Designing Multi-Site Studies for External Validity[1]

Naoki Egami

MIT

New Horizons on Model Transportability and Data Integration
IMSI
June 25, 2026

[1] Funded by the National Science Foundation (SES–2318659). Forthcoming at AJPS.

Internal Validity and External Validity

Internal Validity and External Validity

- Credibility Revolution

Internal Validity and External Validity

- Credibility Revolution
  - Randomized experiments & Observational identification strategies (IV, RDD, DID, ...)

Internal Validity and External Validity

- Credibility Revolution
  - Randomized experiments & Observational identification strategies (IV, RDD, DID, ...)
  - Internal Validity: Unbiased estimates of causal effects within a study

Internal Validity and External Validity

- Credibility Revolution
  - Randomized experiments & Observational identification strategies (IV, RDD, DID, ...)
  - Internal Validity: Unbiased estimates of causal effects within a study
- **External Validity**: Can we generalize causal findings beyond a specific study?

Internal Validity and External Validity

- Credibility Revolution
  - Randomized experiments & Observational identification strategies (IV, RDD, DID, ...)
  - Internal Validity: Unbiased estimates of causal effects within a study
- **External Validity**: Can we generalize causal findings beyond a specific study?
  - Social scientists: Interested in general theories of politics, institutions, and behavior

Internal Validity and External Validity

- Credibility Revolution
  - Randomized experiments & Observational identification strategies (IV, RDD, DID, ...)
  - Internal Validity: Unbiased estimates of causal effects within a study
- **External Validity**: Can we generalize causal findings beyond a specific study?
  - Social scientists: Interested in general theories of politics, institutions, and behavior
  - Policymakers: Interested in scaling up policies to broader populations and contexts

Internal Validity and External Validity

- Credibility Revolution
  - Randomized experiments & Observational identification strategies (IV, RDD, DID, ...)
  - Internal Validity: Unbiased estimates of causal effects within a study
- **External Validity**: Can we generalize causal findings beyond a specific study?
  - Social scientists: Interested in general theories of politics, institutions, and behavior
  - Policymakers: Interested in scaling up policies to broader populations and contexts
  - Long-standing question: at least goes back to the 1960s (Campbell & Stanley, 1963)

Internal Validity and External Validity

- Credibility Revolution
  - Randomized experiments & Observational identification strategies (IV, RDD, DID, ...)
  - Internal Validity: Unbiased estimates of causal effects within a study
- **External Validity**: Can we generalize causal findings beyond a specific study?
  - Social scientists: Interested in general theories of politics, institutions, and behavior
  - Policymakers: Interested in scaling up policies to broader populations and contexts
  - Long-standing question: at least goes back to the 1960s (Campbell & Stanley, 1963)
- **Challenge**: Few studies explicitly address external validity

Framework for External Validity

Framework for External Validity

Four Dimensions of External Validity (Egami & Hartman. 2023. *APSR*).

Framework for External Validity

Four Dimensions of External Validity (Egami & Hartman. 2023. *APSR*).

1. X**-validity: Population**
   - Convenience samples, Non-representative samples

Framework for External Validity

Four Dimensions of External Validity (Egami & Hartman. 2023. *APSR*).

1. X-**validity: Population**
   - Convenience samples, Non-representative samples
2. T-**validity: Treatments**
   - Realistic treatments, Differences in implementations, ...

Framework for External Validity

Four Dimensions of External Validity (Egami & Hartman. 2023. *APSR*).

1. X-**validity: Population**
   - Convenience samples, Non-representative samples
2. T-**validity: Treatments**
   - Realistic treatments, Differences in implementations, ...
3. Y-**validity: Outcome measures**
   - Behavioral or survey outcomes, Short- vs Long-term outcomes, ...

Framework for External Validity

Four Dimensions of External Validity (Egami & Hartman. 2023. *APSR*).

1. X-**validity: Population**
   - Convenience samples, Non-representative samples
2. T-**validity: Treatments**
   - Realistic treatments, Differences in implementations, ...
3. Y-**validity: Outcome measures**
   - Behavioral or survey outcomes, Short- vs Long-term outcomes, ...
4. C-**validity: Contexts**
   - Mechanisms differ across geography, time, political institutions, markets, ...

Multi-Site Studies for External Validity about "Contexts"

- **Multi-Site Causal Studies**: Conduct causal studies in multiple sites

Multi-Site Studies for External Validity about “Contexts”

- **Multi-Site Causal Studies**: Conduct causal studies in multiple sites
  - Example: survey experiments on immigration in the US, UK, and Germany

Multi-Site Studies for External Validity about "Contexts"

- **Multi-Site Causal Studies**: Conduct causal studies in multiple sites
  - Example: survey experiments on immigration in the US, UK, and Germany
  - Experimental: Metaketa (EGAP), JPAL,

Multi-Site Studies for External Validity about "Contexts"

- **Multi-Site Causal Studies**: Conduct causal studies in multiple sites
  - Example: survey experiments on immigration in the US, UK, and Germany
  - Experimental: Metaketa (EGAP), JPAL, Tomz & Weeks (2013), Carnes & Lupu (2016), Kalla & Broockman (2018), Bassan-Nygate et al (2023), Arechar et al (2024), ...

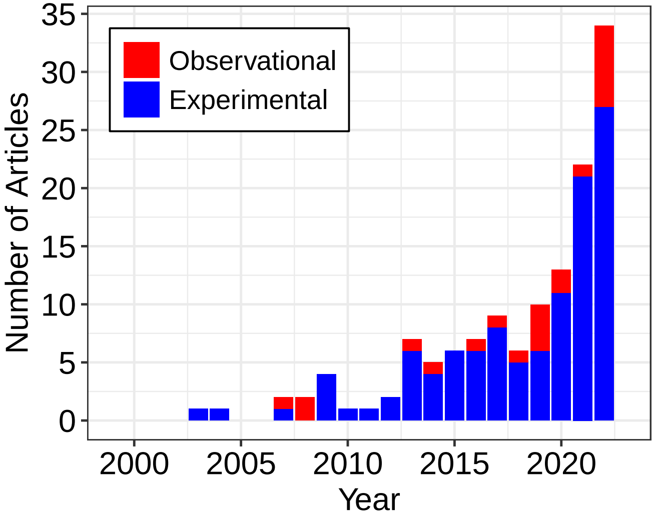
Multi-Site Studies for External Validity about "Contexts"

- **Multi-Site Causal Studies**: Conduct causal studies in multiple sites
  - Example: survey experiments on immigration in the US, UK, and Germany
  - Experimental: Metaketa (EGAP), JPAL, Tomz & Weeks (2013), Carnes & Lupu (2016), Kalla & Broockman (2018), Bassan-Nygate et al (2023), Arechar et al (2024), ...
  - Observational: Eggers et al. (2018), Cavaille & Marshall (2019), Dehejia et al. (2019)

