Political Science Research and Methods (2024), page 1 of 19
doi:10.1017/psrm.2024.57

ORIGINAL ARTICLE

ecolRxC: Ecological inference estimation of R × C tables using latent structure approaches

Jose M. Pavía[1] and Søren Risbjerg Thomsen[2]

[1]GIPEyOP, Area of Quantitative Methods, Universitat de Valencia, Valencia, Spain and [2]Department of Political Science, University of Aarhus, Aarhus, Denmark
Corresponding author: Jose M. Pavía; Email: pavia@uv.es

(Received 24 December 2023; revised 30 March 2024; accepted 26 May 2024)

Abstract
Ecological inference is a statistical technique used to infer individual behavior from aggregate data. A particularly relevant instance of ecological inference involves the estimation of the inner cells of a set of R × C related contingency tables when only their aggregate margins are known. This problem spans multiple disciplines, including quantitative history, epidemiology, political science, marketing, and sociology. This paper proposes new models for solving the problem using the latent structure theory, and presents the `ecolRxC` package, an `R` implementation of this methodology. This article exemplifies, explains, and statistically documents the new extensions and, using real inner cell election data, shows how the new models in `ecolRxC` lead to significantly more accurate solutions than `ecol` and `VTR`, two `Stata` routines suggested within this framework. `ecolRxC` also holds its own against `ei.MD.bayes` and `nslphom`, the two algorithms currently identified in the literature as the most accurate to solve this problem. `ecolRxC` records accuracies as good as those reported for `ei.MD.bayes` and `nslphom`. Besides, from a theoretical perspective, `ecolRxC` stands up for modeling a causal theory of political behavior to build its algorithm. This distinguishes it from other procedures proposed from different frameworks (such as `ei.MD.bayes` and `nslphom`) which model expected behaviors, instead of modeling how voters make choices based on their underlying preferences as `ecolRxC` does.

Keywords: discrete choice models; ecological inference; elections; latent factors; R-package; vote transfers

1. Introduction

Ecological inference is a statistical technique used to infer individual behavior from aggregate data. This methodology has been used to gain insights into how people think, behave, and make decisions in a variety of contexts, such as voter behavior and consumer preferences (King, 1997). A particularly relevant instance of ecological inference comprises the estimation of the interior-cells of a set of R × C related contingency tables when only their aggregate margins are known in a number of subunits. This problem has attracted the interest of researchers for decades (Pavía and Romero, 2024b), chiefly within the disciplines of political science and sociology in connection with voters' electoral behavior. For example, the two-way table could be about the transfer of numbers of individual voters between parties from one election to the next when only the actual marginal results from the two elections in a number of local units (polling stations) are known.

This paper proposes new models for solving this problem using the latent structure approach and describes the `ecolRxC` package, an `R` implementation of this methodology. Compared to previous solutions within this framework (`ecol` and `VTR`), our implementation can generate

both global and unit table estimates and uncertainties, and can lead to significantly more accurate inferences. Our approach also stands up against the two algorithms (`ei.MD.bayes` and `nslphom`) currently considered in the literature as the most accurate (Klima *et al.*, 2016; Plescia and De Sio, 2018; Pavía and Romero, 2023), differentiating itself from them by how it builds its algorithms: our approach models causes of electoral behavior instead of consequences.

In elections, officially reported aggregate statistics are abundant, and usually valid, while individual-level opinion polls are not always available or reliable. Hence, ecological inference algorithms are routinely employed to approximate voter transition matrices between elections, estimate split-ticket voting behaviors, or disentangle racial voting patterns (e.g., Füle, 1994; Park *et al.*, 2014; Barreto *et al.*, 2022). Ecological inference is also used in US Courts on voting rights litigations (Greiner, 2007). The difficulty with ecological inference stems from its intrinsic indeterminacy (Manski, 2007), as there are countless internal cell count distributions compatible with the observed marginal totals. This triggers the potential emergence of the so-called ecological fallacy (Robinson, 1950), sparking much debate over the methodology (Collingwood *et al.*, 2016).

Although there are many more approaches in the literature that deal with the 2×2 problem than with the more general $R \times C$ specification, a significant number of models have also been proposed to solve the latter (e.g., Brown and Payne, 1986; Tziafetas, 1986; Thomsen, 1987; Rosen *et al.*, 2001; Andreadis and Chadjipadelis, 2009; Greiner and Quinn, 2009; Puig and Ginebra, 2014; Pavía, 2024a; Pavía and Romero, 2024a). Some of these models have been implemented in `R` (R Core Team, 2023) packages available on CRAN. Among these, the `eiPack` (Lau *et al.*, 2023) and `lphom` (Pavía and Romero, 2024c) packages stand out for having the functions (`ei.MD.bayes` and `nslphom`, respectively),[1] with the highest reported accuracies to date (Klima *et al.*, 2016; Plescia and De Sio, 2018; Pavía and Romero, 2023).[2]

However, in the same vein as the rest of the models available in `R` packages, these models base their inferences on modeling the expected consequences of voters' political behavior. Their proven practical accuracy is grounded on the particular way they operationalize the assumption of underlying similar/related[3] conditional row probability/fraction[4] distributions across tables. This is an assumption on which almost all the methods rely, founded on the empirical observation that people belonging to the same group tend to vote probabilistically alike (Pavía and Romero, 2024a), mediated by the particular context (Schmitt *et al.*, 2021).

The model proposed by Thomsen (1987), on the contrary, is grounded on a comprehensive theory for behavioral choice which can be used as an instrument for explaining voting behavior as well at the aggregate as at the individual level. Thomsen's methodology is based on a latent

[1]They run respectively a variant of the Bayesian hierarchical Multinomial-Dirichlet model proposed by Rosen *et al.* (2001) and the nslphom linear programming-based algorithm suggested in Pavía and Romero (2024a).

[2]Other methods implemented in `R` to solve the general R×C problem include the iterative version of the 2×2 model proposed by King (Choirat *et al.*, 2017), the multivariate generalization of the Goodman (1953, 1959) regression method (Collingwood *et al.*, 2016; Lau *et al.*, 2023), and the `vottrans` (Gampmayer, 2016), `RxCEcolInf` (Greiner *et al.*, 2021), and `eiCircles` (Forcina and Pavía, 2024) packages. It should be noted that `RxCEcolInf` has not been supported or maintained since November 2022 and has been removed from the CRAN repository. Other related packages include the `eco` (Imai *et al.*, 2008, 2011), `MCMCpack` (Martin *et al.*, 2011), `ei` (King and Roberts, 2016), and `ei.Datasets` (Pavía, 2022) packages.

[3]While some models assume similar conditional row distributions across tables, others prefer to see them as related, considering them as realizations of an underlying probability distribution.

[4]There is a subtle difference between inferring probabilities and fractions. Probabilities can be observed as the underlying propensities that voters have to behave in a certain way, either based on their latent preferences or in a subsequent election conditioned on their behavior in a previous one. Fractions measure the actual behavior of voters in the elections. Under a superpopulation scheme, probabilities serve to model how voters would have behaved if the elections were repeated several times in similar conditions and fractions account for the particular way voters behave in the only realized elections (Pavía, 2024b). In political science the interest is usually in knowing fractions, whereas in epidemiology the goal is estimating probabilities.

structure theory which asserts that voters, having a preferred policy position (usually called an ideal point) in a multidimensional issue space, make choices based on this position but also on some valence issues (Groseclose, 2001) common to all voters, taking into account the different special interests that each party and candidate represent and their "general popularity" caused by valence issues (Thomsen, 2011).

From a domain perspective, the substantial interpretation of the latent model is that the change between elections (or voting among different social groups), apart from stochastic variation, is generated in the same way for all voters, whereas, from an operationalizing perspective, the latent structure approach impels the use of econometric discrete choice models (Train, 2009). When using binary choice models, this leads to a particular functional relationship between the individual and the ecological (aggregate) correlation that Thomsen (1987) exploits for performing cross-level inference after assuming functional homogeneity; an assumption which is supposed to be valid within politically homogenous geographical regions.

Although strictly speaking the model developed in Thomsen (1987) only applies to genuine binary (2×2) choice, as Thomsen (1987) suggests, it can be extended to multivariate ($R \times C$) choice after adjusting initial estimates of binary choice probabilities to reach logical consistency. Two procedures have been proposed to achieve this. First, Thomsen (1987) conceived an innovative method to adjust crude binary probabilities based on an iterative refinement of the initial estimates that exploit the latent structure methodology in each step. Later, Park (2008) suggested doing this by using iterative proportional fitting (Deming and Stephan, 1940). These solutions to estimate $R \times C$ ecological tables from crude binary probabilities using latent structure approaches are, however, incomplete and were (until now) only programmed in some difficult-to-reach (and use) `C++` and `Stata` codes: `ecol` (Thomsen *et al.*, 1995; Siegumfeldt, 2004) and `VTR` (Park, 2002).

