92	12.22	12.99	12.95	12.20	14.02
lphom-integer	13.28	16.88	12.14	12.92	12.22	12.99	12.95	12.20	14.02
				tslphom-based solutions					
tslphom-decimal	11.78	14.81	10.91	11.00	10.89	11.50	11.65	10.92	12.59
tslphom-adjusted	11.83	14.89	10.96	11.02	10.94	11.58	11.71	10.97	12.65
tslphom-integer	11.83	14.89	10.96	11.02	10.94	11.58	11.71	10.97	12.65
				lclphom-based solutions					
lclphom-decimal	10.02	12.89	9.39	8.86	9.05	9.54	9.57	8.84	9.85
lclphom-adjusted	10.11	12.96	9.47	8.92	9.12	9.61	9.63	8.94	9.94
lclphom-integer	10.35	13.56	10.12	9.14	9.63	10.43	10.43	9.90	10.78
				nslphom-based solutions					
nslphom-decimal	9.74	12.97	9.69	8.91	9.37	9.70	9.89	9.13	10.60
nslphom-adjusted	9.81	13.07	9.77	8.98	9.44	9.81	9.98	9.24	10.70
nslphom-integer	10.49	13.29	9.91	9.15	9.66	10.10	10.21	9.55	11.00

Source: Own elaboration. The decimal solutions have been obtained after applying, with default options, the functions lphom, tslphom, lclphom and nslphom of the R package lphom, version 0.3.1–1 (Pavía and Romero, 2022), to the data from the New Zealand electoral commission and the Scotland Electoral Office available in the R package ei.Datasets (Pavía, 2022). The adjusted solutions have been obtained after integer-adjusting the decimal solutions. The integer solutions have been obtained using the ILL procedure described in Section 2. The method adjust2integers of lphom and the argument 'integers = TRUE' have been used to, respectively, attain the adjusted and integer solutions. Please see the reproducible code for the details. The smaller the number, the greater the accuracy.

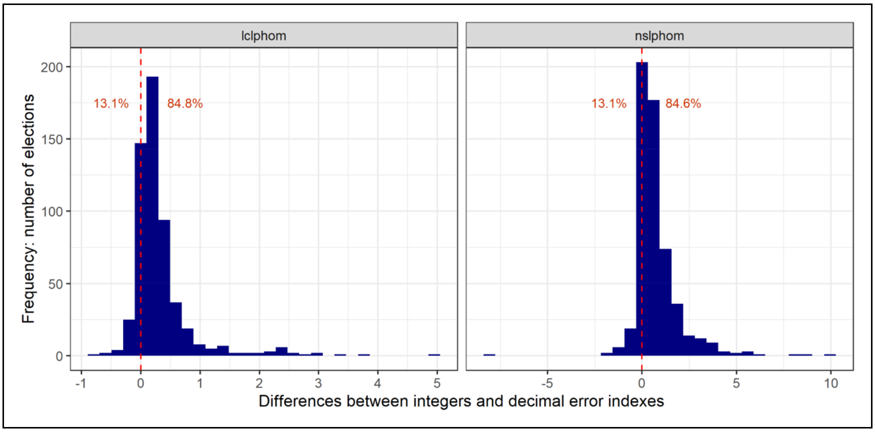

Figure 2. Histograms of the differences between the error indexes of integer and decimal solutions ($EI_{integer} - EI_{decimal}$) attained with lclphom (left panel) and nslphom (right panel). Negative numbers indicate elections for which the integer solutions are more accurate than the decimal ones. The percentages indicate the relative number of times that each of the two variants (integer vs decimal) generates the most accurate solutions

distributions of votes, $[v_{ijk}]$. Furthermore, given the almost equality of decimal and integer solutions when only the global matrix is adjusted (as in the case of the lphom solutions), an additional recommendation would be to adjust only the global matrix when the goal is uniquely focused on $[v_{jk}]$.

Exploring when integer solutions are more accurate than decimal solutions

Although decimal solutions are more accurate than integer solutions in most of the elections, this is not always the case, as can be seen in Figure 2, where histograms of the distributions of the *EI* differences are presented for lclphom and nslphom. In addition to the magnitude of the differences, Figure 2 also shows the percentage of times (elections) that each of the two variants generates more accurate solutions. Integer solutions are more accurate than decimal solutions 13.1% of the time, whatever the base algorithm (lclphom or nslphom),[12] with 2.1% of draws in the case of lclphom and 2.3% in the case of nslphom; the main difference being that there are larger differences between the nslphom solutions. Its standard deviation is more than two times larger than that corresponding to lclphom solutions and the ranges of its positive and negative differences are clearly longer.

