Post Hoc Synthetic Purposive Sampling for Post Hoc External Validity Assessment

Simon Hoellerbauer (UMass Amherst), Isabel Laterzo-Tingley (UT Austin)

Research Objectives

- Can the structure of synthetic purposive sampling (SPS) be adapted for post hoc evaluation of the generalizability of samples when researchers have no site selection ability?

SPS

- Introduced by [1] for use during design stage of studies to select optimal sites for multi-site studies
- Minimize

$$f(\mathbf{S},\mathbf{W}) = \frac{1}{N-N_S}\sum_{k=1}^{N}(1-S_k)\left(\frac{1}{L}\sum_{l=1}^{L}B_{kl}(\mathbf{S},\mathbf{W})\right) \qquad (1)$$

over $\mathbf{S}$ and $\mathbf{W}$, where
$B_{kl}(\mathbf{S},\mathbf{W}) = (X_{kl} - \sum_{j:S_j=1} W_{jk}X_{jl})^2$

Post Hoc SPS

- Post hoc SPS: compare SPS-optimal site selection against actual site selection using objective function:

$$\frac{f(\mathbf{S}=\mathbf{S}_{\text{actual}}, \mathbf{W}=\mathbf{W}^*_{\mathbf{S}_{\text{actual}}})}{f(\mathbf{S}=\mathbf{S}^*, \mathbf{W}=\mathbf{W}^*_{\text{overall}})} \qquad (2)$$

or

$$\frac{f(\mathbf{S}=\mathbf{S}_{\text{actual}}, \mathbf{W}=\mathbf{W}^*_{\text{overall}})}{f(\mathbf{S}=\mathbf{S}^*, \mathbf{W}=\mathbf{W}^*_{\text{overall}})} \qquad (3)$$

where $\mathbf{W}^*_{\text{overall}}$ are the naive SPS-optimal weights and $\mathbf{W}^*_{\mathbf{S}_{\text{actual}}}$ are the SPS-optimal weights when $\mathbf{S}$ is constrained to $\mathbf{S}_{\text{actual}}$
- Definition of target population is **key**: can identify all possible site selections

Example: Naumann et al. 2018

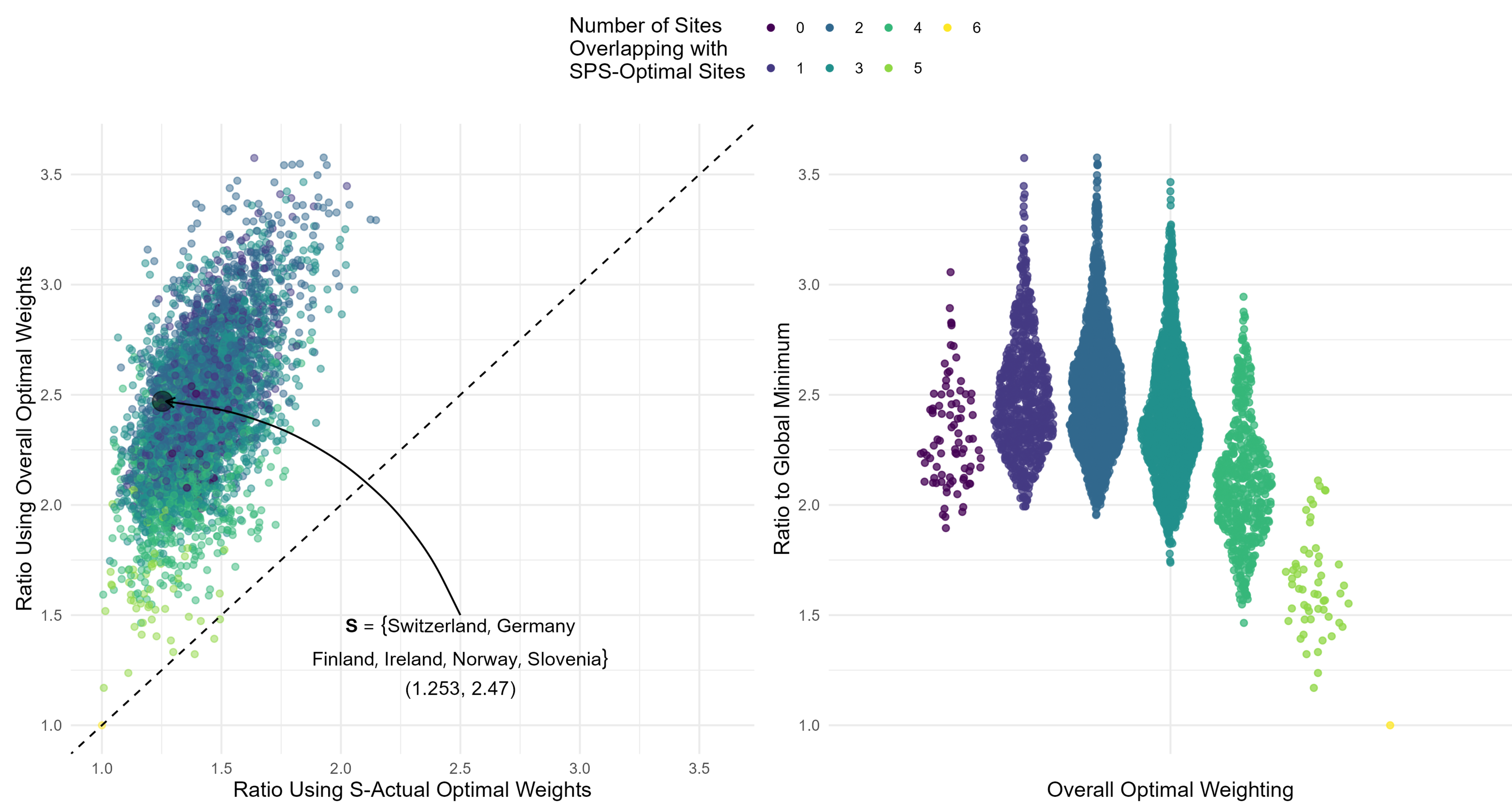
- Investigated attitudes towards immigration in 15 countries *in Europe*
- Austria, Belgium, **Switzerland**, Czechia, **Germany**, Denmark, Spain, **Finland**, France, United Kingdom, **Ireland**, Netherlands, **Norway**, **Slovenia**, Sweden → *target population*
- **Bolded** countries: Randomly chosen site selection from all 5005 possible 6-site selections