Regarding the limitations of `ecol` we find that it (i) does not yield measures of uncertainty (error estimates), (ii) only considers the logit transformation of the marginal observed proportions, (iii) rests on the Yule's Q approximation (Johnson and Kotz, 1972) to derive cross-probability estimates from the marginal proportions and estimated correlations, and (iv) requires the choosing as reference of both a row and a column option, on which the attained solution is dependent. This last feature of the approach differentiates it from multinomial logistic models, where solutions are independent of what option is chosen as reference. Regarding `VTR`, its main restrictions are that it (i) achieves congruence using an ad hoc alien method, (ii) returns ($1 - \alpha = 0.95$) confidence intervals solely in the 2×2 case, and (iii) only estimates global tables when, as it is well-known (e.g., King, 1997; Pavía and Romero, 2024a), solutions attained by combining local solutions tend to be superior.

The aim of this paper is twofold. On the one hand, it introduces the R-package `ecolRxC`, an easy-to-reach, well-documented package, accessible on CRAN, that implements, extends, and improves the solutions proposed by Thomsen (1987) and Park (2008). On the other hand, it statistically documents and explains all its new extensions and shows, using actual inner cell election data, how the new models lead to more accurate solutions. We use real data from several general elections held in New Zealand and Scotland to assess accuracy. Despite the secrecy of the vote, the actual cross-distributions between voting for a party and voting for a candidate are available in these elections.

The `ecolRxC` package, in addition to being able to generate `ecol` and `VTR` outputs, extends Thomsen (1987) and Park (2008) in four directions. It (i) can generate solutions for all local units, also when using Park's approximation; (ii) can estimate uncertainties for both global and local solutions, with both approaches and for the $R \times C$ general case; (iii) can produce estimates that do not depend on choosing a reference row and column; and (iv) can handle as many as eight different scenarios (Pavía, 2023) regarding entries and exists in the electoral lists between elections, in addition to the option of simply adjusting the census changes (Brown and Payne, 1986).

2. Methodological background

Without loss of generality, we consider the problem of inferring voter transition rates/probabilities between two sets of parties across two consecutive elections and assume the same voters ($i = 1, 2, \ldots, T$) participating in both elections. These are restrictive conditions that can obviously be relaxed, for instance, by also considering entries and exits on the census. Later, to test the different methodologies, we also consider the case where voters have two choices in the same election: one for a party and the other for a candidate (who need not come from the chosen party). In that scenario, we can observe the choice of party as the "first election" and the choice of candidate as a choice of a "party" in the "second election."

Let R and C be the number of parties, including the "party" of abstainers, competing in both elections. The goal is to estimate, using ecological inference, the $R \times C$ matrix of joint probabilities/fractions p_{jk} of voting for parties j and k ($j = 1, 2, \ldots, R$ and $k = 1, 2, \ldots, C$) in, respectively, elections 1 and 2 in the whole region. The matrix for the whole region (district) can either be estimated directly or indirectly by first estimating the matrix for each local unit within the district and then adding all the local estimates.

To respond to this challenge, ecological inference exploits the known electoral support (the marginal probabilities/fractions/counts) gained by the competing parties in the two elections in a set of polling units ($u = 1, 2, \ldots, U$) that make up the constituency/district. This defines an under-identified problem that requires some assumptions to be made. Unlike most methods, which assume similar/related p_{jk} across units (i.e., that the p_{jk} are (conditional) independent of u), Thomsen (1987)—the latent structure approach—supposes, grounded on spatial and valence theories of party choice (see, e.g., Sanders *et al.*, 2011), that the individual probability of a certain choice is function of a latent variable (or set of variables) associated with the individual, as well as of the parties' popularities and positions (see, also, Thomsen, 2011). Under these assumptions, each voter's latent position drives her/his choices in the two elections and shapes the observed aggregate outcomes across all individuals in the local unit.

With binary choice and assuming functional homogeneity across all individuals (i.e., constant party positions and popularities across units), Thomsen (1987), Park (2008) and Park *et al.* (2014) demonstrate that the latent structure approach enables (i) the aggregation of individual choices within local units and (ii) the establishment of a latent relationship with observed fractions from which ecological inference can be carried out without the need to estimate the latent variables. They prove that to perform ecological inference it is enough to ascertain a functional relationship between the individual and ecological correlations.

2.1 Binary choice model: the 2 × 2 case

Mathematically, considering an election in which each voter must choose between two parties (1 and 0) and denoting by l_i the d-dimensional vector of latent long-term policy positions (and/or partisanship) of voter i, the binary latent structure model choice states that l_i impacts probabilistically on the voter's choice in both elections, $v_{i,1}$ and $v_{i,2}$, through the equations:

$$\begin{aligned} P(v_{i,1} = 1) &= f(\alpha_1 + \beta_1 l_i) \\ P(v_{i,2} = 1) &= f(\alpha_2 + \beta_2 l_i) \end{aligned} \tag{1}$$

where the coefficient α_t and the d-vector of coefficients β_t ($t = 1, 2$) capture, respectively, the popularities and party positions of the reference party in both elections[5] and f is

[5]In equation (1), voters' preferences and party policy positions are represented as vectors. While a more parsimonious specification can be achieved by representing preferences and positions as scalar ideal points, as is common in the literature on political ideology (Battista *et al.*, 2022), we prefer the current specification because it adds flexibility to the model. Both the number of relevant dimensions at play in each election and the weights assigned by electors to each dimension can vary

a proper function that can either be the cumulative normal function or the logistic function.[6]

Equation (1) models the causal process in which, in our model, the stable opinions of voters are confronted with the often-changing policies of parties and candidates to produce the voting behavior. As argued and tested on cross-national data in Thomsen (2011), the interplay between individual voters and parties is better modelled by the product between the position of the voter and the position of the party (known as "the directional model" in the literature on issue voting; Rabinowitz and Macdonald, 1989) than by the distance between the two (known as "the proximity model"; Downs, 1957). With the directional model, the valence parameters (α_t) are much better predicted by the mean sympathy score for the party than in the proximity model.

When only aggregate information is available, all components in equation (1) are unobserved, so to relate it with the known outcomes, individual probabilities must be aggregated to (averaged at) the polling unit (or constituency) level. In doing so, we consider that the number of voters T_u in each unit is large enough as to make "sampling" errors negligible. This allows to state that the relative marginal outcomes $p_{u,1} = \sum_{i \in u} v_{iu,1}/T_u$ and $p_{u,2} = \sum_{i \in u} v_{iu,2}/T_u$ are (almost) equal to the expected vote fractions in the unit and get:

$$E(v_{iu,t}) = p_{u,\mathrm{t}} = \int_{\Re^d} \Phi(\alpha_t + \beta_t l_{iu}) \phi(l_{iu} | L_u, \Omega) \partial l_{iu}$$

after assuming, as in Thomsen (1987), that the underlying dimension l_{iu} is normally distributed with mean L_u and variance–covariance matrix Ω.

Carrying out the integral (Thomsen, 1987: 56), we obtain:

$$E(v_{iu,t}) = p_{u,\mathrm{t}} = \Phi\left(\frac{\alpha_t + \beta_t L_u}{\sqrt{1 + \beta_t \Omega \beta_t^T}}\right) \tag{2}$$

which is formally the same as (1), except for a rescaling value. Indeed, as α_t, β_t, and Ω are assumed to be constant across units, equations (1) and (2) state that the model for aggregate behavior, apart from rescaling, is equal to the model for individual behavior. What is more, (1) implies that the utilities to vote for a given party in the two elections (their inverse-probit transformed probabilities) are linearly related to each other and, by application of the axiom of local independence at the individual level, that the joint distributions of $\Phi^{-1}(p_{u,1})$ and $\Phi^{-1}(p_{u,2})$ (and of $\Phi^{-1}(p_{u,1})$ and $\Phi^{-1}(1 - p_{u,2})$, $\Phi^{-1}(1 - p_{u,1})$ and $\Phi^{-1}(p_{u,2})$, and $\Phi^{-1}(1 - p_{u,1})$ and $\Phi^{-1}(1 - p_{u,2})$) are binormal. In general:

$$p_{jk} = \Phi_2(\Phi^{-1}(p_{j,1}), \Phi^{-1}(p_{k,2}), \rho_{jk}) \tag{3}$$

between elections. Our specification allows us to capture both issues through the β_t's. Although the multidimensional representation does not play any role in our current implementation, as it does not require the explicit estimation of latent factors, this is recommended in an implementation where latent factors were estimated. Furthermore, to capture the relationships that exist between candidate/party policy positions and valence differentials among candidates/parties, as discussed in the literature (see, e.g., Ansolabehere *et al.*, 2001; Groseclose, 2001), the above specification could be extended under a multichoice model approach by considering a simultaneous multi-equational system for each election.

