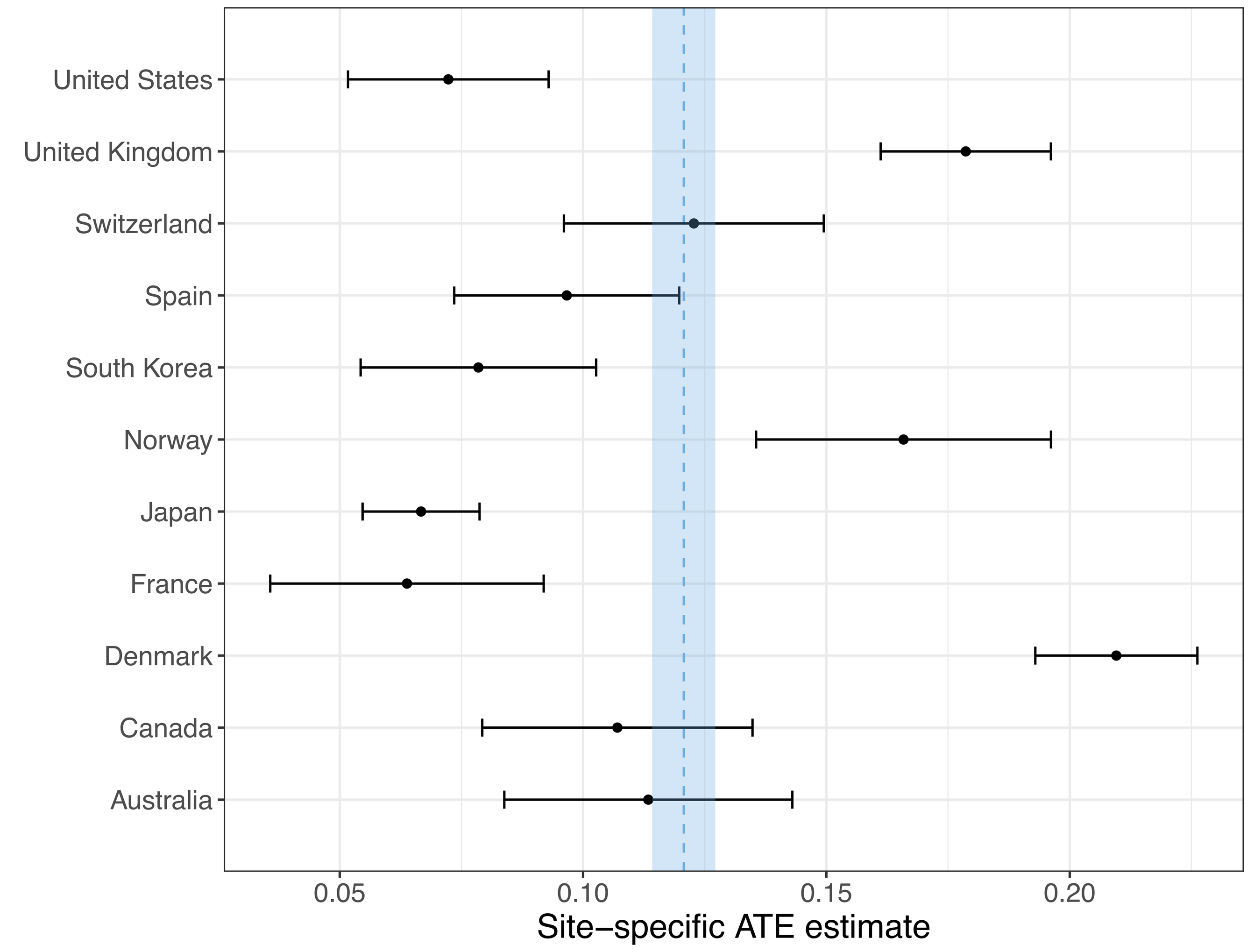
Decomposing Treatment Effect Heterogeneity in Multisite Experiments

Martin Devaux (PhD Candidate, Columbia University)

Problem

What explains the variation across sites in a multisite experiment?



Example from: Valentino, N. A., Soroka, S. N., Iyengar, S., Aalberg, T., Duch, R., Fraile, M., Hahn, K. S., Hansen, K. M., Harell, A., Helbling, M., Jackman, S. D., & Kobayashi, T. (2019). Economic and Cultural Drivers of Immigrant Support Worldwide. *British Journal of Political Science*, *49*(4), 1201–1226.