The above results show that decimal solutions are, as a rule, more accurate than integer ones, although sometimes they are improved by integer solutions; in our study, in a sixth of the times. The issue is why does this happen and when is this going to happen? Given that in real applications we cannot know if this is occurring in a particular instance because actual crosstabs are routinely unknown, the question is whether we could obtain from the available information, $\boldsymbol{X}$ and $\boldsymbol{Y}$, some clues about when the integer solutions would perform well.

In order to seek answers to these questions, linear and logit regression models have been used to try to identify what factors impact, on the one hand, on the differences seen in Figure 2 and, on the other hand, on a dummy variable that summarises their signs.[13]

As explanatory variables, we have considered several features previously identified as determinants of the quality of ecological inference estimates (e.g., King, 1997; Klima et al., 2016; Park et al. 2014; Pavía and Romero, 2023; Plescia and De Sio, 2018). In particular, the list of potential dependent variables to be included in the models are: I (as an indicator of the quantity of information), RC (as indicator of the complexity of the problem), R (the number of relevant parties), C (the number of relevant candidates), R / C (as an indicator of the degree of asymmetry of the transfer matrix), V (the total number of voters), density (the average number of voters by unit, V / I), concentration (the Gini index of the distribution of voters across units, as an indicator of differences of size across units), skewness (the skewness of distribution of voters across units, as an indicator of the (a)symmetry among the sizes of the units), the heterogeneity index (as an indicator of the degree of non-compliance of the homogeneity hypothesis), the standardized χ^2 -Pearson statistic (as an indicator of the degree of dependence between the row and column categories),[14] the compositional total variance (Pawlowsky-Glahn et al., 2015) of the marginal row distributions across units (as an indicator of variability of party support across units), the column compositional total variance (as an indicator of variability of candidacy support across units), and the standard deviations of the distribution of percentages of votes by parties and candidates (as indicators of the degree of vote concentration/variability among parties).

Given the large number of regressors and their correlation structure, the stepwise procedure for variable selection based on the AIC criterion in its backward-forward variant (Hastie et al., 2009) has been utilised as fitting technique, with all the regressors standardized.[15] The final selected models are quite complex, not containing the same relevant variables in all the cases.[16] Although it cannot be ruled out that the factors identified as significant are case-dependent, several regularities emerge. On the one hand, the amount of information, I, is the variable with most impact on the differences. The models indicate that the larger I is, the more accurate are (relatively) the decimal solutions. The significance of this variable, however, vanishes when the dependent variable of the model is the sign of the differences. In this case, the variables with higher impacts are R, RC and R / C. On the one hand, the larger R is, the more probable it is that integer solutions improve decimal ones. On the other hand, larger values of RC and R / C favour decimal solutions.

Finally, regarding computation burden, the higher processing times corresponded to the integer-adjusted solutions, being in any case all of them assumable. For instance, pooling all the elections, the average time in secs was 40.2 for nslphom and 31.3 for lclphom. These figures contrast with the times of 1.8 and 4.8 s required by nslphom and 6.8 and 6.7 s by lclphom for returning their decimal and integer solutions, respectively.[17]

Discussion and concluding remarks

In ecological inference, the normative problem refers to the forecasting of the interior values of a contingency table when only the aggregate margins are known. Despite it being a widely disputed methodology (e.g., Cho and Manski, 2008; Freedman, 1999; Forcina and Pellegrino, 2019; Glynn and Wakefield, 2010; Robinson, 1950), the ecological inference procedures are routinely applied in all those settings where aggregate data are abundant and trustworthy, and individual-level data dubious or hard (impossible) to collect. Political science and sociology are the disciplines where they are more frequently used, mainly in the context of elections.

Ecological inference algorithms have been used to estimate floating voters (Russo, 2014), to reveal split-ticket voting behaviours (Park et al., 2014) or to disclose social or gender voting patterns (Hudson et al., 2010; O'Loughlin, 2000). Furthermore, they are also used to guess vote transfer matrices between elections (Vizcaino and Pavía, 2022), to analyse party loyalty and electoral volatility (Sandoval and Ojeda, 2023) and to disentangle racial voting patterns (Barreto et al., 2022). Indeed, "*the United States Supreme Courts directly recommended ecological inference analysis as the main statistical method to estimate voting preference by racial group (e.g., Thornburg v. Gingles 478 U.S. 30, 1986)*" (Collingwood et al., 2016, p. 92) in voting rights litigations.

However, despite the interior values of contingency tables being whole numbers, almost all the ecological inference models have been devised to infer proportions/fractions/probabilities. Within the ecological inference linear programming framework, this paper proposes an extension of the local method proposed in Pavía and Romero (2024a) to yield integer solutions and uses it to introduce a new variant in the lphom-family algorithms. The interested analyst can easily use this new variant of the algorithms, which is available in the basic functions (lphom, tslphom, nslphom and lclphom) of the lphom package (Pavía and Romero, 2022), by setting their argument *integers* equal to *TRUE*. Moreover, the integer-adjusting approach is also available through the method adjust2integers, devised to transform lphom-class final decimal solutions into integer solutions.