[6]For mathematical convenience, in the rest of this subsection, we assume $f = \Phi$, the cumulative distribution function of the standard normal distribution.

which allows estimation of the joint probabilities when ρ_{jk}, the so-called tetrachoric correlation coefficient, is known.

As Thomsen (1987) shows, when it is assumed that the latent variable variation between individuals has the same structure as the latent variable variation between local units (a reasonable isomorphism assumption when units are not too large within relatively politically homogenous geographical regions) and that the former variation is significantly greater than the latter, ρ_{jk} can be properly approximated by the corresponding ecological probit (or logit) correlation, ρ_e, and be estimated from the observed marginal counts. An alternative identification condition is presented in Park (2008: 34–38).

At this point, joint probabilities can be directly estimated using equation (3) or, as Thomsen (1987) suggests, be approximated using equation (4). Equation (4) is derived using Yule's Q approximation to estimate the tetrachoric correlation (see Thomsen, 1987: 64) and has slightly lesser computational costs.

$$p_{jk} \approx \frac{1 + 2\hat{\rho}_e p_{j,1} + 2\hat{\rho}_e p_{k,2} - \hat{\rho}_e - \sqrt{(1 + 2\hat{\rho}_e p_{j,1} + 2\hat{\rho}_e p_{k,2} - \hat{\rho}_e)^2 - 8\hat{\rho}_e(1 + \hat{\rho}_e)p_{j,1}p_{k,2}}}{4\hat{\rho}_e} \tag{4}$$

2.2. The general R × C case

The greatest limitation to the use of the latent structure theory for bivariate choice on actual elections resides in the non-duality of voters' choices in actual elections. Even in two-party systems (or second round-off presidential elections) the third alternative of "non-voting" is a possible choice. Hence, the above 2 × 2 approach needs to be extended to the R × C case to be useful. Both Thomsen (1987) and Park (2008) each make a proposal, with both proposals departing from crude binary choice estimates.

As a first step, they estimate raw joint probabilities p_{jk} by applying either equation (3) or (4) to the artificial set of binary choices defined by choosing, in election 1, between party j and the other parties and, in election 2, between party k and all other parties. Unfortunately, these crude binary choice-estimated probabilities are not congruent with the observed results. The sum across j (k) of the estimated joint probabilities $\hat{p}_{jk}$ does match the observed marginal fractions p_{j+} (p_{+k}). Hence, as a second step, Park (2008) and Thomsen (1987) propose a way to fix this. Park (2008) suggests using the iterative proportional fitting algorithm (Deming and Stephan, 1940), whereas Thomsen (1987) proposes the use of a more complex algorithm that requires a reference or pivotal party to be chosen in each election.

Denoting by r_1 and r_2 the reference parties in, respectively, election 1 and election 2, the original iterative algorithm of Thomsen works as follows.

(i) First, crude binary choice probabilities $\hat{p}_{jk}^{(0)}$ are computed.

(ii) Second, the temporary estimates of (i) are used to estimate the margins of a set of theoretical 2 × 2 tables composed by the set of parties $\{\{j, r_1\}, \{k, r_2\}\}$, with $j \neq r_1$ and $k \neq r_2$: $\hat{p}_{jk}^{(0)} + \hat{p}_{jr_2}^{(0)}$, $\hat{p}_{r_1k}^{(0)} + \hat{p}_{r_1r_2}^{(0)}$, $\hat{p}_{jk}^{(0)} + \hat{p}_{r_1k}^{(0)}$, and $\hat{p}_{jr_2}^{(0)} + \hat{p}_{r_1r_2}^{(0)}$. And, from them, the joint probabilities $\hat{p}_{jk}^{(0)}$ are updated employing equation (5), which derives from Yule's Q approximation for the tetrachoric correlation.

$$\hat{p}_{jk}^{(1)} = \frac{\hat{p}_{jr_2}^{(0)}\,\hat{p}_{r_1k}^{(0)}}{\hat{p}_{r_1r_2}^{(0)}} \frac{1 + \hat{r}_{jk|r_1,r_2}^{(0)}}{1 + \hat{r}_{jk|r_1,r_2}^{(0)}} \tag{5}$$

where $\hat{r}_{jk|r_1,r_2}^{(0)}$ is the across units ecological correlation between $ln((\hat{p}_{jk}^{(0)} + \hat{p}_{jr_2}^{(0)})/(\hat{p}_{r_1k}^{(0)} + \hat{p}_{r_1r_2}^{(0)}))$ and $ln((\hat{p}_{jk}^{(0)} + \hat{p}_{r_1k}^{(0)})/(\hat{p}_{jr_2}^{(0)} + \hat{p}_{r_1r_2}^{(0)}))$.

(iii) After applying (ii), we have new (updated) estimates for each pair of p_{jk} with $j \neq r_1$ or $k \neq r_2$, but not when $j = r_1$ or $k = r_2$. These probabilities are re-estimated (updated) by re-scaling them using as rates the relative discrepancies between the aggregations of the observed and temporary estimates:

$$\hat{p}^{(1)}_{r_1 k} = \hat{p}^{(0)}_{r_1 k} \frac{p_{+k}}{\tilde{p}^{(1)}_{+k}}$$

$$\hat{p}^{(1)}_{j r_2} = \hat{p}^{(0)}_{j r_2} \frac{p_{j+}}{\tilde{p}^{(1)}_{j+}}$$

where p_{k+} and p_{+j} are the observed marginal fractions and $\tilde{p}^{(1)}_{j+} = \sum_{k \neq r_2} \hat{p}^{(1)}_{jk} + \hat{p}^{(0)}_{j r_2}$ and $\tilde{p}^{(1)}_{+k} = \sum_{j \neq r_1} \hat{p}^{(1)}_{jk} + \hat{p}^{(0)}_{r_1 k}$ temporary marginal fraction estimates.

(iv) Finally, we come back to (i), replace $\hat{p}^{(0)}_{jk}$ by the new estimates and iterate until the process converges.

3. ecolRxC methodological extensions

As stated in the introduction, `ecolRxC` extends previous latent factor ecological inference software in several directions. In this section, we refer to these in more detail.

3.1 Probit transformations and exact estimates of probabilities

The latent factor ecological inference approach as originally suggested in the seminal work of Thomsen (1987) relies on Yule's Q approximation to update joint probabilities and only considers the Pearson correlation across units of the logit transformation of the binary choices. In other words, it uses the ecological logit correlation as an estimator[7] of the individual tetrachoric correlations for all tetrachoric (fourfold) subsets of the voter's choice (Thomsen *et al.*, 1991). `ecolRxC` extends this by also including the options of using the exact equation (3) instead of the approximation equation (4) and working with probit transformations. As we show in section 5 this leads to more accurate estimates, on average.

3.2 Measuring uncertainties

An estimate is not complete without a measurement of its estimation error; that is, its level of associated uncertainty. For the 2×2 case, as referenced in Park (2008), Achen (2000) proposes estimating the standard errors of the binary Thomsen estimator using Fisher's z-transformations. Specifically, after computing $1 - \alpha$ confidence intervals for the ecological correlation

$$[\hat{\rho}_e^-,\ \hat{\rho}_e^+] = \left[tanh\left(\frac{1}{2} ln \frac{1 + \hat{\rho}_e}{1 - \hat{\rho}_e}\right) \mp \frac{z_{\alpha/2}}{\sqrt{U - 2.5}} \right] \tag{6}$$

lower and upper limits of $1 - \alpha$ confidence intervals for $\hat{p}_{jk}$ can be constructed, applying the plug-in principle, replacing in either (3) or (4) the correlation by $\hat{\rho}_e^-$ and $\hat{\rho}_e^+$, respectively.[8]

The extension up to the $R \times C$ case is made in `ecolRxC` via bootstrap (Efron and Tibshirani, 1994) by sampling in the estimated confidence intervals of the crude binary probabilities attained

[7] In this regard, it should be noted that `ecolRxC`, like `ecol` does, computes weighted correlations. Each unit logit/probit transformed marginal fraction is weighed up using as weight the corresponding (observed/estimated) number of voters involved in the computation.

