PyEI: A Python package for ecological inference

Karin C. Knudson[1], Gabe Schoenbach[2], and Amariah Becker[2]

1 Data Intensive Studies Center, Tufts University **2** MGGG Redistricting Lab, Tufts University

Summary

An important question in some voting rights and redistricting litigation in the U.S. is whether and to what degree voting is racially polarized. In the setting of voting rights cases, there is a family of methods called "ecological inference" (see especially King, 1997) that uses observed data, pairing voting outcomes with demographic information for each precinct in a given polity, to infer voting patterns for each demographic group.

More generally, we can think of ecological inference as seeking to use knowledge about the margins of a set of tables (Table 1) to infer associations between the row and column variables, by making (typically probabilistic) assumptions. In the context of assessing racially polarized voting, a table like the one in Table 1 will correspond to a precinct, where each column corresponds to a candidate or voting outcome and each row to a racial group. Ecological inference methods then use the vote counts and demographic data for each precinct to make inferences about the overall voting preferences by demographic group, thus addressing questions like: "What percentage of East Asian voters voted for Hardy?" This example is an instance of what is referred to in the literature as "R by C" ecological inference, where here we have R = 2 groups and C = 3 voting outcomes. PyEI was created to support performing ecological inference with voting data; however, ecological inference methods also applicable in other fields, such as epidemiology (Elliot et al., 2000) and sociology (Goodman, 1953).

Table 1: In ecological inference we have information about the marginal counts for a set of tables like the one here and would like to make inferences about, for example, the number or proportion of East Asian voters who voted for Hardy. The system is underdetermined and ecological inference methods proceed by making statistical assumptions.

	Hardy	Kolstad	Nadeem	
East Asian	?	?	?	Total East Asian
non- East Asian	?	?	?	Total non- East Asian
	Total for Hardy	Total for Kolstad	Total for Nadeem	

Statement of need

The results of ecological inference for inferring racially polarized voting are routinely used in US voting rights cases (King, 1997); therefore, easy to use and high quality tools for performing ecological inference are of practical interest. There is a need for an ecological inference library that brings together a variety of ecological inference methods in one place to facilitate crucial tasks such as: quantifying the uncertainty associated with ecological inference results under a given model; making comparisons between methods; and bringing relevant diagnostic tools to bear on ecological inference methods. To address this need, we introduce PyEI, a Python package for ecological inference.

PyEI is meant to be useful to two main groups of researchers. First, it serves application-oriented researchers and practitioners who seek to run ecological inference on domain data (e.g., voting data), report the results, and understand the uncertainty related to those results. Second, it facilitates exploration and benchmarking for researchers who are seeking to understand properties of existing ecological inference methods in different settings and/or develop new statistical methods for ecological inference.

PyEI brings together the following ecological inference methods in a common framework alongside plotting, reporting, and diagnostic tools:

- Goodman's ecological regression (Goodman, 1953) and a Bayesian linear regression variant
- A truncated-normal based approach (King, 1997)
- Binomial-Beta hierarchical models (King et al., 1999)
- Dirichlet-Multinomial hierarchical models (Rosen et al., 2001)
- A Bayesian hierarchical method for 2×2 EI following the approach of Wakefield (2004)

In several of these cases, PyEI includes modifications to the models as originally proposed in the cited literature, such as reparametrizations or other changes to upper levels of the hierarchical models in order to ease sampling difficulties.

PyEI is intended to be easily extensible, so that additional methods from the literature can continue to be incorporated (for example, work is underway to add the method of James Greiner & Quinn (2009), currently implemented in the R package RxCEcolInf (Greiner et al., 2019)). Newly developed statistical methods for ecological inference can be included and conveniently compared with existing methods.

Several R libraries implementing different ecological inference methods exist, such as eiPack (Lau et al., 2020), RxCEcolInf (Greiner et al., 2019), ei (King & Roberts, 2016), and eiCompare (Collingwood et al., 2020). In addition to presenting a Python-based option that researchers who primarily use Python may appreciate, PyEI incorporates the following key features and characteristics.

First, the Bayesian hierarchical methods implemented in PyEI rest on modern probabilistic programming tooling (Salvatier et al., 2016) and gradient-based MCMC methods such as the No U-Turn Sampler (NUTS) (Betancourt, 2018; Hoffman & Gelman, 2014). Using NUTS where possible should allow for faster convergence than existing implementations that rest primarily on Metropolis-Hastings and Gibbs sampling steps. Consider effective sample size, which is a measure of how the variance of a Monte Carlo estimate of a posterior expectation computed from dependent samples compares to the variance of the corresponding estimate computed from independent samples from the posterior distribution (or, very roughly, how "effective" the samples are for estimating a posterior expectation, compared to independent samples) (Gelman et al., 2013). Under certain assumptions on the target posterior distribution, in Metropolis-Hastings the number of evaluations of the log-posterior required for a given effective sample size scales linearly with the dimensionality of the parameter space, while in Hamiltonian Monte Carlo approaches such as NUTS, the number of required evaluations of the gradient of the log-posterior scales only as the fourth root of the dimension (Neal, 2011). Reasonable scaling with the dimensionality of the parameter space is important in ecological inference, as that dimensionality is large when there are many precincts.

Second, integration with the existing tools PyMC3 (Salvatier et al., 2016) and ArviZ (Kumar et al., 2019) makes the results amenable to state of the art diagnostics (e.g. convergence diagnostics) and some reasonable checks are automatically performed.