This new approach was proposed with the expectation that by reducing the search space for the solutions from a continuous simplex space to a discrete simplex space, to which indeed the actual outcomes belong, more accurate solutions would be obtained. The 565 elections from New Zealand and Scotland available in the R package ei.Datasets (Pavía, 2022), where the actual cross-distributions of votes of each election are known, have been used to assess it. The assessment is made by comparing the accuracy of the default (decimal) solutions with two new alternative solutions obtained after (i) integer-adjusting the decimal solutions (i.e., after running adjust2integers to the decimal solutions) and (ii) applying integer linear programming in all the steps (i.e., by using the option 'integers = TRUE'). The evaluation is made considering the solutions achieved with four lphom-based algorithms (lphom, tslphom, nslphom and lclphom).

The results attained suggest that, at least for the datasets analysed, the use of the integer approach in all the steps does not lead to more accurate global solutions. For instance, the more accurate solutions reached using nslphom with default options show, on average, 7.7% less error than its integer counterparts. In light of these findings, it could be affirmed that, contrary to initial expectations, estimates are not more accurate on average when obtained working in a simplex discrete space and that, therefore, the recommendations would be: (i) to compute decimal solutions when the goal is to estimate global (and local) transfer probabilities/fractions, (ii) to integer-adjust decimal solutions when the focus lies in forecasting the local (and global) cross-distributions of votes and (iii) to only integer-adjust the decimal global solution when the aim is simply focused on forecasting the global crosstabs of votes.

Although displaying proportions is enough in the immense majority of ecological inference problems and for analytical purposes, sometimes it may be more convenient to show the transfer matrices in whole numbers, such as when the estimates are presented to the general public in a unit by unit fashion. For example, when they are offered in an interactive map which citizens can use to select the census section (unit) they are interested in so as to see the results and the transition matrix.

Acknowledgements

The author wishes to thank M. Hodkinson for revising the English of the paper and two anonymous reviewers and the editor, Ahmed Fouad El Haddad, for valuable comments and suggestions.

Availability of data and material

The data used in this research is publicly available on the R-package ei.Datasets (version 0.0.1–1) accessible on CRAN.

Code availability

The reproducible ad-hoc R-code employed, based on functions included in the R-packages eiPack (version 0.2–1) and lphom (version 0.3.1–1) is available in the supplementary files.

Declaration of conflicting interests

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

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has been supported by Consellería de Educación, Universidades y Empleo, Generalitat Valenciana [grant numbers AICO/2021/257, CIAICO/2023-GVRTE/2023/4572860] and by the Ministerio de Ciencia e Innovación [grant number PID2021-128228NB-I00].

Supplementary material

Supplementary material is available in the online version of this article.

Notes

1. In these cases, it is recommendable to combine both sources of information (polls and aggregate data), through the so-called hybrid models (Greiner and Quinn, 2010; Klima et al., 2019; Pavía, 2024b), as this leads to more accurate estimates.
2. On the one hand, the ei.MD.bayes function from the eiPack package implements (a version of) the hierarchical model suggested in Rosen et al. (2001). On the other hand, the nslphom function of the lphom package runs the algorithm of the same name proposed in Pavía and Romero (2024a). Although the lphom package has several other (basic) functions for solving the

ecological inference problem (lphom, tslphom and lclphom), nslphom has been identified as the most accurate (Romero and Pavía, 2021; Pavía and Romero, 2022a, 2023; Pavía, 2024b).

3 An exception is the model devised by Greiner and Quinn (2009), which proposes a RxC procedure that models counts and, according to the authors, "respects the bounds deterministically" (see also, Greiner and Quinn, 2010). It should be noted, nevertheless, that Greiner and Quinn's model also relies on proportion estimates in its intermediate steps and that its implementation in the software R, RxCEcolInf (Greiner et al., 2021), does not generate estimates congruent with the observed margins.

4 For instance, after estimating with ei.MD.bayes the 3×3 precinct tables of political registration by race of the dataset senc (also available in the R-package eiPack), one can obtain, comparing the sums by columns of actual and estimated counts, (i) discrepancies of ~150 in absolute terms and ~0.04% in relative terms for the (mean of the posterior) global solution, (ii) maximum discrepancies of ~1623 and ~0.44% among the samples of the global solutions or (iii) maximum discrepancies of ~271 and ~21.52% among the samples of precincts Harnett-3 (county of New Hanover) and Canetuck (county of Pender), respectively. These results have been achieved using the default options of ei.MD.bayes and after setting the random seed in 123. Details are available in the reproducible code "Example of inconsistency of ei.MD.Bayes from eiPack" of the supplementary material.