[8] In equation (6) *tanh* stands for the hyperbolic tangent function and $z_{\alpha/2}$ for the $1 - \alpha/2$ percentile of a standard normal distribution.

using the 2×2 approach. Specifically, `ecolRxC` computes $1 - \alpha$ confidence intervals for the estimated probabilities by (i) randomly extracting B resamples from each estimated $1 - \alpha$ crude binary probability confidence interval ($\tilde{p}_{jk}^{(0),b}$ $b = 1, 2, \ldots, B$), (ii) making each set of resamples $\{\tilde{p}_{jk}^{(0),b}\}$ congruent/compatible with the known outcomes, using either the iterative proportional fitting algorithm or the Thomson algorithm detailed in subsection 2.2, and (iii) calculating for each set of final congruent estimates their $\alpha/2$ and $1 - \alpha/2$ percentiles. An alternative for estimating uncertainties would be to directly bootstrap polling units. We consider our proposal more in line with the approach.

3.3 Estimation of unit transfer tables

`ecolRxC` estimates both local (polling unit) and global (constituency) vote transfer matrices. In section 2, and in order not to overwhelm the exposition and notation, we choose to remain ambiguous as to whether the p_{jk} probabilities refer to a polling unit or to the whole district. As a rule, `ecolRxC` applies the methods presented in section 2 working at the polling unit level, obtaining the global matrices as aggregation (composition) of local matrices. As in the case of Park's solution, nevertheless, `ecolRxC` also offers the possibility of directly estimating global matrices by just applying either equation (3) or (4) to the constituency known margins.

3.4 Eliminating indeterminacy implied by pivotal cells

As detailed in subsection 2.2, the original algorithm proposed by Thomsen (1987) reaches consistency/congruency in the final $R \times C$ estimates by choosing a row and a column as reference. This means that when the Thomsen procedure is employed, the solution attained depends on which row–column pair is chosen as pivotal. `ecolRxC`, in addition to retaining this option, avoids this indeterminacy by building its final solution as a combination of all potential solutions that can be reached considering as reference all the possible pairs of a row and a column.

This raises the question of how to combine the RC attained solutions, where R is the number of rows and C the number of columns. As default, `ecolRxC` builds its composite (local and global) solutions as a weighted average of the RC reference solutions with weights equal to the absolute values of the crude ecological correlations, $\hat{\rho}_{r_1 r_2}^{(0)}$. We call this combined solution AVCR.

More specifically, `ecolRxC` computes eight different global solutions which differ in the way they weight unit solutions. These eight composite solutions can be grouped into two families, according to whether the weights depend only on the reference row–column pair or if they are also a function of the unit. The general formulae for both cases are given by equations (7) and (8), respectively:

$$\sum_{u=1}^{U} \frac{1}{\sum_{r_1=1}^{R} \sum_{r_2=1}^{C} \omega_{r_1 r_2}} \sum_{r_1=1}^{R} \sum_{r_2=1}^{C} \omega_{r_1 r_2} [\hat{v}_{jk}^{u}]_{r_1}^{r_2} \tag{7}$$

$$\sum_{u=1}^{U} \frac{1}{\sum_{r_1=1}^{R} \sum_{r_2=1}^{C} \omega_{r_1 r_2}^{u}} \sum_{r_1=1}^{R} \sum_{r_2=1}^{C} \omega_{r_1 r_2}^{u} [\hat{v}_{jk}^{u}]_{r_1}^{r_2} \tag{8}$$

where $[\hat{v}_{jk}^{u}]_{r_1}^{r_2}$ denotes the (final) estimated matrix of transfer votes (counts) achieved for unit u when the r_1 row and the r_2 column are used as reference, and $\omega_{r_1 r_2}$ and $\omega_{r_1 r_2}^{u}$ stand for a generic global and local weight, respectively.

Different solutions are reached depending on how weights are defined. Specifically, in addition to considering constant weights (which is equivalent to taking a simple average and thus is called the "Mean" solution), `ecolRxC` considers four possibilities for global weights, $\omega_{r_1 r_2}$:

- $\hat{v}^{(0)}_{r_1 r_2}$: Reference cell number of voters, RCNV
- $\sqrt{\hat{v}^{(0)}_{r_1 r_2}}$: Square root reference cell number of voters, SQRCNV
- $\sqrt{v_{r_1+} \cdot v_{+r_2}}$: Square root reference margins, SQRM
- $|\hat{\rho}^{(0)}_{r_1 r_2}|$: Absolute values of reference correlations, AVCR

and three options for local weights, $\omega^{u}_{r_1 r_2}$:

- $\hat{v}^{u,(0)}_{r_1 r_2}$: Local reference cell number of voters: LRCNV
- $\sqrt{\hat{v}^{u,(0)}_{r_1 r_2}}$: Local square root reference cell number of voters: LSQRCNV
- $\sqrt{v^{u}_{r_1+} \cdot v^{u}_{+r_2}}$: Local square root reference margins: LSQRM

where, on the one hand, $\hat{v}^{(0)}_{r_1 r_2}$ is the crude estimate of the global total votes for the (r_1, r_2)-cell, $\hat{\rho}^{(0)}_{r_1 r_2}$ is the crude (logit/probit)-estimated ecological correlation linked to the (r_1, r_2)-cell, and v_{r_1+} and v_{+r_2} are the observed global margins (number of votes) corresponding to row r_1 and column r_2, respectively. And, on the other hand, $\hat{v}^{u,(0)}_{r_1 r_2}$ is the crude estimate of total votes for the (r_1, r_2)-cell of table u, and $v^{u}_{r_1+}$ and $v^{u}_{+r_2}$ are the observed margins (number of votes) corresponding to row r_1 and column r_2 of table u, respectively.

3.5 Census changes

Finally, other extensions included in `ecolRxC` are the options of either adjusting censuses or estimating census changes between elections, with as many as eight different scenarios being considered for the latter. More details are in Pavía (2023) or in the package documentation.

4. An application example

Using `ecolRxC` is quite simple. The user only needs two objects (matrices or data frames) with the observed row and column margin counts in a set of U related contingency tables and to customize, if desired, its other arguments, as described in Appendix I; where details on the function arguments and outputs can be found. To exemplify how the function works we consider the problem of estimating the vote transition fractions between a set of parties and a set of candidates in a mixed-member proportional election in which voters vote simultaneously for a party and a candidate.

As an example, we apply `ecolRxC` with default options to the voting data recorded in the electorate of Northland during the 2017 New Zealand general elections. In that election, the electors of Northland were called to choose among 19 parties and 9 candidates, and a total of 40102 vote-tickets were recorded as distributed across 136 polling units. We use the data available on that election in the R-package `ei.Datasets` (Pavía, 2022), but before applying ecological inference, as is usual practice (e.g., Klima *et al.*, 2016; Plescia and De Sio, 2018; Pavía and Romero, 2024b), we merge small parties and candidates together in "Others." We aggregate together those parties or candidates that individually do not gain at least 3 percent of the total constituency vote. This simplifies the problem by going from estimating a 19×9 matrix to estimating a 5×5 matrix. The interested reader can find the code for this example in Appendix II. The code ends calling the function `plot`, which shows a graphic summary of the value of `ecolRxC` (see Figure 1). Interested readers can find estimated confidence intervals of the row-fraction estimates displayed in Figure 1 in Appendix III.

5. An assessment of `ecolRxC`

As previously stated, `ecolRxC` extends the former implementations of the latent structure model for ecological inference: `ecol` and `VTR`. This section gauges its practical performance with real

Figure 1. Graphical summary example of an output of `ecolRxC`. The global total counts are presented in the margins of the plot table and the estimated transition row-standardized fractions in the inner-cells of the table. The sizes of the numbers in each interior cell are (in log-scale) proportional to its corresponding estimated counts and the intensity of the color of each cell within each row is proportional to the fraction of voters of the corresponding row option that switch to the corresponding column option.

data. Data and accuracy measures are presented in subsections 5.1 and 5.2, respectively. Subsection 5.3 is devoted to evaluating `ecolRxC` with different specifications. First, we assess whether the new approaches improve previous solutions. Second, we study the impact of weights in `ecolRxC` composite solutions, as defined in subsection 3.4. Third, we explore whether the observed election features could be employed to automatically determine which `ecolRxC` specification produces the most accurate solution. Finally, we end the section by pondering the relative performance of `ecolRxC` by comparing its accuracy with that reported for `ei.MD.bayes` and `nslphom` in other studies.