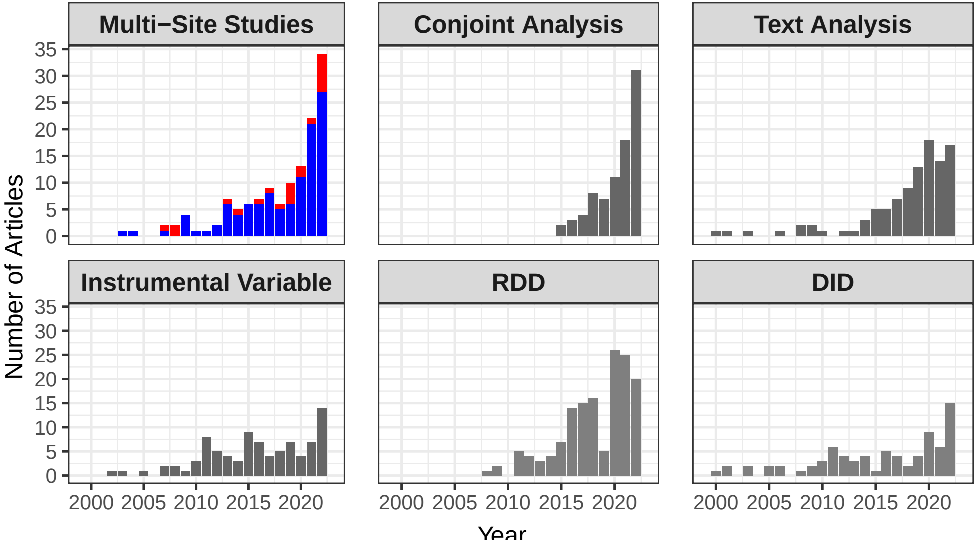
Multi-Site Studies for External Validity about "Contexts"

- **Multi-Site Causal Studies**: Conduct causal studies in multiple sites
  - Example: survey experiments on immigration in the US, UK, and Germany
  - Experimental: Metaketa (EGAP), JPAL, Tomz & Weeks (2013), Carnes & Lupu (2016), Kalla & Broockman (2018), Bassan-Nygate et al (2023), Arechar et al (2024), ...
  - Observational: Eggers et al. (2018), Cavaille & Marshall (2019), Dehejia et al. (2019)
  - Sites can be countries, states (within the US), markets, villages, schools, ...

Multi-Site Causal Studies in Top 10 Political Science Journals



Multi-Site Causal Studies in Top 10 Political Science Journals



Site Selection: The Essential Research Design Question

Site Selection: The Essential Research Design Question

Where should we run experiments?

Site Selection: The Essential Research Design Question

Where should we run experiments?

From where should we collect data for observational causal inference?

Conventional Approaches to Site Selection for External Validity

Conventional Approaches to Site Selection for External Validity

- **Approach 1: Random Sampling** (an obvious but uncommon solution)

Conventional Approaches to Site Selection for External Validity

- **Approach 1: Random Sampling** (an obvious but uncommon solution)
  - **Challenge**: Infeasible due to logistical and ethical constraints

Conventional Approaches to Site Selection for External Validity

- **Approach 1: Random Sampling** (an obvious but uncommon solution)
  - **Challenge**: Infeasible due to logistical and ethical constraints
  - Less than 2% of studies use random sampling

Conventional Approaches to Site Selection for External Validity

- **Approach 1: Random Sampling** (an obvious but uncommon solution)
  - **Challenge**: Infeasible due to logistical and ethical constraints
  - Less than 2% of studies use random sampling
- **Approach 2: Purposive Sampling** (Seawright & Gerring, 2008)

Conventional Approaches to Site Selection for External Validity

- **Approach 1: Random Sampling** (an obvious but uncommon solution)
  - **Challenge**: Infeasible due to logistical and ethical constraints
  - Less than 2% of studies use random sampling
- **Approach 2: Purposive Sampling** (Seawright & Gerring, 2008)
  - Include diverse sites to have wide variation in contextual factors (e.g., regime types)

Conventional Approaches to Site Selection for External Validity

- **Approach 1: Random Sampling** (an obvious but uncommon solution)
  - **Challenge**: Infeasible due to logistical and ethical constraints
  - Less than 2% of studies use random sampling
- **Approach 2: Purposive Sampling** (Seawright & Gerring, 2008)
  - Include diverse sites to have wide variation in contextual factors (e.g., regime types)
  - **Challenges**:

Conventional Approaches to Site Selection for External Validity

- **Approach 1: Random Sampling** (an obvious but uncommon solution)
  - **Challenge**: Infeasible due to logistical and ethical constraints
  - Less than 2% of studies use random sampling
- **Approach 2: Purposive Sampling** (Seawright & Gerring, 2008)
  - Include diverse sites to have wide variation in contextual factors (e.g., regime types)
  - **Challenges**:

    (1) Often focus on just 1 or 2 site-level variables

Conventional Approaches to Site Selection for External Validity

- **Approach 1: Random Sampling** (an obvious but uncommon solution)
  - **Challenge**: Infeasible due to logistical and ethical constraints
  - Less than 2% of studies use random sampling
- **Approach 2: Purposive Sampling** (Seawright & Gerring, 2008)
  - Include diverse sites to have wide variation in contextual factors (e.g., regime types)
  - **Challenges**:

    (1) Often focus on just 1 or 2 site-level variables

    (2) Less transparent and reproducible (Fearon and Laitin, 2008; Nielsen, 2016)

Conventional Approaches to Site Selection for External Validity

- **Approach 1: Random Sampling** (an obvious but uncommon solution)
  - **Challenge**: Infeasible due to logistical and ethical constraints
  - Less than 2% of studies use random sampling
- **Approach 2: Purposive Sampling** (Seawright & Gerring, 2008)
  - Include diverse sites to have wide variation in contextual factors (e.g., regime types)
  - **Challenges**:

    (1) Often focus on just 1 or 2 site-level variables

    (2) Less transparent and reproducible (Fearon and Laitin, 2008; Nielsen, 2016)

    (3) No statistical guarantee

Conventional Approaches to Site Selection for External Validity

- **Approach 1: Random Sampling** (an obvious but uncommon solution)
  - **Challenge**: Infeasible due to logistical and ethical constraints
  - Less than 2% of studies use random sampling
- **Approach 2: Purposive Sampling** (Seawright & Gerring, 2008)
  - Include diverse sites to have wide variation in contextual factors (e.g., regime types)
  - **Challenges**:

    (1) Often focus on just 1 or 2 site-level variables

    (2) Less transparent and reproducible (Fearon and Laitin, 2008; Nielsen, 2016)

    (3) No statistical guarantee
  - Stratified convenience sampling (Olsen et al, 2013)

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**
  - Combines Synthetic Control (Abadie et al, 2010) and Purposive Sampling

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**
  - Combines Synthetic Control (Abadie et al, 2010) and Purposive Sampling
  - SPS selects diverse sites such that

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**
  - Combines Synthetic Control (Abadie et al, 2010) and Purposive Sampling
  - SPS selects diverse sites such that

    non-selected sites can be well approximated by the weighted average of selected sites

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**
  - Combines Synthetic Control (Abadie et al, 2010) and Purposive Sampling
  - SPS selects diverse sites such that

    non-selected sites can be well approximated by the weighted average of selected sites
  - SPS can make the weighted average of selected sites representative

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**
  - Combines Synthetic Control (Abadie et al, 2010) and Purposive Sampling
  - SPS selects diverse sites such that

    non-selected sites can be well approximated by the weighted average of selected sites
  - SPS can make the weighted average of selected sites representative
- **Desirable properties**

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**
  - Combines Synthetic Control (Abadie et al, 2010) and Purposive Sampling
  - SPS selects diverse sites such that

    non-selected sites can be well approximated by the weighted average of selected sites
  - SPS can make the weighted average of selected sites representative

- **Desirable properties**

(1) Flexible: Incorporate practical constraints

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**
  - Combines Synthetic Control (Abadie et al, 2010) and Purposive Sampling
  - SPS selects diverse sites such that

    non-selected sites can be well approximated by the weighted average of selected sites
  - SPS can make the weighted average of selected sites representative

- **Desirable properties**

(1) Flexible: Incorporate practical constraints + domain knowledge

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**
  - Combines Synthetic Control (Abadie et al, 2010) and Purposive Sampling
  - SPS selects diverse sites such that

    non-selected sites can be well approximated by the weighted average of selected sites
  - SPS can make the weighted average of selected sites representative