5 These properties are not exclusive to the ecological inference models proposed within the mathematical programming framework. Congruence with decimal estimates are also found in the models available from information theory and latent factor approaches, as well as in some frequentist statistical models.

6 Commonly, each table of the set of related tables refers to a unit of a non-overlapping geographical partition of the area of interest. We call each individual table (unit) local and the aggregation (union) of all the local tables (units) global.

7 For a discussion of how to handle more complex scenarios see, for instance, Pavía (2023).

8 Note that $v_{jk} = p_{jk} \sum_i x_{ij}$, $v_{ijk} = p_{ijk}\, x_{ij}$ and $p_{jk} = \sum_i x_{ij} p_{ijk} / \sum_i x_{ij}$, $\forall i, j, k$.

9 In this characterization the lphom-estimate is the only transfer matrix that, in addition to satisfying all the aggregate row and column constraints, minimizes the sum of expected count discrepancies across units, measured in L_1-norm, between the observed and expected unit column totals when the estimated matrix is applied to the observed unit row totals.

10 In this case, to be fully coherent, the iterative process is initiated using the integer version of the lphom model, defined by the set of equations (A14) to (A18) and (A19) to (A24) of the Technical Appendix.

11 The only difference between the integer-adjusted and integer solutions attained with lphom and tslphom is that, in the former, the solutions are obtained in two steps, using two different functions, with the second function (adjust2integers) taking as input the decimal output previously returned by the first function, whereas, in the latter, all the computations are performed within a sole function, with all the calculations being therefore performed using all the available computer precision.

12 Although both figures coincide, they only intersect in a third of the cases.

13 When modelling the nslphom differences, the outlier corresponding to the minimum *EI* has been excluded from the fitting.

14 Although the heterogeneity index (Romero et al., 2020) and the standardized χ^2-Pearson statistic (Pavía, 2022) cannot be computed in real applications because they depend on the unknown actual global cross-classification distribution, they can be estimated from the achieved solution.

15 The standardization eases the interpretation of the estimated coefficients of the fitted models, making it possible to directly assess the relative importance of each variable.

16 All the details regarding (i) the computation of the regressors and (ii) the fitting, specification and estimation of the models are available in the reproducible code of the supplementary material.
17 These times have been recorded performing the computations on a laptop with a CPU processor Intel® Core™ i7-6820K (8 cores) 2.70GHz and 64GB of RAM.
18 It is computationally unfeasible to try to solve this problem in just one step, using ILP directly, due to the NP-hard nature of the problem (Kleinberg, 2006). Probably, this feature of the problem explains why the ecological inference problem has not been tackled within a discrete space from the maximum likelihood approach.

References

Antweiler W (2007) Estimating voter migration in Canada using generalized maximum entropy. *Electoral Studies* 26: 756-771.

Audemard J (2024) Understanding vote transfers in two-round elections without resorting to declared data. The contribution of ecological inference, consolidated with factual information from a case study of the 2014 municipal elections in Montpellier. *Bulletin de Méthodologie Sociologique* 161–162: 13-34.

Barreto M, Collingwood L, Garcia-Rios S, et al. (2022) Estimating candidate support in Voting Rights Act cases: Comparing iterative EI and EI-RC methods. *Sociological Methods & Research* 51: 271-304.

Bernardini-Papalia R and Fernández-Vázquez E (2020) Entropy-based solutions for ecological inference problems: A composite estimator. *Entropy* 22: 81.

Brown C (1982) The Nazi vote: A national ecological study. *The American Political Science Review* 76: 285-302.

Brown PJ and Payne CD (1986) Aggregate data, ecological regression and voting transition. *Journal of the American Statistical Association* 81: 452-460.

Cho WKT and Manski CF (2008) *Cross level/ecological inference*. In: Box-Steffensmeier JM, Brady HE and Collier D (eds) *The Oxford Handbook of Political Methodology*. Oxford: University Press, 547-569.

Collingwood L, Oskooii K, Garcia-Rios S, et al. (2016) Eicompare: Comparing ecological inference estimates across EI and EI:R×C. *The R Journal* 8: 92-101.

Corominas A, Lusa A and Valvet MD (2015) Computing voter transitions: The elections for the Catalan Parliament, from 2010 to 2012. *Journal of Industrial Engineering and Management* 8: 122-136.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39: 1-38.

El Deeb O (2023) Entropic spatial auto-correlation of voter uncertainty and voter transitions in parliamentary elections. *Physica A: Statistical Mechanics and its Applications* 617: 128675.