Contribution

Existing methods:

- Meta-analysis: is there between-study variance, net of sampling variation?
- Meta-regressions: does variation in specific site-level covariates correlate with cross-site variation?
- Reweighting: Lu et al. (2023) - how much variation would remain if all sites had the same distribution of observed unit-level covariates?
  Lu, B., Ben-Michael, E., Feller, A., & Miratrix, L. (2023). Is It Who You Are or Where You Are? Accounting for Compositional Differences in Cross-Site Treatment Effect Variation. *Journal of Educational and Behavioral Statistics*, *48*(4), 420–453.

New question: based on observed covariates, do site-level or unit-level features explain more of the heterogeneity?

Quantity of interest: how much variation would remain if covariates at one level were held identical in expectation across sites?

Helps answer:

1. Are population or context differences driving the heterogeneity?
2. How do the modeled unit- and site-explained variations compare to the total systematic heterogeneity?

Formal Setup

- Outcome Y, treatment T, sites $\{1, \ldots, K\}$
- Observed unit-level covariates X, observed site-level covariates M
- Unobserved unit-level covariates U_x, site-level covariates U_m
- CATE on observed covariates:

$$\tau(\mathrm{X,M}) = \mathbb{E}_{U_x,U_m}[\tau(\mathrm{X,\ M},U_x,\ U_m) \mid \mathrm{X,\ M}]$$

Estimands:

$$\tau^2_{site} = \mathrm{Var}(\mathbb{E}[\tau(\mathrm{X,M}) \mid X = \tilde{x}, M])$$
$$\tau^2_{unit} = \mathrm{Var}(\mathbb{E}[\tau(\mathrm{X,M}) \mid X, M = \tilde{m}])$$

Key assumption: conditional cross-level independence

$$U_x \perp M \mid X, M \text{ and } U_m \perp X \mid X, M$$

Under this assumption, while τ^2_{site} and τ^2_{unit} might also capture variation explained by unobserved covariates, they do not inadvertently capture variation explained by the other level.

Estimation

Algorithm - estimation of τ^2_{site} and τ^2_{unit}

Input: pooled.data (experimental data pooled across sites)
Output: τ^2_{site}, τ^2_{unit}

1. M, X, Y,T ← pooled.data[M], pooled.data[X], pooled.data[Y], pooled.data[T];
2. $\tau(\cdot)$ ← outcome model estimated using M, X, Y, T ;

Shuffle covariates at each level
3. data.site ← sample(X), M, Y;
4. data.unit ← X, sample(M), Y;

Predict potential outcomes
5. data.site[$\tilde{Y}_1$, $\tilde{Y}_0$] ← τ(data.site);
6. data.unit[$\tilde{Y}_1$, $\tilde{Y}_0$] ← τ(data.unit);

Estimate site average treatment effects
7. **for** site j **do**
   8 | $\widehat{ATE}_{site,j}$ ← mean(data.site[$\tilde{Y}_1$] – data.site[$\tilde{Y}_0$]);
   9 | $\widehat{ATE}_{unit\ j}$ ← mean(data.unit[$\tilde{Y}_1$] – data.unit[$\tilde{Y}_0$]);
10. **end**

Estimate cross-site variances
11. $\hat{\tau}^2_{site}$ ← estimated between−site variance of $\widehat{ATE}_{site}$;
12. $\hat{\tau}^2_{unit}$ ← estimated between−site variance of $\widehat{ATE}_{unit}$;