- **Desirable properties**

(1) Flexible: Incorporate practical constraints + domain knowledge

(2) Transparent: Clarify all factors that affected site selection

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**
  - Combines Synthetic Control (Abadie et al, 2010) and Purposive Sampling
  - SPS selects diverse sites such that

    non-selected sites can be well approximated by the weighted average of selected sites
  - SPS can make the weighted average of selected sites representative

- **Desirable properties**

(1) Flexible: Incorporate practical constraints + domain knowledge

(2) Transparent: Clarify all factors that affected site selection

(3) Statistical foundation: Minimize the worst-case mean squared errors in generalization

Synthetic Purposive Sampling (Egami and Lee, 2025. *AJPS*)

- **Synthetic Purposive Sampling (SPS)**
  - Combines Synthetic Control (Abadie et al, 2010) and Purposive Sampling
  - SPS selects diverse sites such that

    non-selected sites can be well approximated by the weighted average of selected sites
  - SPS can make the weighted average of selected sites representative

- **Desirable properties**

(1) Flexible: Incorporate practical constraints + domain knowledge

(2) Transparent: Clarify all factors that affected site selection

(3) Statistical foundation: Minimize the worst-case mean squared errors in generalization

- SPS offers statistical foundation to select diverse sites for external validity

Introduction to SPS

Setup

- N potential sites: Target population of sites against which external validity is evaluated

Setup

- N potential sites: Target population of sites against which external validity is evaluated

  $\approx$ Scope condition of a substantive theory

Setup

- N potential sites: Target population of sites against which external validity is evaluated

  $\approx$ Scope condition of a substantive theory

- N_S: the number of study sites

Setup

- N potential sites: Target population of sites against which external validity is evaluated

  $\approx$ Scope condition of a substantive theory

- N_S: the number of study sites (the median $= 3$ & the 80th percentile $= 6.6$)

Setup

- N potential sites: Target population of sites against which external validity is evaluated

  $\approx$ Scope condition of a substantive theory
- N_S: the number of study sites (the median $= 3$ & the 80th percentile $= 6.6$)
- **Site-specific Average Treatment Effect (Site-specific ATE)**

$$\theta_k := \mathbb{E}\{Y_{ik}(1) - Y_{ik}(0)\} \text{ for each site } k \in \{1, \ldots, N\}.$$

Setup

- N potential sites: Target population of sites against which external validity is evaluated

  $\approx$ Scope condition of a substantive theory

- N_S: the number of study sites (the median $= 3$ & the 80th percentile $= 6.6$)

- **Site-specific Average Treatment Effect (Site-specific ATE)**

$$\theta_k := \mathbb{E}\{Y_{ik}(1) - Y_{ik}(0)\} \text{ for each site } k \in \{1, \ldots, N\}.$$

- **Average Site Average Treatment Effect (Average-Site ATE)**

Subgroup ATEs

$$\theta_{AS} := \frac{1}{N} \sum_{k=1}^{N} \theta_k$$

Setup

- N potential sites: Target population of sites against which external validity is evaluated

  $\approx$ Scope condition of a substantive theory
- N_S: the number of study sites (the median $= 3$ & the 80th percentile $= 6.6$)
- **Site-specific Average Treatment Effect (Site-specific ATE)**

$$\theta_k := \mathbb{E}\{Y_{ik}(1) - Y_{ik}(0)\} \text{ for each site } k \in \{1, \ldots, N\}.$$

- **Average Site Average Treatment Effect (Average-Site ATE)** Subgroup ATEs

$$\theta_{AS} := \frac{1}{N} \sum_{k=1}^{N} \theta_k$$

- For selected sites, we have conventional estimators (e.g., difference-in-means)

Setup

- N potential sites: Target population of sites against which external validity is evaluated

  $\approx$ Scope condition of a substantive theory

- N_S: the number of study sites (the median $= 3$ & the 80th percentile $= 6.6$)

- **Site-specific Average Treatment Effect (Site-specific ATE)**

$$\theta_k := \mathbb{E}\{Y_{ik}(1) - Y_{ik}(0)\} \text{ for each site } k \in \{1, \ldots, N\}.$$

- **Average Site Average Treatment Effect (Average-Site ATE)**

Subgroup ATEs

$$\theta_{AS} := \frac{1}{N} \sum_{k=1}^{N} \theta_k$$

- For selected sites, we have conventional estimators (e.g., difference-in-means)

- **Question**: How can we select N_S sites to credibly estimate the Average-Site ATE?

Illustration of SPS: Two Site-level Variables

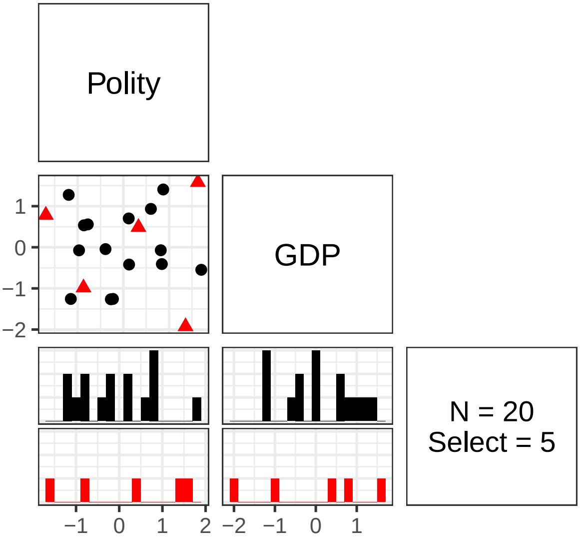
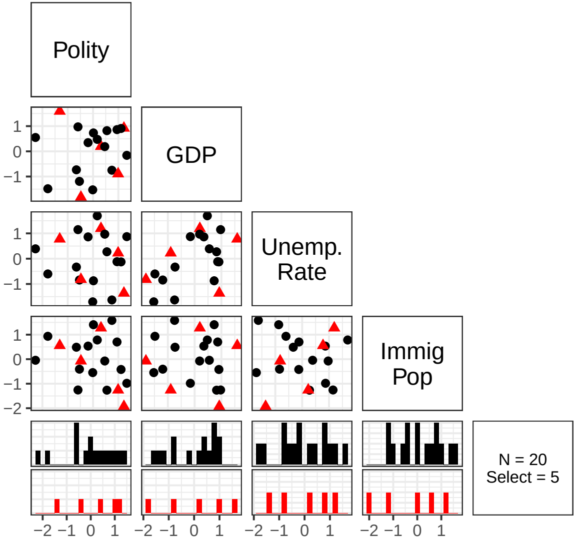


Illustration of SPS: Four Site-level Variables



Step 1: Choosing Site-Level Variables

Step 1: Choosing Site-Level Variables

- **Step 1**: Choose site-level variables $\mathbf{X}$ to diversify across sites

Step 1: Choosing Site-Level Variables

- **Step 1**: Choose site-level variables $\mathbf{X}$ to diversify across sites
  - Contextual factors that are theoretically expected to explain across-site heterogeneity

Step 1: Choosing Site-Level Variables

- **Step 1**: Choose site-level variables $\mathbf{X}$ to diversify across sites
  - Contextual factors that are theoretically expected to explain across-site heterogeneity
  - e.g., Polity, GDP, Unemployment rate, Size of immigrant population

Step 1: Choosing Site-Level Variables

- **Step 1**: Choose site-level variables $\mathbf{X}$ to diversify across sites
  - Contextual factors that are theoretically expected to explain across-site heterogeneity
  - e.g., Polity, GDP, Unemployment rate, Size of immigrant population