5.1 Data

For assessing the accuracy of ecological inference estimates, the closeness between estimates and true cross-distributions needs to be measured. The problem with behavioral data is that it is not always easy to define what "true" means. Fortunately, this seems to be a less of a problem with voting behavior: to discern the actual electoral behavior of a voter, all that is necessary is to know how the voter votes. Unfortunately, because of the principle of voting secrecy, this is not possible: the actual behavior of individual voters is by definition unknown.

In some elections, however, such as when voters cast multiple votes in the same ballot, actual vote flows can be known. This is the case of the 2007 Scottish Parliament election and of the Parliament elections of New Zealand since 2002. In those elections, the actual constituency party-to-candidate cross-distributions of votes were disclosed by the electoral authorities and later gathered, together with the marginal distributions of votes across polling stations, in the R-package `ei.Datasets` (Pavía, 2022). In both countries, a mixed-member system that combines first-past-the-post voting and party-list proportional representation is used to elect Parliament representatives, with voters, grouped into districts, casting two votes in the same ballot: one for a district candidate and another for a (regional/national) party list. Constituency/district cross-vote distributions are built from this.

As district candidates vary from district to district (and parties sometimes also vary by region), a different cross-table is available for each district and year. To be specific, `ei.Datasets`

contains a total of 565 datasets/elections grouped into eight sets—as all elections that took place in the same country and year share a similar political environment. This comprises a large number of examples that embrace "a broad diversity of electoral contexts" (Pavía, 2022: 253). We rely on these datasets to assess `ecolRxC`.

Indeed, the datasets in `ei.Datasets` are becoming a standard to evaluate ecological inference algorithms. For example, a large number of these datasets were employed in the ecological inference comparative studies performed in Plescia and De Sio (2018) and Pavía and Romero (2023, 2024b). Before using the data, however, we merge less popular (in number of votes) parties and candidates. As is usual practice (e.g., Klima *et al.*, 2016; Pavía and Romero 2023; Pavía, 2024a), those parties and candidates that individually did not reach a minimum of the district share of votes were grouped in "Others." As in the example, we set this minimum at 3 percent.

5.2 Measures of accuracy

We assess accuracy by measuring distances between global (constituency) estimated and true vote transfer tables, using the error and discrepancy indices, *EI* and *EPW* (equations (9) and (10))[9] as well as an index, *EQ*, based on quadratic differences (equation (11)). *EI* can be interpreted as the proportion of votes which must be relocated in one table to construct the other table, *EPW* as the mean average of the errors estimating the row-standardized vote transfer rates, and *EQ* is an index that penalizes larger errors. The smaller these indices, the closer the estimated and actual tables.

$$EI = 100 \times \frac{0.5 \sum_{j=1}^{R} \sum_{k=1}^{C} | v_{jk} - \hat{v}_{jk}|}{\sum_{j=1}^{R} \sum_{k=1}^{C} v_{jk}} \tag{9}$$

$$EPW = 100 \times \frac{\sum_{j=1}^{R} \sum_{k=1}^{C} v_{jk} | p_{k|j} - \hat{p}_{k|j}|}{\sum_{j=1}^{J} \sum_{k=1}^{K} v_{jk}} \tag{10}$$

$$EQ = 100 \times \frac{\sqrt{\sum_{j=1}^{R} \sum_{k=1}^{C} (v_{jk} - \hat{v}_{jk})^2}}{\sum_{j=1}^{R} \sum_{k=1}^{C} v_{jk}} \tag{11}$$

where v_{jk} ($\hat{v}_{jk}$) denotes the actual (estimated) number of voters who simultaneously voted for party j and candidate k in the entire population and $p_{k|j}$ the row-standardized proportion of voters in the entire electoral space who voted for candidate k among those who voted for party j.

5.3 Results

The function `ecolRxC` allows an important level of customization simply by varying three of its main arguments: `scale`, `method`, and `Yule.aprox`. Different versions of ecological inference latent structure models/procedures emerge depending on the values chosen for these arguments. With `scale` determining just what transformation is applied to the known fraction margins, `method` and `Yule.aprox` have a greater impact on the particular algorithm performed by `ecolRxC`. In order to make the analysis easier as well as the presentation that follows, Table 1 lists and names the different procedures that emerge by combining all the possible values for the `method` and `Yule.aprox` arguments.

[9] *EI* and *EPW* are two popular distance matrix indices in ecological inference (Thomsen *et al.*, 1991; Klima *et al.*, 2016; Pavía and Romero, 2024a).

Table 1. Basic ecological inference latent structure procedures available in `ecolRxC`

Procedure acronym	Arguments' values		Comments
	`Method`	`Yule.aprox`	
VTR	`'IPF'`	`FALSE`	This corresponds to Park (2002) `VTR` procedure.
VTR-local	`'IPF'`	`FALSE`	This procedure also requires `local = TRUE`.
VTR-Yule	`'IPF'`	`TRUE`	As VTR, but using (4) instead of (3).
VTR-local-Yule	`'IPF'`	`TRUE`	This procedure also requires `local = TRUE`.
ecol	`'Thomsen'`	`TRUE`	This corresponds to Thomsen *et al*. (1995) and Siegumfeldt (2004) `ecol` procedure when the `reference` argument is set equal to a vector.
ecolRxC-Yule	`'Thomsen'`	`TRUE`	As ecolRxC, but using (4) instead of (3).
ecol-biN	`'Thomsen'`	`FALSE`	As ecol, but using (3) instead of (4).
ecolRxC	`'Thomsen'`	`FALSE`	This corresponds to the default method when `scale = 'probit'`.

Source: compiled by the authors from `ecolRxC` (version 0.1.1-10).

In the case of ecol, the final attained estimate depends on which party and candidate is used as reference. Thomsen (1987: 74) recommends choosing a neutral option, such as abstention, as reference at both elections. This is not possible with our data as that information is missing. Hence, as an alternative, we decided to choose all possible combinations of reference options and attach to ecol in the assessments the average error across all these combinations. This entails considering extreme combinations as references. As we shall see later, more accurate solutions could be attained for ecol with a clever selection of references in the spirit of Thomsen's recommendation.

5.3.1 Comparing the basic latent factor procedures in `ecolRxC`

A summary of the accuracy of the different specifications/procedures listed in Table 1 is presented in Figure 2 and Table 2. Figure 2 graphically shows the overall average accuracy of the different procedures measured with *EI*, *EPW*, and *EQ* when both transformations (logit and probit) are employed for scaling the observed proportion margins. In Table 2, only *EI* errors are presented, with the elections grouped by country and by year of celebration. We consider this the most logical way to group these elections, since all datasets from the same year and country reflect a shared political environment.

Several findings emerge when analyzing the different panels in Figure 2. First, overall the scale/transformation used has a really small impact on the accuracy of the estimates, with probit transformations tending to yield, on average, slight better estimates (see also Table 2). Second, all error measures (*EI*, *EPW*, and *EQ*) draw almost the same order of preferences among the different procedures, with the default extended model proposed in this paper (the ecolRxC procedure) clearly outperforming the rest of the configurations. Third, overall, reaching congruence utilizing the Thomsen algorithm leads to more accurate solutions than employing the iterative proportional fitting algorithm. Fourth, as a rule, using the Yule approximation deteriorates the accuracy of the estimates, with ecol-biN (which uses Yule approximation) and VTR (which does not) generating solutions of relatively similar quality. Fifth, the estimation of unit (local) solutions when employing VTR has only a slight impact in terms of global accuracy. Consideration should be given, nevertheless, as to the value of having estimates for each unit in some applications. In any case, whatever the specification considered, we can affirm that the ecological inference approach adds significant value to solving this problem, since simply assuming independence between the rows and columns yields an average error of 36.98, as measured by the *EI* coefficient.

The analysis of results of Table 2 reinforces the previous findings. Similar conclusions to the ones attained pooling all the elections are reached when the elections are grouped by country and year.[10] On average, the ecolRxC procedure is the one generating by far the most accurate

[10]Interested readers can find the details about the *EPW* and *EQ* errors in Appendix IV.