Third, summary and plotting utilities for reporting, visualizing, and comparing results are included (e.g. Figure 1, Figure 2), with an emphasis on visualizations and reports that clarify the uncertainty of estimates under a model.

Lastly, clear documentation is provided, including a set of introductory and example notebooks.
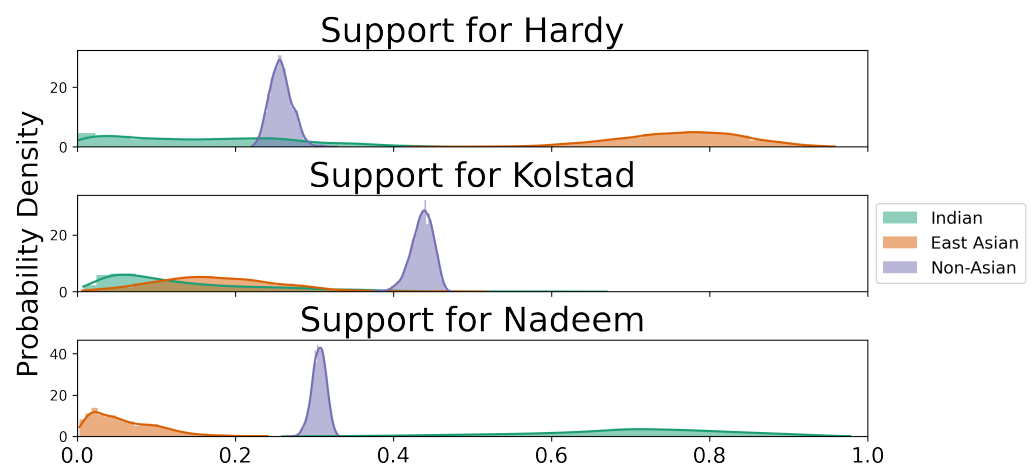

Figure 1: Kernel density estimation plots for visualizing uncertainty of support for candidates within each group.
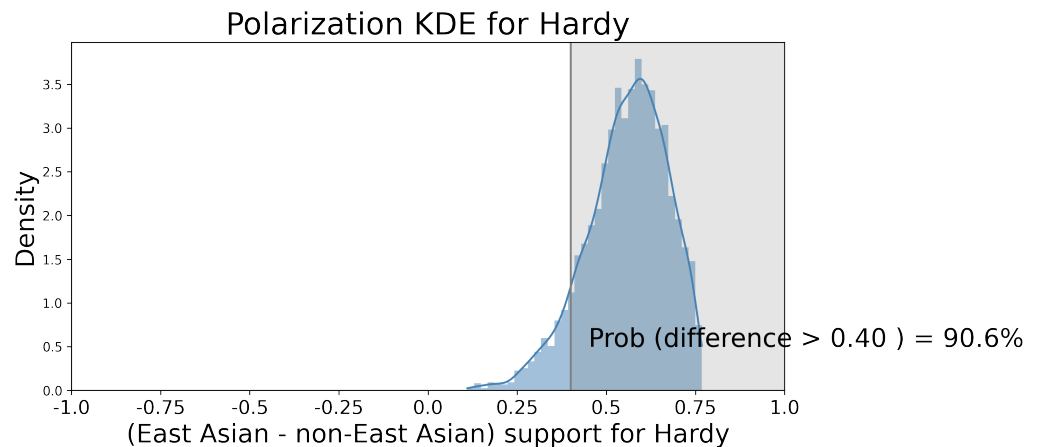
Figure 2: Visualizing and quantifying degree of polarization.

Acknowledgments

This software development is part of a research project comparing methods, joint with Moon Duchin and Thomas Weighill. We thank Colin Carroll, JN Matthews, and Matthew Sun for their helpful contributions to PyEI.

References

Betancourt, M. (2018). *A conceptual introduction to hamiltonian monte carlo*. http://arxiv.org/abs/1701.02434

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

Elliot, P., Wakefield, J. C., Best, N. G., Briggs, D. J., & others. (2000). *Spatial epidemiology: Methods and applications*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198515326.001.0001

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press. https://doi.org/10.1201/b16018

Goodman, L. A. (1953). Ecological regressions and behavior of individuals. *American Sociological Review*. https://doi.org/10.2307/2088121

Greiner, D. J., Baines, P., & Quinn, K. M. (2019). *RxCEcolInf: 'R x C ecological inference with optional incorporation of survey information'*. https://CRAN.R-project.org/package=RxCEcolInf

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623. http://jmlr.org/papers/v15/hoffman14a.html

James Greiner, D., & Quinn, K. M. (2009). R× C ecological inference: Bounds, correlations, flexibility and transparency of assumptions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 67–81. https://doi.org/10.1111/j.1467-985X.2008.00551.x

King, G. (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press. https://doi.org/10.1515/9781400849208

King, G., & Roberts, M. (2016). *ei: Ecological inference*. https://CRAN.R-project.org/package=ei

King, G., Rosen, O., & Tanner, M. A. (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research*, *28*(1), 61–90. https://doi.org/10.1177/0049124199028001004

Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, *4*(33), 1143. https://doi.org/10.21105/joss.01143

Lau, O., Moore, R. T., & Kellermann, M. (2020). *eiPack: Ecological inference and higher-dimension data management*. https://CRAN.R-project.org/package=eiPack

Neal, R. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, *2*(11), 2. https://doi.org/10.1201/b10905-7

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

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55. https://doi.org/10.7717/peerj-cs.55

Wakefield, J. (2004). Ecological inference for 2× 2 tables. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *167*(3), 385–425. https://doi.org/10.1111/j.1467-985x.2004.02046.x