- Approximate $\mathbf{X}$ of non-selected sites with the weighted average of $\mathbf{X}$ of selected sites

Step 1: Choosing Site-Level Variables

- **Step 1**: Choose site-level variables $\mathbf{X}$ to diversify across sites
  - Contextual factors that are theoretically expected to explain across-site heterogeneity
  - e.g., Polity, GDP, Unemployment rate, Size of immigrant population

- Approximate $\mathbf{X}$ of non-selected sites with the weighted average of $\mathbf{X}$ of selected sites
  - e.g., GDP of Germany (non-selected) $\approx$ the weighted average GDP of selected countries

Step 2: SPS Solves the Optimization Problem

Step 2: SPS Solves the Optimization Problem

- $X_{k\ell}$: ℓ th site-level variable of site k

Step 2: SPS Solves the Optimization Problem

- $X_{k\ell}$: ℓ th site-level variable of site k
- S_k: Binary variable taking 1 if site k is selected and taking 0 otherwise

Step 2: SPS Solves the Optimization Problem

- $X_{k\ell}$: ℓ th site-level variable of site k
- S_k: Binary variable taking 1 if site k is selected and taking 0 otherwise
- W_{jk}: Weight on selected site j to predict non-selected site k

Step 2: SPS Solves the Optimization Problem

- $X_{k\ell}$: ℓ th site-level variable of site k
- S_k: Binary variable taking 1 if site k is selected and taking 0 otherwise
- W_{jk}: Weight on selected site j to predict non-selected site k
- $B_{k\ell} := (X_{k\ell} - \sum_{j:S_j=1} W_{jk} X_{j\ell})^2$: Approximation error for $X_{k\ell}$ of non-selected site k

Step 2: SPS Solves the Optimization Problem

- $X_{k\ell}$: ℓ th site-level variable of site k
- S_k: Binary variable taking 1 if site k is selected and taking 0 otherwise
- W_{jk}: Weight on selected site j to predict non-selected site k
- $B_{k\ell} := (X_{k\ell} - \sum_{j:S_j=1} W_{jk} X_{j\ell})^2$: Approximation error for $X_{k\ell}$ of non-selected site k
- **Problem**: Minimize the average approximation error by site-selection $\mathbf{S}$ and weights $\mathbf{W}$

Step 2: SPS Solves the Optimization Problem

- $X_{k\ell}$: ℓ th site-level variable of site k
- S_k: Binary variable taking 1 if site k is selected and taking 0 otherwise
- W_{jk}: Weight on selected site j to predict non-selected site k
- $B_{k\ell} := (X_{k\ell} - \sum_{j:S_j=1} W_{jk} X_{j\ell})^2$: Approximation error for $X_{k\ell}$ of non-selected site k
- **Problem**: Minimize the average approximation error by site-selection $\mathbf{S}$ and weights $\mathbf{W}$

$$\min_{(\mathbf{S},\ \mathbf{W})} \quad \frac{1}{N - N_S} \underbrace{\sum_{k=1}^{N} (1 - S_k)}_{\text{over Non-selected sites}} \times \underbrace{\left(\frac{1}{L} \sum_{\ell=1}^{L} B_{k\ell}(\mathbf{S}, \mathbf{W}) \right)}_{\text{over Variables}}$$

Step 2: SPS Solves the Optimization Problem

- $X_{k\ell}$: ℓ th site-level variable of site k
- S_k: Binary variable taking 1 if site k is selected and taking 0 otherwise
- W_{jk}: Weight on selected site j to predict non-selected site k
- $B_{k\ell} := (X_{k\ell} - \sum_{j:S_j=1} W_{jk} X_{j\ell})^2$: Approximation error for $X_{k\ell}$ of non-selected site k
- **Problem**: Minimize the average approximation error by site-selection $\mathbf{S}$ and weights $\mathbf{W}$

$$\min_{(\mathbf{S},\ \mathbf{W})} \quad \frac{1}{N - N_S} \underbrace{\sum_{k=1}^{N} (1 - S_k)}_{\text{over Non-selected sites}} \times \underbrace{\left(\frac{1}{L} \sum_{\ell=1}^{L} B_{k\ell}(\mathbf{S}, \mathbf{W}) \right)}_{\text{over Variables}}$$

such that (a) the number of selected sites $= N_S$ and (b) weights are positive and sum to 1

Step 2: SPS Solves the Optimization Problem

- $X_{k\ell}$: ℓ th site-level variable of site k
- S_k: Binary variable taking 1 if site k is selected and taking 0 otherwise
- W_{jk}: Weight on selected site j to predict non-selected site k
- $B_{k\ell} := (X_{k\ell} - \sum_{j:S_j=1} W_{jk} X_{j\ell})^2$: Approximation error for $X_{k\ell}$ of non-selected site k
- **Problem**: Minimize the average approximation error by site-selection $\mathbf{S}$ and weights $\mathbf{W}$

$$\min_{(\mathbf{S},\ \mathbf{W})} \quad \frac{1}{N - N_S} \underbrace{\sum_{k=1}^{N} (1 - S_k)}_{\text{over Non-selected sites}} \times \underbrace{\left(\frac{1}{L} \sum_{\ell=1}^{L} B_{k\ell}(\mathbf{S}, \mathbf{W}) \right)}_{\text{over Variables}}$$

  such that (a) the number of selected sites $= N_S$ and (b) weights are positive and sum to 1
- **Output**: site selection $\widehat{\mathbf{S}}$ and weights $\widehat{\mathbf{W}}$

SPS is Flexible

SPS is Flexible

1. Incorporate (non-theory-driven) practical and ethical constraints

SPS is Flexible

1. Incorporate (non-theory-driven) practical and ethical constraints
   - e.g., No online samples in certain European countries: $S_k = 0$ for infeasible sites

SPS is Flexible

1. Incorporate (non-theory-driven) practical and ethical constraints
   - e.g., No online samples in certain European countries: $S_k = 0$ for infeasible sites
   - e.g., Funding opportunities, collaborators, and knowledge of the local context

SPS is Flexible

1. Incorporate (non-theory-driven) practical and ethical constraints
   - e.g., No online samples in certain European countries: $S_k = 0$ for infeasible sites
   - e.g., Funding opportunities, collaborators, and knowledge of the local context
2. Incorporate qualitative and theoretical knowledge

SPS is Flexible

1. Incorporate (non-theory-driven) practical and ethical constraints
   - e.g., No online samples in certain European countries: $S_k = 0$ for infeasible sites
   - e.g., Funding opportunities, collaborators, and knowledge of the local context
2. Incorporate qualitative and theoretical knowledge
   - e.g., Select some sites with domain knowledge $\rightarrow$ Remaining sites with SPS

SPS is Flexible

1. Incorporate (non-theory-driven) practical and ethical constraints
   - e.g., No online samples in certain European countries: $S_k = 0$ for infeasible sites
   - e.g., Funding opportunities, collaborators, and knowledge of the local context
2. Incorporate qualitative and theoretical knowledge
   - e.g., Select some sites with domain knowledge $\rightarrow$ Remaining sites with SPS
   - Combine sampling method of your choice with SPS

SPS is Flexible

1. Incorporate (non-theory-driven) practical and ethical constraints
   - e.g., No online samples in certain European countries: $S_k = 0$ for infeasible sites
   - e.g., Funding opportunities, collaborators, and knowledge of the local context
2. Incorporate qualitative and theoretical knowledge
   - e.g., Select some sites with domain knowledge $\rightarrow$ Remaining sites with SPS
   - Combine sampling method of your choice with SPS
3. Prioritize important site-level variables via stratification (Recommended)