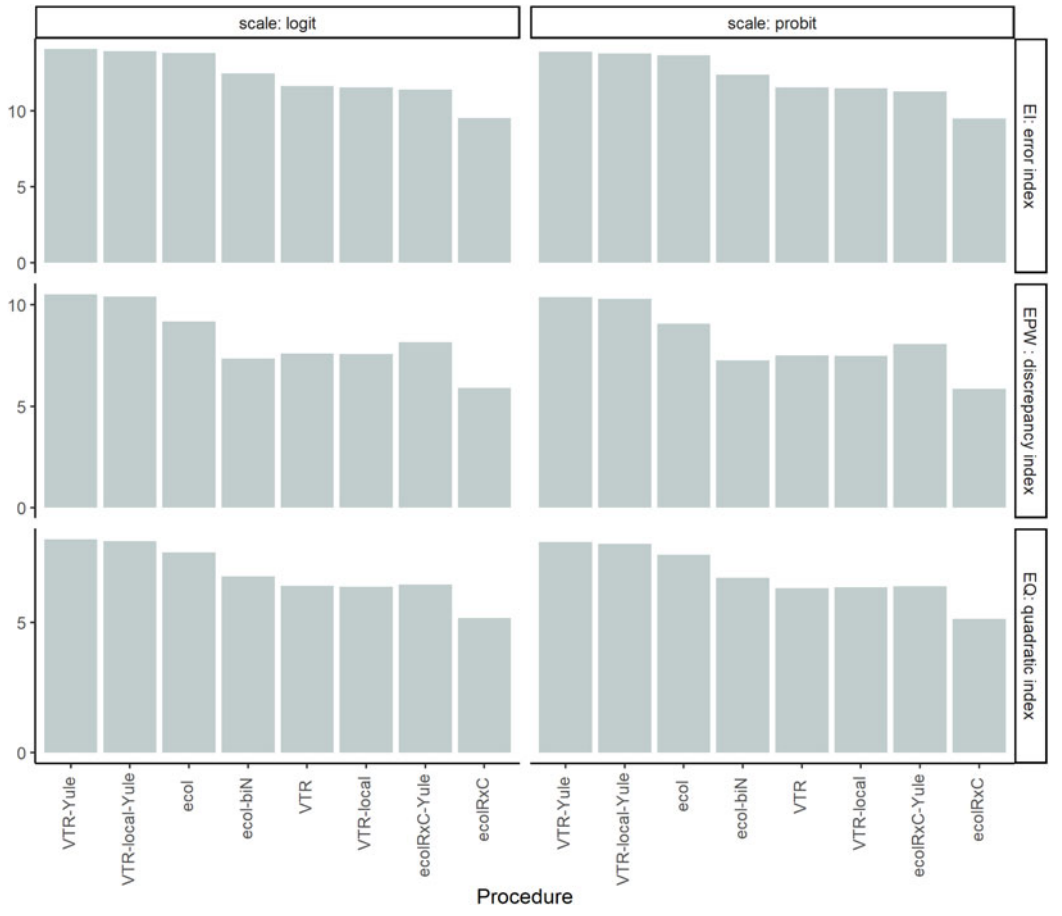

Figure 2. Graphical representation of average values of *EI* (upper panels), *EPW* (intermediate panels), and *EQ* (lower panels) errors by procedure (specification) using either the logit (left panels) or the probit (right panels) fraction-transformations. The correspondence between the acronyms of the procedures and its `ecolRxC` specification is detailed in Table 1. In the ecol specification, errors are computed as simple averages of the *RC* errors corresponding to the *RC* possible reference solutions. The smaller the number, the better the accuracy.

solutions. The results by group of elections, however, are heterogeneous. In general, the current implementations of the methodology encounter significantly more problems in the group of the Scottish datasets.[11]

An analysis of the features affecting the accuracy of estimates reveals that `ecolRxC`, like other ecological inference models, faces challenges when the number of polling stations is small and the dimension of the contingency tables (the number of coefficients to be estimated) increases. Equally, the examination also confirms that the accuracy of its solutions deteriorates when unit tables are more heterogeneous and the relationships between row and column options weaken. Furthermore, the scrutiny also shows that our implementations of the latent structure approach using binary choice models suffer more, in comparative terms, when there are smaller

[11]The relative bad performance of `ecolRxC` in the Scottish data is an issue that it shares with `ei.MD.bayes`. This function even encounters more problems than `ecolRxC` in this subset of elections. Pavía and Romero (2023) report an average *EI* error of 23.09 for `ei.MD.bayes` in this subset, even after manually improving all its tuning parameters.

Table 2. Averages of *EI* errors by group of elections

Country year # of Elections Avg. # of units Avg. # of cells	NZ 2002 N = 69 $\bar{U}$ = 83.2 $\overline{RC}$ = 39.5	NZ 2005 N = 69 $\bar{U}$ = 81.8 $\overline{RC}$ = 23.8	SCO 2007 N = 73 $\bar{U}$ = 70.2 $\overline{RC}$ = 35.2	NZ 2008 N = 70 $\bar{U}$ = 84.1 $\overline{RC}$ = 23.4	NZ 2011 N = 70 $\bar{U}$ = 85.7 $\overline{RC}$ = 26.2	NZ 2014 N = 71 $\bar{U}$ = 81.2 $\overline{RC}$ = 27.9	NZ 2017 N = 71 $\bar{U}$ = 101.9 $\overline{RC}$ = 24.8	NZ 2020 N = 72 $\bar{U}$ = 134.9 $\overline{RC}$ = 24.5	NZ + SCO N = 565 $\bar{U}$ = 90.5 $\overline{RC}$ = 28.2
Logit transformations									
VTR-Yule	14.61	11.59	25.71	10.76	11.98	13.37	11.61	12.57	14.08
VTR-local-Yule	14.48	11.44	25.56	10.62	11.84	13.24	11.48	12.43	13.94
ecol	15.39	11.51	24.02	10.59	11.99	13.26	11.16	12.23	13.81
ecol-biN	15.02	10.84	22.21	9.53	10.91	11.63	9.43	10.00	12.48
VTR	13.19	9.98	23.90	8.46	9.68	10.29	8.28	9.05	11.65
VTR-local	13.08	9.86	23.73	8.36	9.58	10.25	8.26	9.02	11.57
ecolRxC-Yule	12.09	9.82	19.87	8.93	9.70	10.91	9.19	10.36	11.40
ecolRxC	10.92	8.85	17.50	7.58	8.21	8.62	6.80	7.63	9.54
Probit transformations									
VTR-Yule	14.50	11.45	25.55	10.63	11.83	13.23	11.45	12.35	13.93
VTR-local-Yule	14.39	11.32	25.42	10.50	11.70	13.10	11.34	12.22	13.80
ecol	15.25	11.38	23.86	10.47	11.84	13.10	11.04	12.03	13.66
ecol-biN	14.92	10.76	22.07	9.47	10.83	11.51	9.37	9.87	12.38
VTR	13.15	9.92	23.78	8.40	9.62	10.21	8.22	8.91	11.57
VTR-local	13.04	9.80	23.62	8.31	9.52	10.17	8.20	8.89	11.49
ecolRxC-Yule	12.02	9.73	19.72	8.85	9.61	10.81	9.11	10.19	11.29
ecolRxC	10.90	8.81	17.38	7.56	8.20	8.59	6.79	7.55	9.50

Source: compiled by the authors after applying with different specifications the function `ecolRxC` to the 565 datasets of the R package `ei.Datasets` (Pavía, 2022). The correspondence between the acronyms of the procedures and its `ecolRxC` specification is listed in Table 1. For the ecol specification, errors are computed as simple averages of the errors attained using as reference all the *RC* possible combinations with a row and a column. The smaller the number, the better the accuracy.

variabilities among row and column options. All this helps to explain, at least in part, the relatively poor estimates obtained for Scotland.[12]

5.3.2 On the impact of weights in `ecolRxC` composite solutions

The previous analysis clearly points to the ecolRxC specification as the one yielding more accurate results. Our proposal of choosing the weights of the absolute values of the ecological correlations as default to combine the *RC* reference solutions follows in the footsteps of Thomsen (1987), who recommended using the "party of abstainers" as reference in both elections. The "party of abstainers" is not only quite stable (i.e., it shows a strong ecological correlation across elections), but it also tends to be sizeable. In this respect, it merits an analysis of whether more accurate results could be obtained using other weights that put more emphasis on the size (in number of votes) of the reference options.

Table 3 presents the averages of *EI* errors[13] by group of elections for the eight composite solutions defined in subsection 3.4 when the ecolRxC procedure is employed to attain the polling unit estimates. Overall, the most accurate solutions are clearly obtained when weights are defined as the absolute values of the ecological correlations, although sporadically other composite solutions show a smaller average error in some groups of elections. According to these results, the decision to take the AVCR solution as default solution of `ecolRxC` appears to be an accurate choice, although the simple mean solution also provides quite accurate estimates.

5.3.3 Can observed features be employed to determine the most accurate `ecolRxC` specification?

The comparisons between ecol and ecolRxC specifications clearly show that the errors of the solutions built as (weighted) averages of the *RC* reference solutions are significantly smaller

[12]On one hand, Scotland's districts have, on average, fewer polling units and larger table sizes. On the other hand, Scotland's elections exhibit lower levels of marginal variability. The mean district within-unit diversities for Scotland, measured by averages of the standard deviations of across-unit marginal distributions, are 0.13 and 0.17 for parties and candidates, respectively. These figures are significantly smaller than the corresponding values for NZ, which are 0.20 and 0.25, respectively.