Forcina A, Gnaldi M and Bracalente B (2012) A revised Brown and Payne model of voting behaviour applied to the 2009 elections in Italy. *Statistical Methods & Applications* 21: 109-119.

Forcina A and Pellegrino D (2019) Estimation of voter transitions and the ecological fallacy. *Quality & Quantity* 53: 1859-1874.

Freedman DA, Klein SP, Sacks J, et al. (1991) Ecological regression and voting rights. *Evaluation Review* 15: 673-711.

Freedman DA, Ostland M, Roberts MR, et al. (1999) Response to King's comment. *Journal of the American Statistical Association* 94: 355-357.

Füle E (1994) Estimating voter transitions by ecological regression. *Electoral Studies* 13: 313-330.

Glynn A and Wakefield J (2010) Ecological inference in the social sciences. *Statistical Methodology* 7: 307-322.

Gnaldi M, Tomaselli V and Forcina A (2018) Ecological fallacy and covariates: New insights based on multilevel modelling of individual data. *International Statistical Review* 86: 119-135.

Goodman LA (1953) Ecological regressions and the behaviour of individuals. *American Sociological Review* 18: 663-664.

Goodman LA (1959) Some alternatives to ecological correlation. *American Journal of Sociology* 64: 610-625.

Gosnell HF and Gill NN (1935) An analysis of the 1932 presidential vote in Chicago. *The American Political Science Review* 29: 967-984.

Greiner DJ (2007) Ecological inference in voting rights act disputes: Where are we now, and where do we want to be? *Jurimetrics* 47: 115-167.

Greiner DJ, Baines P and Quinn KM (2021) RxCEcolInf: RxC Ecological Inference with Optional Incorporation of Survey Information (R package version 0.1–5) [Computer software]. https://cran.r-project.org/package=RxCEcolInf.

Greiner DJ and Quinn KM (2009) Rxc ecological inference: Bounds, correlations, flexibility and transparency of assumptions. *Journal of the Royal Statistical Society, Series A* 172: 67-81.

Greiner DJ and Quinn KM (2010) Exit polling and racial bloc voting: Combining individual-level and RxC ecological data. *The Annals of Applied Statistics* 4: 1774-1796.

Greselin F and Ingrassia S (2015) Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers. *Statistical Computing* 25: 215-226.

Hastie T, Tibshirani R and Friedman J (2009) *The Elements of Statistical Learning*. New York: Springer.

Hawkes A (1969) An approach to the analysis of electoral swing. *Journal of the Royal Statistical Society, Series A* 132: 68-79.

Hudson IL, Moore L, Beh EJ, et al. (2010) Ecological inference techniques: An empirical evaluation using data describing gender and voter turnout at New Zealand elections, 1893–1919. *Journal of the Royal Statistical Society, Series A* 173: 185-213.

Imai K, Lu Y and Strauss A (2008) Bayesian And likelihood inference for 2 (2 ecological tables: An incomplete data approach. *Political Analysis* 16: 41-69.

Irwin GA and Meeter DA (1969) Building voter transition models from aggregate data. *Midwest Journal of Political Science* 13: 545-566.

Jiang W, King G, Schmaltz A, et al. (2020) Ecological regression with partial identification. *Political Analysis* 28: 65-86.

Johnston R and Pattie C (2000) Ecological inference and entropy-maximizing: An alternative estimation procedure for split-ticket voting. *Political Analysis* 8: 333-345.

Johnston RJ and Hay AM (1983) Voter transition probability estimates: An entropy-maximizing approach. *European Journal of Political Research* 11: 93-98.

Judge GG, Miller DJ and Cho WK (2003) An Information Theoretic Approach to Ecological Estimation and Inference. UC Berkeley: Department of Agricultural and Resource Economics.

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

King G, Rosen O and Tanner MA (1999) Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research* 28: 61-90.

King G, Rosen O and Tanner MA (eds) (2004) *Ecological inference: New methodological strategies*. New York, NY: Cambridge University Press.

King G, Rosen O, Tanner M, et al. (2008) Ordinary economic voting behavior in the extraordinary election of Adolf Hitler. *The Journal of Economic History* 68: 951-996.

Kleinberg J (2006) *Algorithm Design*. Pearson/Addison-Wesley.

Klima A, Schlesinger T, Thurner PW, et al. (2019) Combining aggregate data and exit polls for the estimation of voter transitions. *Sociological Methods & Research* 48: 296-325.

Klima A, Thurner PW, Molnar C, et al. (2016) Estimation of voter transitions based on ecological inference: An empirical assessment of different approaches. *AStA - Advances in Statistical Analysis* 100: 133-159.