SPS is Flexible

1. Incorporate (non-theory-driven) practical and ethical constraints
   - e.g., No online samples in certain European countries: $S_k = 0$ for infeasible sites
   - e.g., Funding opportunities, collaborators, and knowledge of the local context
2. Incorporate qualitative and theoretical knowledge
   - e.g., Select some sites with domain knowledge $\rightarrow$ Remaining sites with SPS
   - Combine sampling method of your choice with SPS
3. Prioritize important site-level variables via stratification (Recommended)
   - e.g., Make sure to have at least 2 democracies and at least 2 autocracies

How to add constraints

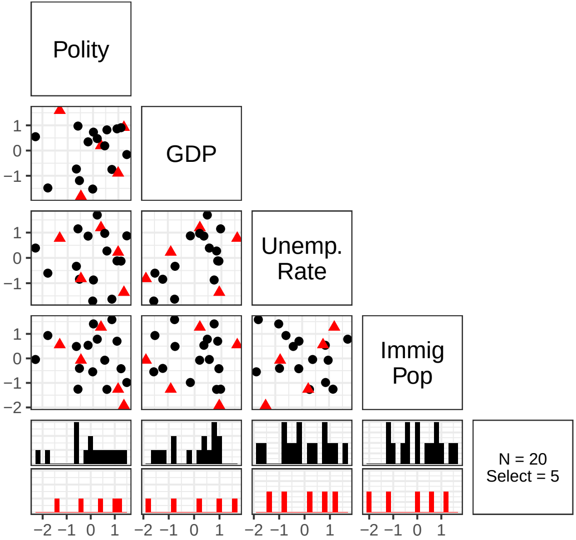
SPS is Flexible

1. Incorporate (non-theory-driven) practical and ethical constraints
   - e.g., No online samples in certain European countries: $S_k = 0$ for infeasible sites
   - e.g., Funding opportunities, collaborators, and knowledge of the local context
2. Incorporate qualitative and theoretical knowledge
   - e.g., Select some sites with domain knowledge $\rightarrow$ Remaining sites with SPS
   - Combine sampling method of your choice with SPS
3. Prioritize important site-level variables via stratification (Recommended)
   - e.g., Make sure to have at least 2 democracies and at least 2 autocracies

How to add constraints

- Related Literature: Abadie and Zhao (2021) and Doudchenko et al (2021)

Users Can Visually Assess Site Selection



External Validity Analysis After Site Selection

External Validity Analysis After Site Selection

- For selected site $j \in \mathcal{S}$, use a conventional causal estimator $\widehat{\theta}_j$ (e.g., difference-in-means)

External Validity Analysis After Site Selection

- For selected site $j \in \mathcal{S}$, use a conventional causal estimator $\widehat{\theta}_j$ (e.g., difference-in-means)
- For non-selected site $k \in \mathcal{R}$, use a weighted average estimator

$$\widehat{\theta}_k^W := \sum_{j \in \mathcal{S}} \widehat{W}_{jk} \widehat{\theta}_j$$

where W_{jk} are SPS weights

External Validity Analysis After Site Selection

- For selected site $j \in \mathcal{S}$, use a conventional causal estimator $\widehat{\theta}_j$ (e.g., difference-in-means)
- For non-selected site $k \in \mathcal{R}$, use a weighted average estimator

$$\widehat{\theta}_k^W := \sum_{j \in \mathcal{S}} \widehat{W}_{jk} \widehat{\theta}_j$$

  where W_{jk} are SPS weights

- For the **Average-Site ATE**, use

$$\widehat{\theta}_{AS} \; := \; \frac{1}{N} \Big(\sum_{j \in \mathcal{S}} \widehat{\theta}_j + \sum_{k \in \mathcal{R}} \widehat{\theta}_k^W \Big)$$

External Validity Analysis After Site Selection

- For selected site $j \in \mathcal{S}$, use a conventional causal estimator $\widehat{\theta}_j$ (e.g., difference-in-means)
- For non-selected site $k \in \mathcal{R}$, use a weighted average estimator

$$\widehat{\theta}_k^W := \sum_{j \in \mathcal{S}} \widehat{W}_{jk} \widehat{\theta}_j$$

  where W_{jk} are SPS weights
- For the **Average-Site ATE**, use

$$\widehat{\theta}_{AS} \; := \; \frac{1}{N} \Big(\sum_{j \in \mathcal{S}} \widehat{\theta}_j + \sum_{k \in \mathcal{R}} \widehat{\theta}_k^W \Big)$$

- We can obtain the conservative variance estimator

External Validity Analysis After Site Selection

- For selected site $j \in \mathcal{S}$, use a conventional causal estimator $\widehat{\theta}_j$ (e.g., difference-in-means)
- For non-selected site $k \in \mathcal{R}$, use a weighted average estimator

$$\widehat{\theta}_k^W := \sum_{j \in \mathcal{S}} \widehat{W}_{jk} \widehat{\theta}_j$$

  where W_{jk} are SPS weights

- For the **Average-Site ATE**, use

$$\widehat{\theta}_{AS} \ := \ \frac{1}{N}\Big(\sum_{j \in \mathcal{S}} \widehat{\theta}_j + \sum_{k \in \mathcal{R}} \widehat{\theta}_k^W\Big)$$

- We can obtain the conservative variance estimator

- **Theorem:** The SPS estimator minimizes the worst-case mean squared error, within a large class of weighted average estimators

Theorem

SPS Minimizes the Worst-Case MSE

Estimation

- SPS optimization problem

$$
\begin{aligned}
\min_{(\mathbf{S}\in\{0,1\},\ \mathbf{W})} \quad & \lambda_1 \times \frac{1}{N-N_S}\sum_{k=1}^{N}(1-S_k)\left(\frac{1}{L}\sum_{\ell=1}^{L} B_{k\ell}(\mathbf{W},\mathbf{S})\right) \\
& + \lambda_2 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N} \underbrace{W_{jk}S_j(1-S_k)\frac{1}{L}\sum_{\ell=1}^{L}(X_{j\ell}-X_{k\ell})^2}_{\text{Weighted Distance between Selected and Non-selected Sites}} \\
& + \lambda_3 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N} \underbrace{S_j(1-S_k)W_{jk}^2}_{\text{Deviation from Uniform Weights}} \\
\text{such that} \quad & \sum_{k=1}^{N} S_k = N_S, \quad \mathbf{W} \geq 0, \quad \sum_{j} S_j W_{jk} = 1 \ \text{ for all sites } k \text{ with } S_k = 0.
\end{aligned}
$$

SPS Minimizes the Worst-Case MSE

- MSE of any weighted-average estimators are bounded

$$
\begin{aligned}
\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left\{\left(\theta_k - \widehat{\theta}_k(\mathbf{W})\right)^2\right\} \;\leq\; & \lambda_1 \times \frac{1}{N-N_S}\sum_{k=1}^{N}(1-S_k)\left(\frac{1}{L}\sum_{\ell=1}^{L} B_{k\ell}(\mathbf{W},\mathbf{S})\right) \\
& + \lambda_2 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N} W_{jk} S_j (1-S_k)\frac{1}{L}\sum_{\ell=1}^{L}(X_{j\ell}-X_{k\ell})^2 \\
& + \lambda_3 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N} S_j(1-S_k)W_{jk}^2 + C
\end{aligned}
$$

- If a linear model of X can explain a larger amount of across-site heterogeneity, λ_1 should be larger
- If the DGP is highly non-linear, λ_2 should be larger
- If unmeasured moderators have larger effect, λ_3 should be larger

Empirical Applications

Attitudes Toward Immigration

- Naumann et al (2018). "Attitudes toward highly skilled and low-skilled immigration in Europe: A survey experiment in 15 European countries."