[13]*EPW* and *EQ* errors lead to similar conclusions. They can be found in Appendix V.

Table 3. Averages of *EI* errors by group of elections for the eight composite solutions

Country year	NZ 2002	NZ 2005	SCO 2007	NZ 2008	NZ 2011	NZ 2014	NZ 2017	NZ 2020	NZ + SCO
Mean	11.21	8.82	18.33	7.46	8.38	8.87	6.97	7.76	9.76
RCNV	11.94	9.50	17.47	8.16	9.01	9.22	7.26	7.73	10.06
SQRCNV	11.60	9.13	18.44	7.78	8.64	8.96	7.03	7.69	9.94
SQRM	11.38	9.11	17.81	7.70	8.55	8.90	6.99	7.70	9.79
AVCR	10.90	8.81	17.38	7.56	8.20	8.59	6.79	7.55	9.50
LRCNV	11.82	9.44	17.33	8.09	8.93	9.15	7.21	7.67	9.98
LSQRCNV	11.57	9.11	18.38	7.78	8.62	8.94	7.02	7.69	9.92
LSQRM	11.35	9.10	17.80	7.68	8.53	8.88	6.98	7.70	9.78

Source: compiled by the authors after applying the function `ecolRxC` with default options (`method = 'Thomsen'`, `scale = 'probit'`, `Yule.aprox = FALSE`) to the 565 datasets of the R package `ei.Datasets` (Pavía, 2022). The definition and acronyms of the different composite solutions are detailed in subsection 3.4. The smaller the number, the better the accuracy.

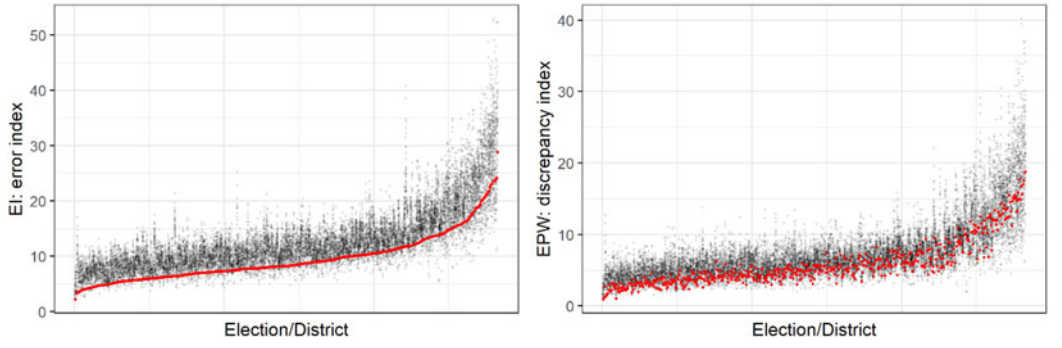

Figure 3. Estimated *EI* (left panel) and *EPW* (right panel) errors by election corresponding to the `ecolRxC` default solution (red points) and its linked *RC* solutions (black points) attained choosing as reference all the *RC* possible pairs with a row and a column. Elections have been ordered from smallest to largest *EI*.

than the average errors of the *RC* reference solutions. In other words, the error of the mean is smaller than the mean of the errors—overall, 9.50 versus 13.66 in terms of *EI* errors and using probit transformations. The issue is whether, conditioned to the election, this happens for all the reference solutions. That is, are all the reference solutions (almost) always worse than the ecolRxC solution? And, if not, is there any observable feature that permits identifying the reference options that beat the combined solution? The data presented in Figure 3 help answer these questions.

Figure 3 displays the estimated ecolRxC EI and EPW errors along with their corresponding *RC* reference errors for each election. The elections are ordered from smallest to largest ecolRxC *EI* errors, with the left panel showing the *EI* errors and the right panel the *EPW* errors.[14] Two clear patterns emerge from the figure. On the one hand, the ecolRxC solution systematically improves the majority of the reference solutions—on average, 89 percent of the time per election. On the other hand, for the majority of the elections (almost 76 percent), there is a (r_1, r_2)-reference solution with smaller error than the corresponding ecolRxC solution. In fact, if the (r_1, r_2)-reference solution with the smallest error were chosen in each election, the average EI error would decrease to 8.18.

The question, therefore, is whether the best reference solution of each election could be identified from the observed election features. Our answer is that this is not possible. Despite we are able to improve the average ecol solution by properly selecting (r_1, r_2) by, for instance, exploiting

[14] *EQ* errors have been omitted as they lead to similar conclusions.

the fact that the average correlations (across elections) between the *EI* (r_1, r_2)-reference errors and $|\hat{\rho}^{(0)}_{r_1 r_2}|$ and $v_{r_1 r_2}$ are −0.23 and −0.11, respectively, we did not find any pattern presented in the observed data which is able to improve ecolRxC solutions. For example, when either the cell with the highest (estimated) number of votes or the pair with the highest crude ecological correlation is utilized to decide the (r_1, r_2)-pair to be employed as reference, the *EI* average errors attained are 11.38 and 11.64, respectively—noticeably smaller than 13.81, but still clearly above 9.50. Similarly, if inspired by Thomson's recommendation of choosing abstainers as the reference party at each election (given its neutral and commonly large size) we assess ecol accuracy using the largest party and candidate as references, we again find that although better solutions can be attained by avoiding the extreme combinations, they still do not improve ecolRxC. For example, the smallest average error attained with this specification is 11.78, which is reached choosing a probit-transformation and the Yule approximation.

5.3.4 Comparing `ecolRxC` with `ei.MD.bayes` and `nslphom`

All the analyses performed point to the ecolRxC specification (`ecolRxC` default) as the best approach among the ones available in `ecolRxC`. The remaining question is how this approach compare to the other two algorithms, `ei.MD.bayes` and `nslphom`, previously identified in the literature as the most accurate (Klima *et al.*, 2016; Plescia and De Sio, 2018; Pavía and Romero, 2023). In other words, what is the relative performance of `ecolRxC` compared to the performances of `ei.MD.bayes` and `nslphom`? To answer this question, we compare the *EI* errors reported in Pavía and Romero (2023)—who analyze the same elections considered in this application except for the group of elections corresponding to New Zealand in 2020—with the *EI* errors we obtain here. For the 493 elections analyzed in Pavía and Romero (2023), `ecolRxC`, with default options, records an average *EI* error of 9.78, a figure quite similar to the numbers 10.52 and 9.77 reported in Pavía and Romero (2023) for `ei.MD.bayes` and `nslphom`, respectively. Our conclusion is therefore clear: `ecolRxC` shows an accuracy in line with those found for `ei.MD.bayes` and `nslphom` and, consequently, it deserves to be recognized as having a place among the best approaches for estimating R × C ecological inference tables.

6. Conclusions

The objective of ecological inference is to build "data" on the individual level from data on the aggregate ecological level. Within ecological inference, a particularly relevant challenge involves consistently filling the interior cells of a set of R × C contingency tables when only their margins are known. This is a particular instance of cross-level (ecological) inference that appears in many disciplines, including quantitative history, marketing and epidemiology, with political science and sociology being the areas where this challenge emerges more frequently.

Over time, many algorithms have been proposed to solve this problem from frameworks as diverse as mathematical linear programming, Bayesian and frequentist statistics, linear regression, or entropy theory. In our opinion, however, it is not enough to just construct models that provide accurate statistical solutions, the models also need to be well suited to substantial interpretations. The ecological inference models based on the latent structural theory fit this requirement, since the underlying latent factors can be estimated from the aggregate results. This paper extends and improves Thomsen's solution and describes a new `R`-package `ecolRxC` that permits accurate solutions to be obtained within this framework.

`ecolRxC` has not only programmed the previous versions of this methodology described in Thomsen (1987) and Park (2008), but it improves and extends them by offering new capabilities (for instance, the estimation of uncertainties or the automatic treatment of inconsistencies between margin aggregates), also yielding more accurate solutions. In the 565 real datasets analyzed in this paper, the overall average *EI* errors with Thomsen's and Park's algorithms

have been 13.81 (11.64 if the pair with largest ecological correlation had been used as reference with `scale = 'logit'`) and 11.38, respectively. These are at least 20 percent worse than the 9.50 *EI* average error recorded by `ecolRxC` with default options. Furthermore, `ecolRxC` also stands up against comparison with `ei.MD.bayes` and `nslphom`—the two algorithms currently identified in the literature as the most accurate. `ecolRxC` records accuracies in line with those found for `ei.MD.bayes` and `nslphom`.