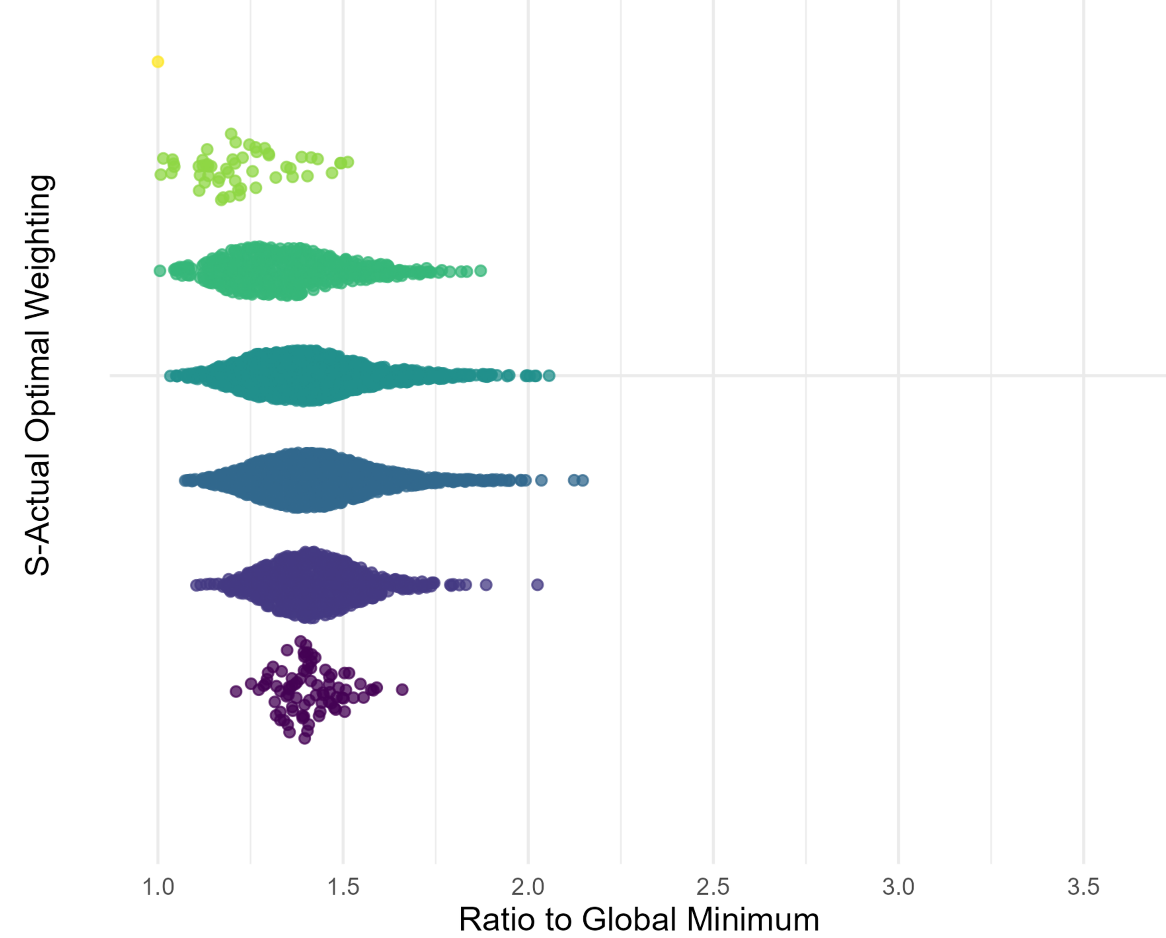
Illustrating Post Hoc SPS

How well can we generalize to target population – i.e. approximate non-selected sites – using Switzerland, Germany, Finland, Ireland, Norway, and Slovenia?

Because we know target population (all 15 countries), we can also compare against all possible site selections → allows us to evaluate relatively how well a site selection minimizes imbalance in non-selected sites.

$\frac{f(\mathbf{S}=\mathbf{S}_{\text{actual}},\mathbf{W}=\mathbf{W}^*_{\mathbf{S}_{\text{actual}}})}{f(\mathbf{S}=\mathbf{S}^*,\mathbf{W}=\mathbf{W}^*_{\text{overall}})}$ vs $\frac{f(\mathbf{S}=\mathbf{S}_{\text{actual}},\mathbf{W}=\mathbf{W}^*_{\text{overall}})}{f(\mathbf{S}=\mathbf{S}^*,\mathbf{W}=\mathbf{W}^*_{\text{overall}})}$ for all 5005 possible 6-site selections





Comparing Optimal Naumann et al. 2018 Site Selection to Randomly Chosen Site Selection

SPS Optimal	Site Overlap w/ Optimal	Sites Selected	Objective Function Value (W-Overall)	Objective Function Value (W-S-Actual)	Percentile Rank (Overall)	Percentile Rank (S-Actual)
Yes	6	Switzerland, Czechia, Germany, Denmark, Spain, Netherlands	0.944	0.944	1.000	1.000
No	2	Switzerland, Germany, Finland, Ireland, Norway, Slovenia	2.332	1.183	0.415	0.852

Conclusions

- Can use SPS post hoc to compare a set of sites for a study against the optimal set of sites