Attitudes Toward Immigration

- Naumann et al (2018). "Attitudes toward highly skilled and low-skilled immigration in Europe: A survey experiment in 15 European countries."

- **Experimental Design**
  - Outcome: Support for migrants
  - Treatment: highly skilled or low-skilled

Attitudes Toward Immigration

Naumann

- Naumann et al (2018). "Attitudes toward highly skilled and low-skilled immigration in Europe: A survey experiment in 15 European countries."
- **Experimental Design**
  - Outcome: Support for migrants
  - Treatment: highly skilled or low-skilled
- **Empirical Validation**

Attitudes Toward Immigration

Naumann

- Naumann et al (2018). "Attitudes toward highly skilled and low-skilled immigration in Europe: A survey experiment in 15 European countries."
- **Experimental Design**
  - Outcome: Support for migrants
  - Treatment: highly skilled or low-skilled
- **Empirical Validation**
  - We pretend that we can only select 6 sites out of 15 sites
  - We compare to the actual experimental benchmarks

Synthetic Purposive Sampling

Synthetic Purposive Sampling

- **Step 1**: Choose key site-level variables.
  $\rightsquigarrow$ 7 variables discussed in the paper.

Synthetic Purposive Sampling

- **Step 1**: Choose key site-level variables.
  $\rightsquigarrow$ 7 variables discussed in the paper.

- GDP, Size of migrant population,
  Unemployment rates, Support for immigration,
  Fiscal exposure, age, education

Synthetic Purposive Sampling

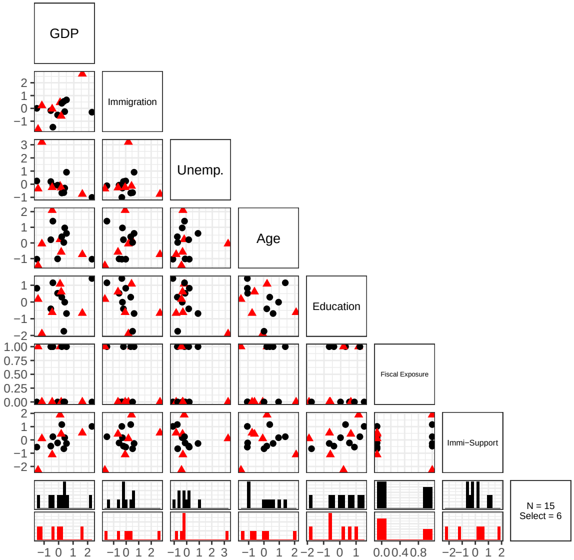
- **Step 1**: Choose key site-level variables.
  $\rightsquigarrow$ 7 variables discussed in the paper.
- GDP, Size of migrant population,
  Unemployment rates, Support for immigration,
  Fiscal exposure, age, education
- **Step 2**: Run Stratified SPS.

Synthetic Purposive Sampling

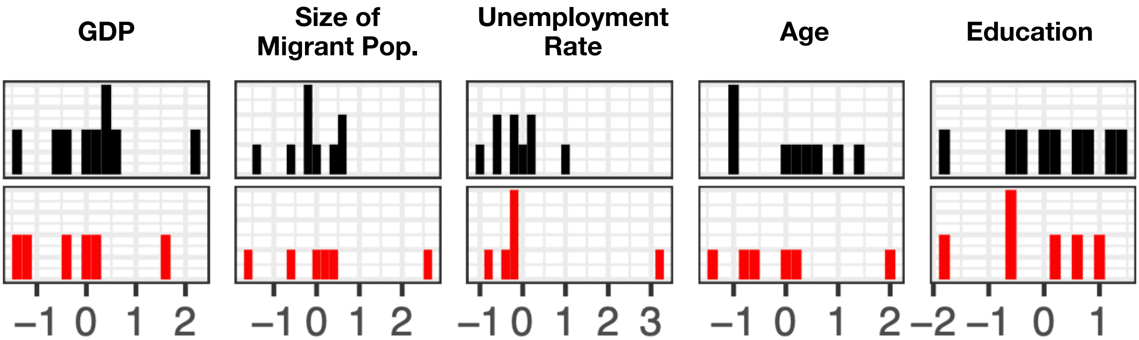
- **Step 1**: Choose key site-level variables.
  $\rightsquigarrow$ 7 variables discussed in the paper.
- GDP, Size of migrant population,
  Unemployment rates, Support for immigration,
  Fiscal exposure, age, education
- **Step 2**: Run Stratified SPS.
- Selected 6 sites
  (Czech Republic, Denmark, Spain, Sweden,
  Switzerland, United Kingdom)

Synthetic Purposive Sampling

- **Step 1**: Choose key site-level variables.
  $\rightsquigarrow$ 7 variables discussed in the paper.
- GDP, Size of migrant population, Unemployment rates, Support for immigration, Fiscal exposure, age, education
- **Step 2**: Run Stratified SPS.
- Selected 6 sites (Czech Republic, Denmark, Spain, Sweden, Switzerland, United Kingdom)



Synthetic Purposive Sampling



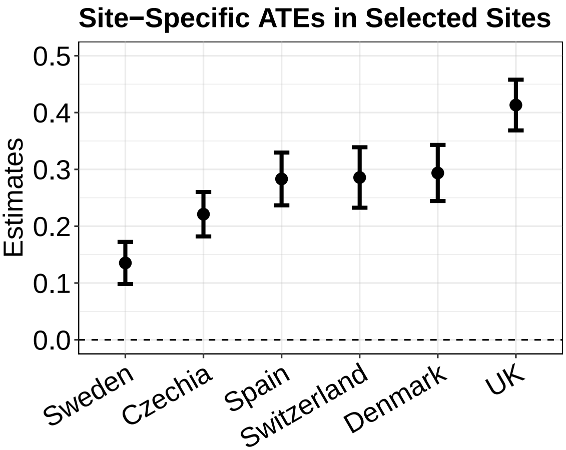
Analysis (Step 1): Internal Validity Analysis

Analysis (Step 1): Internal Validity Analysis

- **Step 1**: Usual analysis of experiments
- Estimate the ATEs in the selected 6 sites
  (Czech Republic, Denmark, Spain, Sweden,
  Switzerland, United Kingdom)

Analysis (Step 1): Internal Validity Analysis

- **Step 1**: Usual analysis of experiments
- Estimate the ATEs in the selected 6 sites (Czech Republic, Denmark, Spain, Sweden, Switzerland, United Kingdom)



Analysis (Step 2): Average-Site Average Treatment Effects

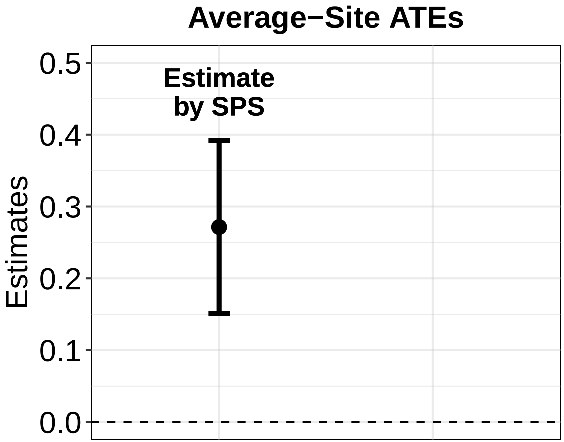
Analysis (Step 2): Average-Site Average Treatment Effects

