# Ecological Inference and Higher-Dimension Data Management

*Olivia Lau*[1] *Ryan T. Moore*[2] *Mike Kellermann*[3]

## Abstract

Ecological inference (EI) takes contingency tables as the unit of analysis. These tables are described by marginal row and column totals (or proportions); the goal of inference is to determine the joint intra-table relationship between the rows and columns. Using a common example from political science, let the unit of analysis be a voting precinct in a state-wide election:

| | Democrat | Republican | No Vote | Total |
|---|---|---|---|---|
| Black | ? | ? | ? | 423 |
| White | ? | ? | ? | 219 |
| Hispanic | ? | ? | ? | 14 |
| Total | 317 | 156 | 183 | 656 |

For a given election, we estimate each cell (e.g., the number of Blacks who voted for the Democratic candidate) for each precinct $i = 1, \ldots, I$, then aggregate across precincts to obtain election-wide results.

While some existing R packages (`MCMCpack` by Andrew Martin and Kevin Quinn and `eco` by Kosuke Imai and Ying Lu, for example) offer functions that analyze $2 \times 2$ models, we implement more general methods that can take more than two rows or columns. Our package will include:

- Extreme case analysis, or the method of bounds, suggested by Duncan and Davis (1953).
- Ecological regression described in Goodman (1953) using both frequentist point estimates and a Bayesian estimator that produces correct standard errors.
- R $\times$ C model described in Rosen et al. (2001) using three estimators: a Bayesian Markov-chain Monte Carlo algorithm, maximum likelihood, and penalized least squares.

Since the unit of analysis (each ecological table) is a matrix rather than a vector, studying the statistical problem of ecological inference requires computational innovation in higher-dimension data management. For example, in the case of the Bayesian R $\times$ C estimator, we need to keep track of an array of dimension:

$$\text{rows} \times \text{columns} \times \text{precincts} \times \text{simulations}$$

In typical electoral data, there may be four rows (Black, White, Hispanic, Other), three columns (Democrat, Republican, No Vote), and 11,366 precincts (in the case of Ohio in 2004), for a total of 136,392 cell parameters (about 1 GB of memory) *per simulation.* It is thus impossible to store every simulation, or even a substantially thinned number of simulations, without several terabytes of memory (supposing that R could handle that much). We propose to deal with this memory management issue for higher-dimension data in several ways:

- For each iteration (or every iteration saved), the user may specify a quantity of interest to be calculated and stored (rather than the parameter draws themselves).
- Rather than storing draws in the workspace, the Bayesian methods will have the option to `sink` draws to a file. Since these multi-dimensional data need to be formatted in two dimensions for disk storage, we will provide functions to reconstruct the higher dimensions upon reading the sunk file.

In addition, this package will provide wrapper functions to operate on the margins of higher-dimension arrays, providing useful summary and print functions.

[1]Ph.D. Candidate, Department of Government, and M.A. student, Department of Statistics, Harvard University, `olau@fas.harvard.edu`. An alpha version of this software may be found at `http://www.fas.harvard.edu/~olau/software/`.

[2]Ph.D. Candidate, Department of Government, and M.A. student, Department of Statistics, Harvard University, `rtmoore@fas.harvard.edu`

[3]Ph.D. Candidate, Department of Government, Harvard University, `kellerm@fas.harvard.edu`

## References


Duncan, O. D. and Davis, B. (1953), "An Alternative to Ecological Correlation," *American Sociological Review*, 18, 665–666.

Goodman, L. (1953), "Ecological Regressions and the Behavior of Individuals," *American Sociological Review*, 18, 663–666.

Imai, K. and Lu, Y. (2005), *eco: R Package for Fitting Bayesian Models of Ecological Inference in 2x2 Tables*, R package version 2.2-1.

Martin, A. D., , and Quinn, K. M. (2005), *MCMCpack: Markov chain Monte Carlo (MCMC) Package*, R package version 0.7-1.

Rosen, O., Jiang, W., King, G., and Tanner, M. A. (2001), "Bayesian and Frequentist Inference for Ecological Inference: The $R \times C$ Case," *Statistica Neerlandica*, 55, 134–156, http://gking.harvard.edu/files/abs/rosen-abs.shtml.