- Using $\mathbf{S}_{\text{actual}}$-optimized weights (3) is better measurement of minimum imbalance in a site selection than using the overall SPS weights

Next Steps

- For [2], compare ATE estimates for suboptimal site selections
- Illustrate comparing *between* studies
- Optimize generalizable population

Optimizing Generalizable Population

- Invert SPS optimization to choose the set of non-selected sites to which we can generalize with fixed selected sites
- Fix $\mathbf{S} = \mathbf{S}_{\text{Actual}}$; Let $\mathbf{M} = (M_1, M_2, \ldots, M_N)$, where $M_j = 1$ if a site is included in the generalizable population
- $$\min_{\mathbf{m},\mathbf{W}} \frac{1}{N_M - N_S}\sum_{k:M_k=1}(1-S_k)\left(\frac{1}{L}\sum_{l=1}^{L}B_{kl}(\mathbf{S},\mathbf{W})\right) \qquad (4)$$
- Can constrain N_M to be in a certain range
- Can add $\lambda * N_M$ or $\lambda * \frac{1}{N_M}$ as penalty terms to favor smaller and larger generalizable populations, respectively

Email: hoellerbauers@gmail.com
Website: `https://hoellers.github.io`

References

[1] Naoki Egami and Diana Da In Lee. Designing multi-site studies for external validity: Site selection via synthetic purposive sampling. Working Paper, 2024.

[2] Elias Naumann, Lukas F. Stoetzer, and Giuseppe Pietrantuono. Attitudes towards highly skilled and low-skilled immigration in Europe: A survey experiment in 15 European countries. *European Journal of Political Research*, 57(4):1009–1030, 2018.