- **Step 2**: Estimate the Average-Site ATEs

  via the SPS estimator

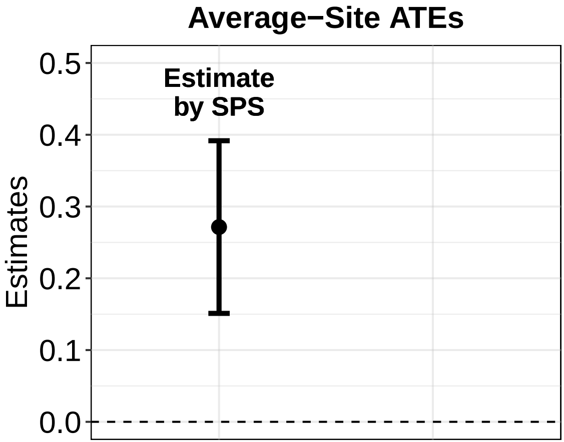
Analysis (Step 2): Average-Site Average Treatment Effects

- **Step 2**: Estimate the Average-Site ATEs via the SPS estimator



Analysis (Step 2): Average-Site Average Treatment Effects

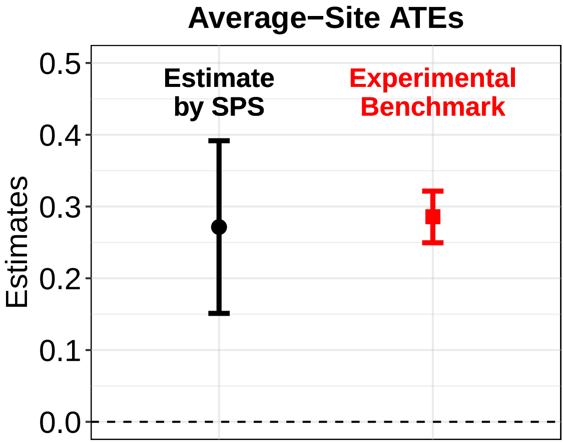
- **Step 2**: Estimate the Average-Site ATEs via the SPS estimator
- Empirical Validation: Compare with the experimental benchmark



Analysis (Step 2): Average-Site Average Treatment Effects

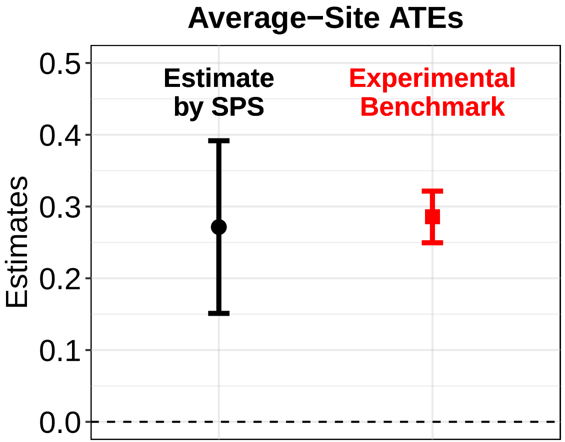
- **Step 2**: Estimate the Average-Site ATEs
  via the SPS estimator

- Empirical Validation:
  Compare with the experimental benchmark



Analysis (Step 2): Average-Site Average Treatment Effects

- **Step 2**: Estimate the Average-Site ATEs
  via the SPS estimator
- Empirical Validation:
  Compare with the experimental benchmark
- Site-level Cross-Validation: p-value $= 0.99$



Synthetic Purposive Sampling (SPS)

Synthetic Purposive Sampling (SPS)

- A general-purpose method to select diverse sites for external validity

Synthetic Purposive Sampling (SPS)

- A general-purpose method to select diverse sites for external validity
- Practicality + Transparency + Statistical Foundation

Synthetic Purposive Sampling (SPS)

- A general-purpose method to select diverse sites for external validity
- Practicality + Transparency + Statistical Foundation
- R package: `spsR` (`https://github.com/naoki-egami/spsR`)

Synthetic Purposive Sampling (SPS)

- A general-purpose method to select diverse sites for external validity
- Practicality + Transparency + Statistical Foundation
- R package: `spsR` (`https://github.com/naoki-egami/spsR`)
- Extension: SPS for selecting Treatments

Concluding Remarks

Research Agenda on External Validity

Research Agenda on External Validity

1. **Formal Framework for External Validity** (Egami & Hartman, 2023. APSR.)
   - (1) Population, (2) Treatment, (3) Outcome, (4) Context

Research Agenda on External Validity

1. **Formal Framework for External Validity** (Egami & Hartman, 2023. APSR.)
   - (1) Population, (2) Treatment, (3) Outcome, (4) Context
2. **Research Design for External Validity** (Egami & Lee, 2025. AJPS.)
   - Site selection via Synthetic Purposive Sampling

Research Agenda on External Validity

1. **Formal Framework for External Validity** (Egami & Hartman, 2023. APSR.)
   - (1) Population, (2) Treatment, (3) Outcome, (4) Context
2. **Research Design for External Validity** (Egami & Lee, 2025. AJPS.)
   - Site selection via Synthetic Purposive Sampling
3. **Estimators for External Validity** (Devaux and Egami, 2022; Jin et al., 2025. PNAS.)
   - Robustness to external validity bias

Research Agenda on External Validity

1. **Formal Framework for External Validity** (Egami & Hartman, 2023. APSR.)
   - (1) Population, (2) Treatment, (3) Outcome, (4) Context
2. **Research Design for External Validity** (Egami & Lee, 2025. AJPS.)
   - Site selection via Synthetic Purposive Sampling
3. **Estimators for External Validity** (Devaux and Egami, 2022; Jin et al., 2025. PNAS.)
   - Robustness to external validity bias
4. **External Validity as "Public Goods"** (Work in Progress)

Research Agenda on External Validity

1. **Formal Framework for External Validity** (Egami & Hartman, 2023. APSR.)
   - (1) Population, (2) Treatment, (3) Outcome, (4) Context
2. **Research Design for External Validity** (Egami & Lee, 2025. AJPS.)
   - Site selection via Synthetic Purposive Sampling
3. **Estimators for External Validity** (Devaux and Egami, 2022; Jin et al., 2025. PNAS.)
   - Robustness to external validity bias
4. **External Validity as "Public Goods"** (Work in Progress)
   - Collective action problem

Research Agenda on External Validity

1. **Formal Framework for External Validity** (Egami & Hartman, 2023. APSR.)
   - (1) Population, (2) Treatment, (3) Outcome, (4) Context
2. **Research Design for External Validity** (Egami & Lee, 2025. AJPS.)
   - Site selection via Synthetic Purposive Sampling
3. **Estimators for External Validity** (Devaux and Egami, 2022; Jin et al., 2025. PNAS.)
   - Robustness to external validity bias
4. **External Validity as "Public Goods"** (Work in Progress)
   - Collective action problem $\rightsquigarrow$ An initiative to fund multi-country survey experiments

Research Agenda on External Validity

1. **Formal Framework for External Validity** (Egami & Hartman, 2023. APSR.)
   - (1) Population, (2) Treatment, (3) Outcome, (4) Context
2. **Research Design for External Validity** (Egami & Lee, 2025. AJPS.)
   - Site selection via Synthetic Purposive Sampling
3. **Estimators for External Validity** (Devaux and Egami, 2022; Jin et al., 2025. PNAS.)
   - Robustness to external validity bias
4. **External Validity as "Public Goods"** (Work in Progress)
   - Collective action problem $\rightsquigarrow$ An initiative to fund multi-country survey experiments

Goal: Statistical & Institutional Foundations for External Validity

Thank You!