Lau O, Moore ORT and Kellermann M (2020) *eiPack: Ecological Inference and Higher-Dimension Data Management* (R package Version 0.2–1) [Computer software]. https://cran.r-project.org/package=eiPack.

Manski CF (2007) *Identification for Prediction and Decision*. Harvard University Press.

McCarthy C and Terence MR (1977) Estimates of voter transition probabilities from the British General Elections of 1974. *Journal of the Royal Statistical Society, Series A* 140: 78-85.

Ogburn WF and Goltra I (1919) How women vote. *Political Science Quarterly* 34: 413-433.

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* 63: 168-192.

Park W, Hanmer MJ and Biggers DR (2014) Ecological inference under unfavorable conditions: Straight and split-ticket voting in diverse settings and small samples. *Electoral Studies* 36: 192-203.

Pavía JM (2022) Ei.datasets: Real datasets for assessing ecological inference algorithms. *Social Science Computer Review* 40: 247-260.

Pavía JM (2023) Adjustment of initial estimates of voter transition probabilities to guarantee consistency and completeness. *SN Social Sciences* 3: 75.

Pavía JM (2024a) A local convergent ecological inference algorithm for RxC tables. *The Journal of Mathematical Sociology*. forthcoming.

Pavía JM (2024b) Integrating survey and aggregate data in ecological inference estimation of RxC tables.

Pavía JM and Romero R (2024a) Improving estimates accuracy of voter transitions. Two new algorithms for ecological inference based on linear programming. *Sociological Methods and Research* 53: 1491-1533.

Pavía JM and Romero R (2022) *lphom: Ecological Inference by Linear Programming under Homogeneity* (R package Version 0.3.1–1) [Computer software]. https://cran.r-project.org/package=lphom

Pavía JM and 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: 151-186.

Pavía JM and Romero R (2024b) Symmetry estimating R×C vote transfer matrices from aggregate data. *Journal of the Royal Statistical Society, Series A*. online available.

Pavía JM and Thomsen SR (2024) ecolRxC: Ecological inference estimation of R×C tables using latent structure approaches, *Political Science Research and Methods*. forthcoming.

Pawlowsky-Glahn V, Egozcue JJ and Tolosana-Delgado R (2015) *Modeling and Analysis of Compositional Data*. Chichester: John Wiley & Sons, Ltd.

Penadés A and Pavía (2024) An ecological inference approach to the origins of proportional representation. *Social Science Information* 63: 168-192.

Petropoulos F, Apileti D and Assimakopoulos V (2022) Forecasting: Theory and practice. *International Journal of Forecasting* 38: 705-871.

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

Puig X and Ginebra J (2014) A Bayesian cluster analysis of election results. *Journal of Applied Statistics* 41: 73-94.

Robinson WS (1950) Ecological correlations and the behavior of individuals. *American Sociological Review* 15: 351-357.
Romero R and Pavía JM (2021) Estimating vote party entries and exits by ecological inference. Mathematical programming versus Bayesian statistics. *BEIO* 37: 85-97.
Romero R, Pavía JM, Martín J, et al. (2020) Assessing uncertainty of voter transitions estimated from aggregated data: Application to the 2017 French presidential election. *Journal of Applied Statistics* 47: 2711-2736.
Rosen O, Jiang W, King G, et al. (2001) Bayesian and frequentist inference for ecological inference: The RxC case. *Statistica Neerlandica* 55(2): 134-156.
Russo L (2014) Estimating floating voters: A comparison between the ecological inference and the survey methods. *Quality & Quantity* 48: 1667-1683.
Sandoval P and Ojeda S (2023) Estimation of electoral volatility parameters employing ecological inference methods. *Quality & Quantity* 57: 405-426.
Thomsen SR (1987) *Danish Elections, 1920-1979: A Logit Approach to Ecological Analysis and Inference*. Aarhus: Politica.
Tziafetas G (1986) Estimation of the voter transition matrix. *Optimization* 17: 275-279.
Upton GJG (1978) A note on the estimation of voter transition probabilities. *Journal of the Royal Statistical Society, Series A* 141: 507-512.
Vangrevelinghe G (1961) Étude statistique comparée des résultats des référendums de 1958 et 1961. *Revue de Statistique Applique* 9: 83-100.
Vizcaino A and Pavía JM (2022) New parties and matryoshka scissions in Spain. The case of Podemos and Más Madrid. *Social Sciences & Humanities Open* 6: 100307.
Wakefield J (2004) Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society, Series A* 167: 385-445.
Weihs C, Mersmann O and Ligges U (2014) *Foundations of Statistical Algorithms. With References to R Packages*. Boca Raton, FL: CRC Press.