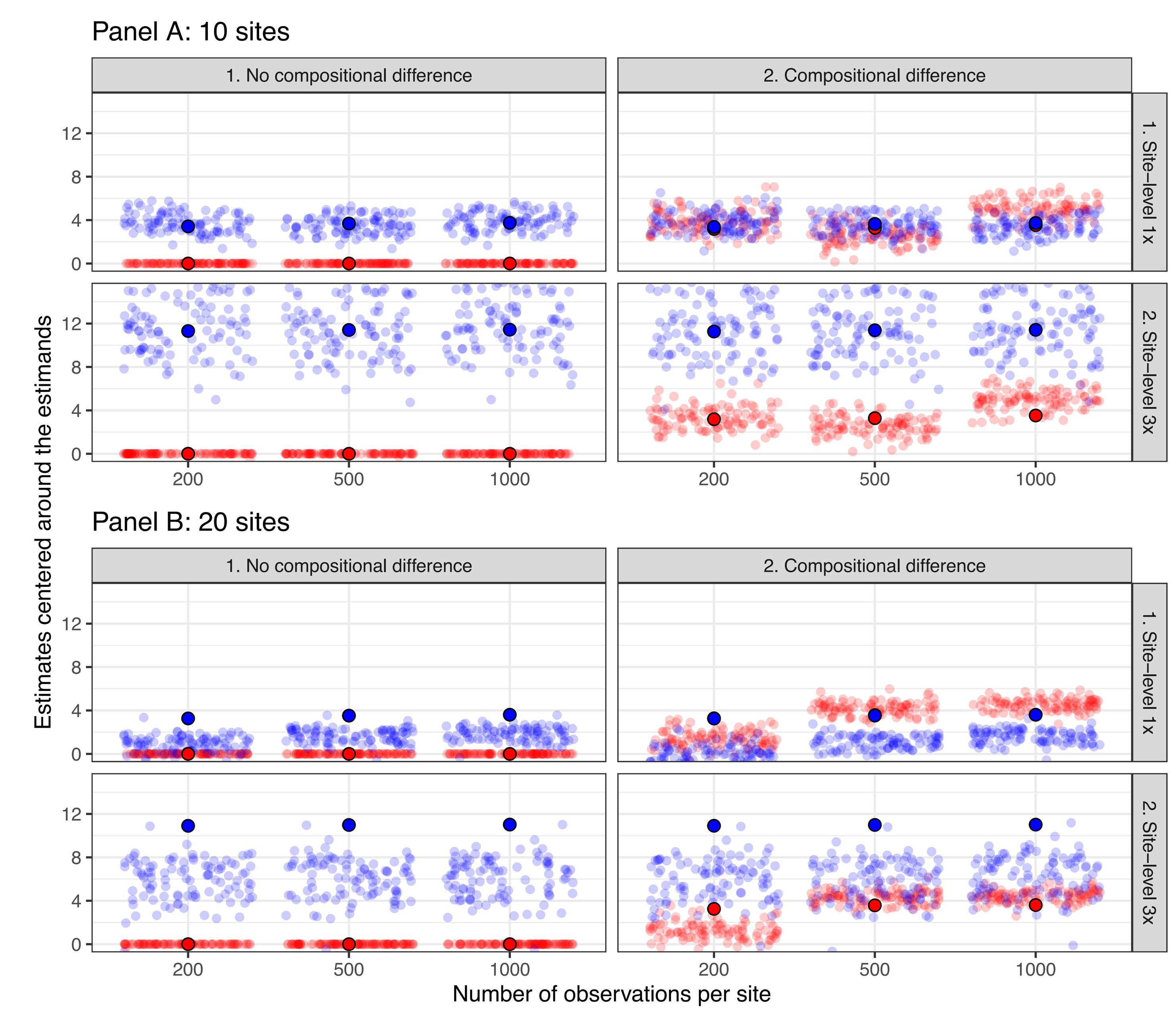
Permutation test - difference between $\hat{\tau}^2_{site}$ and $\hat{\tau}^2_{unit}$:

The site-level datasets $\widehat{ATE}_{site}$ and $\widehat{ATE}_{unit}$ are combined. The "site" and "unit" labels are randomly permuted. Each time, $\hat{\tau}^2_{site}$ - $\hat{\tau}^2_{unit}$ is computed, providing a distribution under the null of no difference.

Simulation

With different coefficients, compositional differences, number of observations, and number of sites. Outcome model selected: BART

1. How do the estimands (τ^2_{site} and τ^2_{unit}) behave?
2. How well does the estimator recover them?



With 20 sites, $\hat{\tau}^2_{site}$ is consistently underestimated.

Tentative reason: data suggests that with more sites, BART gives less importance to the unit-level covariates and might not be adapted to the nested structure of the data.

Tentative solution: use a multilevel outcome model.

Application

STAR experiment (Tennessee, 1985-1989): Does reducing class-size lead to better educational outcomes?

- Reports focus on how site-level (e.g. inner-city versus rural) moderate the treatment effect.
- The decomposition suggests that $\hat{\tau}^2_{unit}$ is actually three times as large as $\hat{\tau}^2_{site}$ and significantly so.
- **Limitation**: can composition really be distinguished from context?

Word, E., Johnston, J., Pate Bain, H., DeWayne Fulton, B., Boyd Zaharias, J., Achilles, C., Nannette Lintz, M., Folger, J., & Breda, C. (1990). *The State Of Tennessee's Student/Teacher Achievement Ratio (Star) Project, Technical Report*. Tennessee State Department of Education.