In this paper, we have focused on assessing the accuracy of the new proposals, leaving other relevant issues for further investigation. On the one hand, despite the enormous computational burden involved, we consider that comparing the precision (estimated uncertainties) of the latent structure approach with their main competitors (`ei.MD.bayes` and `nslphom`) could provide valuable insights into their relative strengths and weaknesses. On the other hand, given the limited literature on the latent structure approach for ecological inference, we consider that exploring how the model's assumptions can be tested and the sensitivity of inferences to them presents an interesting avenue for future research.

Finally, it is worth to pointing out that, although our exposition has focused on the problem of estimating vote transfer matrices, `ecolRxC` could also be utilized to estimate other types of vote-related cross-distributions (such as social class and vote, race and vote, or age-gender group and vote) as well as other general cross-tables (such as caste and educational level, wealth and home ownership, or age-gender group and cultural consumption). Despite it not being possible to establish a supporting behavioral theory that impacts both categorizations for these examples, we are convinced that our implementation of the latent factor approach can effectively be employed on them. This confidence stems from its foundation in exploiting correlations among row and column categories. Certainly, although in almost all of the examples listed, one of the variables corresponds to an intrinsic characteristic of the individual, which should therefore be considered exogenous due to its factual nature, it is not hard to imagine the existence in all the examples of some latent dimensions, naturally associated with the factual variable, that impact the response variable in the same way as in the specified model.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2024.57. To obtain replication material for this article, https://doi.org/10.7910/DVN/KOZI2C

Data. The data used in this paper are publicly available on the R-package `ei.Datasets` (version 0.0.1-1) accessible on CRAN in the URL https://CRAN.R-project.org/package=ei.Datasets.

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

Financial support. Generalitat Valenciana (Consellería de Educación, Cultura, Universidades y Empleo), grant GVRTE/2023/4655315, and Ministerio de Ciencia e Innovación, grant PID2021-128228NB-I00.

Competing interests. None.

Code availability. The reproducible ad hoc R-code employed, which permits replicating the numbers and findings reported in the paper, is available at https://doi.org/10.7910/DVN/KOZI2C.

References

Achen C (2000) *The Thomsen Estimator for Ecological Inference*. Unpublished manuscript.

Andreadis I and Chadjipadelis T (2009) A method for the estimation of voter transition rates. *Journal of Elections, Public Opinion and Parties* **19**, 203–218.

Ansolabehere S, Snyder JM and Stewart C (2001) Candidate positioning in U.S. House elections. *American Journal of Political Science* **45**, 136–159.

Barreto M, Collingwood L, Garcia-Rios S and Oskooii KAR (2022) Estimating candidate support in voting rights act cases: comparing iterative EI and EI-R_C methods. *Sociological Methods & Research* **51**, 271–304.

Battista JC, Peress M and Richman J (2022) Estimating the locations of voters, politicians, policy outcomes, and status quos on a common scale. *Political Science Research and Methods* **10**, 806–822.

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

Choirat C, Honaker J, Imai K, King G and Lau O (2017) *Zelig: Everyone's Statistical Software* [Computer software]. Available at http://zeligproject.org/

Collingwood L, Oskooii K, Garcia-Rios S and Barreto M (2016) eiCompare: comparing ecological inference estimates across EI and EI:R×C. *The R Journal* **8**, 92–101.

Deming WE and Stephan FF (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* **11**, 427–444.

Downs A (1957) An economic theory of political action in a democracy. *Journal of Political Economy* **65**, 135–150.

Efron B and Tibshirani RJ (1994) *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC.

Forcina A and Pavía JM (2024) *eiCircles: Ecological Inference of RxC Tables by Overdispersed-Multinomial Models* (R package version 0.1-7) [Computer software]. Available at https://CRAN.R-project.org/package=eiCircles

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

Gampmayer M (2016) *vottrans: Voter Transition Analysis* (R package version 1.0) [Computer software]. Available at https://CRAN.R-project.org/package=vottrans

Goodman LA (1953) Ecological regressions and the behavior of individuals. *American Sociological Review* **18**, 663–664.

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

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 and Quinn KM (2009) R×C ecological inference: bounds, correlations, flexibility, and transparency of assumptions. *Journal of the Royal Statistical Society, Series A* **172**, 67–81.

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]. Available at https://CRAN.R-project.org/package=RxCEcolInf

Groseclose T (2001) A model of candidate location when one candidate has a valence advantage. *American Journal of Political Science* **45**, 862–886.

Imai K, Lu Y and Strauss A (2008) Bayesian and likelihood inference for 2x2 ecological tables: an incomplete data approach. *Political Analysis* **16**, 41–69.

Imai K, Lu Y and Strauss A (2011) eco: R package for ecological inference in 2x2 tables. *Journal of Statistical Software* **42**, 1–23.

Johnson NL and Kotz S (1972) *Distributions in Statistics: Continuous Multivariate Distributions*. New York: Wiley.

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

King G and Roberts M (2016) *ei: Ecological Inference* (R package version 1.3-3) [Computer software]. Available at https://CRAN.R-project.org/package=ei

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

Lau O, Moore ORT and Kellermann M (2023) *eiPack: Ecological Inference and Higher-Dimension Data Management* (R package version 0.2-2) [Computer software]. Available at https://CRAN.R-project.org/package=eiPack

Manski CF (2007) *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

Martin AD, Quinn KM and Park JH (2011) MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software* **42**, 1–21.

Park W-h (2002) *VTR and Ecoline (Version 1.0)* [Computer software]. Ann Arbor, MI: University of Michigan.

Park W-h (2008) Ecological Inference and Aggregate Analysis of Elections (PhD dissertation). The University of Michigan.

Park W-h, 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) Integer estimation of inner-cell values in RxC ecological tables. *Bulletin of Sociological Methodology*, forthcoming.

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 (2024a) Improving estimates accuracy of voter transitions. Two new algorithms for ecological inference based on linear programming. *Sociological Methods & Research* **53**, 1491–1533.

Pavía JM and Romero R (2024b) Symmetry estimating RxC vote transfer matrices from aggregate data. *Journal of the Royal Statistical Society – Series A*, online available. https://doi.org/10.1093/jrsssa/qnae013

Pavía JM and Romero R (2024c) *lphom: Ecological Inference by Linear Programming under Homogeneity* (R package version 0.3.5-5) [Computer software]. Available at https://CRAN.R-project.org/package=lphom

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 cluster analysis of vote transitions. *Computational Statistics and Data Analysis* **70**, 328–344.

Rabinowitz G and Macdonald SE (1989) A directional theory of issue voting. *American Political Science Review* **83**, 77–91.

R Core Team (2023) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, Available at https://www.R-project.org/

Robinson WS (1950) Ecological correlations and the behavior of individuals. *American Sociological Review* **15**, 351–357.

Rosen O, Jiang W, King G and Tanner MA (2001) Bayesian and frequentist inference for ecological inference: the RxC case. *Statistica Neerlandica* **55**, 134–156.

Sanders D, Clark HD, Stewart MC and Whiteley P (2011) Downs, Stokes and the dynamics of electoral choice. *British Journal of Political Science* **41**, 287–314.

Schmitt H, Segatti P and van der Ejik C (2021) *Consequences of Context: How Social, Political and Economic Environments Affects Voting*. London: ECPR Press.

Siegumfeldt F (2004) *User's Guide to Ecol for Stata*. Aarhus: Aarhus University.

Thomsen SR (1987) *Danish Elections, 1920–79: A Logit Approach to Ecological Analysis and Inference*. Aarhus: Politica.

Thomsen SR (2011) The cultural component in voting behaviour. Paper presented in the 2009 Annual Meeting of the Mid-West Political Science Association. Available at https://rb.gy/8fhyvt

Thomsen SR, Berglund S and Wörlund I (1991) Assessing the validity of the logit method for ecological inference. *European Journal of Political Research* **19**, 441–477.

Thomsen SR, Frandsen AG, Kristmar T, Lauristsen P and Sørensen MB (1995) *Ecol (Version 3)* [Computer software]. Aarhus: Aarhus University.

Train KE (2009) *Discrete Choice Methods with Simulation*. New York: Cambridge University Press.

Tziafetas G (1986) Estimation of the voter transition matrix. *Optimization* **17**, 275–279.

Cite this article: Pavía JM and Thomsen SR (2024) `ecolRxC`: Ecological inference estimation of R × C tables using latent structure approaches. *Political Science Research and Methods* 1–19. https://doi.org/10.1017/psrm.2024.57