Technical appendix: the ILL model

The ILL model is an extension of the LL model, which is a two-step procedure that, given a global matrix $[p^G_{jk}]$ of proportions, returns as output I local matrices $[p_{ijk}]$ $(i = 1, \ldots, I)$ congruent with $\boldsymbol{X}$ and $\boldsymbol{Y}$, with each of them being as close as possible to $[p^G_{jk}]$. This process, however, does not guarantee that integer count estimates will eventually be reached. This paper proposes the inclusion of a third step in the LL procedure to ensure this. A new step in which the LL solutions are refined using integer linear programming (ILP) to guarantee integer counts.[18] This allows expansion of the properties of the methods recently suggested in Pavía and Romero (2024a) and Pavía (2024a) by allowing them the possibility of delivering integer solutions.

Mathematically, the three-step linear program algorithm detailed in equations (A1)-(A13) describes the ILL (integer_lphom_local) procedure for a generic unit i. The first step delimits the sets of p_{ijk} values and of non-negative auxiliary quantities ε^+_{ijk} and ε^-_{ijk} ($j = 1, \ldots, R$; $k = 1, \ldots, C$) that, satisfying the constraints (A1)-(A4), minimize (A5). The second step returns as solution the p_{ijk} proportions, and the non-negative auxiliary quantities δ^+_{ijk} and δ^-_{ijk}, that minimize (A8), verifying the constraints (A1)-(A4) and (A6)-(A7). These two steps make up the LL algorithm. Finally, the third step finds, conditioned on the proportions p_{ijk} attained in step two, the v_{ijk} integers that minimize (A13) subject to the constraints (A9)-(A12), where λ^+_{ijk} and λ^-_{ijk} are auxiliary non-negative variates.

$$p_{ijk} \geq 0 \quad for \quad j = 1, \ldots, R \quad k = 1, \ldots, C \tag{A1}$$

$$\sum_{k=1}^{K} p_{ijk} = 1 \quad for \ j = 1, \ldots, J \tag{A2}$$

$$\sum_{j=1}^{J} x_{ij} p_{ijk} = y_{ik} \quad for \ k = 1, \ldots, C \tag{A3}$$

$$(p_{jk}^{G} - p_{ijk}) x_{ij} = \varepsilon_{ijk}^{+} - \varepsilon_{ijk}^{-} \quad for\, j = 1, \ldots, R \quad k = 1, \ldots, C \tag{A4}$$

$$minimize \sum_{j=1}^{J} \sum_{k=1}^{K} (\varepsilon_{ijk}^{+} + \varepsilon_{ijk}^{-}) \tag{A5}$$

$$W = \sum_{j=1}^{J} \sum_{k=1}^{K} (\varepsilon_{ijk}^{+} + \varepsilon_{ijk}^{-}) \tag{A6}$$

$$p_{ijk} = p_{jk}^{G} + \delta_{ijk}^{+} - \delta_{ijk}^{-} \quad for \ j = 1, \ldots, R \quad k = 1, \ldots, C \tag{A7}$$

$$minimize \ \sum_{j=1}^{J} \sum_{k=1}^{K} (\delta_{ijk}^{+} + \delta_{ijk}^{-}) \tag{A8}$$

$$v_{ijk} \in \mathbb{N} \quad for \ j = 1, \ldots, R \quad k = 1, \ldots, C \tag{A9}$$

$$\sum_{k=1}^{K} v_{ijk} = x_{ij} \quad for \ j = 1, \ldots, R \tag{A10}$$

$$\sum_{j=1}^{J} v_{ijk} = y_{ik} \quad for \ k = 1, \ldots, C \tag{A11}$$

$$v_{ijk} = p_{ijk} x_{ij} + \lambda_{ijk}^{+} - \lambda_{ijk}^{-} \quad for \ j = 1, \ldots, R \quad k = 1, \ldots, C \tag{A12}$$

$$minimize \ \sum_{j=1}^{J} \sum_{k=1}^{K} (\lambda_{ijk}^{+} + \lambda_{ijk}^{-}) \tag{A13}$$

The solution of the (A1)-(A13) system depends on $[p_{jk}^{G}]$. The lphom-family algorithms are characterized by their use of the matrix, $[\widehat{p}_{jk}^{(0)}]$, solution of the basic lphom linear system (Romero et al., 2020), defined by equations (A14)-(A18), as initial global matrix. In the (A14)-(A18) system, the unknowns are the p_{jk} unobserved global proportions as well as the (non-negative) auxiliary quantities e_{ik}^{+} and e_{ik}^{-}.

$$p_{jk} \geq 0 \quad for \ j = 1, \ldots, R \quad k = 1, \ldots, C \tag{A14}$$

$$\sum_{k=1}^{K} p_{jk} = 1 \quad for \;\; j = 1, \ldots, R \tag{A15}$$

$$\sum_{j=1}^{J}\left(\sum_{i=1}^{I} x_{ij}\right) p_{jk} = \left(\sum_{i=1}^{I} y_{ik}\right) \quad for \;\; k = 1, \ldots, C \tag{A16}$$

$$y_{ik} = \sum_{j=1}^{J} x_{ij}\, p_{jk} + e_{ik}^{+} - e_{ik}^{-} \quad for \;\; k = 1, \ldots, C \quad i = 1, \ldots, I \tag{A17}$$

$$mimimize \;\; \sum_{i=1}^{I}\sum_{k=1}^{K}(e_{ik}^{+} + e_{ik}^{-}) \tag{A18}$$

In particular, with LL, lphom, tslphom, nslphom and lclphom operate as follows. First, lphom only generates a global solution: the matrix $[\widehat{p}_{jk}^{(0)}]$ that is obtained minimizing (A18) subject to (A14)-(A17). Second, tslphom returns (i) as local solutions the matrices $[\widehat{p}_{ijk}^{(t)}]$ (for $i = 1, \ldots, I$ and $t = 1$) that are attained after solving the system (A1)-(A8) for each unit using $[p_{jk}^{G}] = [\widehat{p}_{jk}^{(t-1)}]$ as global matrix and (ii) as global solution the matrix $[\widehat{p}_{jk}^{(t)}] = \sum_{i=i}^{I} [\widehat{p}_{ijk}^{(t)}]\, x_{ij} \,/\, \sum_{i=i}^{I} x_{ij}$. Third, after applying iteratively n times (for $t = 1, \ldots, n$) the tslphom process with temporary global matrix $[\widehat{p}_{jk}^{(t-1)}]$, nslphom returns as local and global solutions the matrices obtained in the iteration t^* (with $1 \leq t^* \leq n$) that minimizes (A19). Fourth, lclphom is a bit more complex. For $t = 1$, it starts with the tslphom solution and defines for each unit i a temporary solution $[\widehat{p}_{ijk}] = [\widehat{p}_{ijk}^{(t)}]$ and a temporary error $e_i = \sum_{k=1}^{K} \left| y_{ik} - \sum_{j=1}^{J} x_{ij}\, \hat{p}_{jk}^{(t)} \right|$. Afterwards, for $t > 1$, the tslphom process is applied repeatedly updating e_i and $[\widehat{p}_{ijk}]$ for those units for which a smaller error is observed and a new $[\widehat{p}_{jk}^{(t)}]$ computed using the current temporary local solutions available. This process, whose convergence is guaranteed (Pavía, 2024a), is repeated until no error is reduced. The last step of lclphom consists in applying the tslphom procedure once again using as $[p_{jk}^{G}]$ the global matrix attained after convergence. The local and global matrices achieved in this last step comprise the lclphom-solution.

$$\widehat{\mathrm{HET}}^{(t)} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} |\hat{p}_{ijk}^{(t)} x_{ij} - \hat{p}_{jk}^{(t)} x_{ij}|}{2\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}} \tag{A19}$$

Finally, the lphom model could be also modified, in the same way as the lphom_local model, to also yield integer solutions. To do that, we propose to solve the ILP problem defined by equations (A20)-(A24), which taking as input the observed matrices of counts and the matrix $[\widehat{p}_{jk}^{(0)}]$ generated by the lphom_model produces a matrix $[\check{p}_{jk}^{(0)}]$ of row-standardized proportions that generates integer counts when multiplied by observed votes. That is, $\breve{p}_{jk}^{(0)} = \breve{v}_{jk} / \sum_i x_{ij}$, $\forall j, k$, with $\breve{v}_{jk}$ being the integer numbers obtained solving the ILP defined by equations (A20)-(A24).

$$v_{jk} \in \mathbb{N} \quad for\ \ j = 1, \ldots, R \quad k = 1, \ldots, C \tag{A20}$$

$$\sum_{k=1}^{K} v_{jk} = \sum_{i=1}^{I} x_{ij} \quad for \quad j = 1, \ldots, R \tag{A21}$$

$$\sum_{j=1}^{J} v_{jk} = \sum_{i=1}^{I} y_{ik} \quad for \quad k = 1, \ldots, C \tag{A22}$$

$$v_{jk} = \widehat{p}_{jk}^{(0)} \sum_{i=1}^{I} x_{ij} + \xi_{jk}^{+} - \xi_{jk}^{-} \quad for\ \ j = 1, \ldots, R \quad k = 1, \ldots, C \tag{A23}$$

$$minimize \sum_{j=1}^{J} \sum_{k=1}^{K} (\xi_{jk}^{+} + \xi_{jk}^{-}) \tag{A24}$$