Please send comments to
Egami@MIT.Edu

More information on
https://naokiegami.com

Supplementary Slides

- **Formal Framework** Framework
- **Collective Action Problem** Collective
- Subgroup Average-Site ATEs Sub
- Site-Level Cross-Validation CV
- Incorporate Additional Knowledge Const
- Theoretical Guarantees
  - MSE MSE
  - Variance Variance
- Target Population Target
- Practical Guides Guide
- Simulation Study Simulation
- Connections
  - SCM SCM
  - Case Selection Qualitative
  - Meta-Analysis Meta
- Limitation Limitation
- Empirical Literature Review Review
- Empirical Applications
  - Naumann et al (2018) Naumann
  - Blair et al (2021) Blair
  - Gift and Gift (2015) Gift
  - Bisbee et al (2017) Bisbee

Unobserved Site-level Variables

Unobserved Site-level Variables

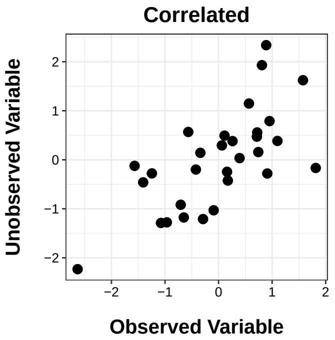
1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables

Unobserved Site-level Variables

1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated

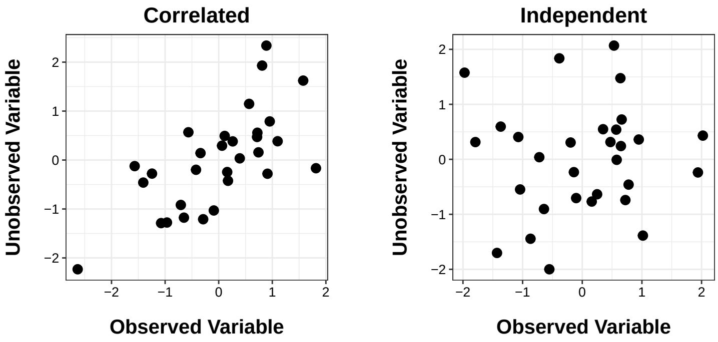
Unobserved Site-level Variables

1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated

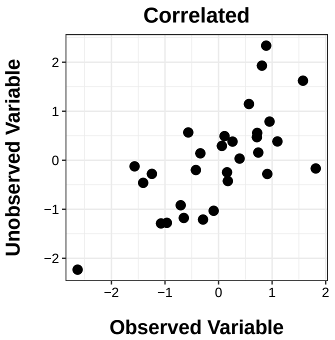


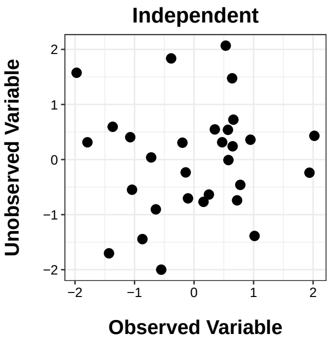
Unobserved Site-level Variables

1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
    - when key site-level variables are correlated



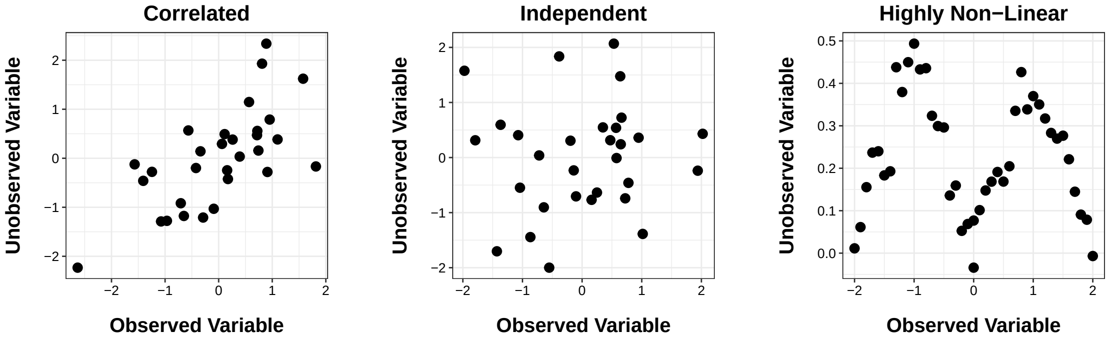
1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated
   - SPS makes the problem worse only when complicated non-linear relationships





Unobserved Site-level Variables

1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated
   - SPS makes the problem worse only when complicated non-linear relationships



Unobserved Site-level Variables

1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated
   - SPS makes the problem worse only when complicated non-linear relationships
2. **Site-Level Cross-Validation**

Site-CV

Unobserved Site-level Variables

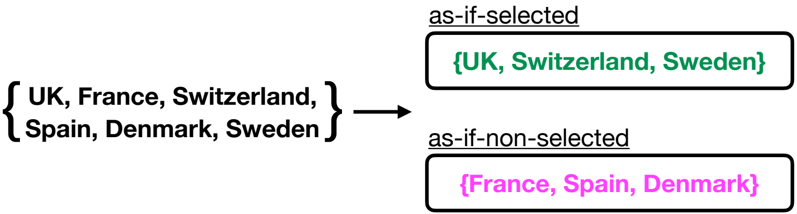
1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated
   - SPS makes the problem worse only when complicated non-linear relationships
2. **Site-Level Cross-Validation** Site-CV
   - Split selected sites into "as-if-selected" sites and "as-if-non-selected" sites

Unobserved Site-level Variables

1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated
   - SPS makes the problem worse only when complicated non-linear relationships
2. **Site-Level Cross-Validation**

   Site-CV

   - Split selected sites into “as-if-selected” sites and “as-if-non-selected” sites



Unobserved Site-level Variables

1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated
   - SPS makes the problem worse only when complicated non-linear relationships

2. **Site-Level Cross-Validation**

   Site-CV

   - Split selected sites into "as-if-selected" sites and "as-if-non-selected" sites
   - Test the predicted ATE from "as-if-selected" $=$ the ATE in "as-if-non-selected"?

{ UK, France, Switzerland, Spain, Denmark, Sweden } →

as-if-selected

{UK, Switzerland, Sweden}

as-if-non-selected

{France, Spain, Denmark}

Unobserved Site-level Variables

1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated
   - SPS makes the problem worse only when complicated non-linear relationships
2. **Site-Level Cross-Validation** Site-CV
   - Split selected sites into "as-if-selected" sites and "as-if-non-selected" sites
   - Test the predicted ATE from "as-if-selected" $=$ the ATE in "as-if-non-selected"?
3. Interpretation: SPS minimizes the prediction error in generalization

Unobserved Site-level Variables

1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated
   - SPS makes the problem worse only when complicated non-linear relationships
2. **Site-Level Cross-Validation** Site-CV
   - Split selected sites into "as-if-selected" sites and "as-if-non-selected" sites
   - Test the predicted ATE from "as-if-selected" = the ATE in "as-if-non-selected"?
3. Interpretation: SPS minimizes the prediction error in generalization
   - Without (often infeasible) random sampling, no method can guarantee unbiasedness

Unobserved Site-level Variables

1. Diversity of observed variables $\rightsquigarrow$ Diversity of *unobserved* variables
   - when key site-level variables are correlated
   - SPS makes the problem worse only when complicated non-linear relationships
2. **Site-Level Cross-Validation** Site-CV
   - Split selected sites into "as-if-selected" sites and "as-if-non-selected" sites
   - Test the predicted ATE from "as-if-selected" = the ATE in "as-if-non-selected"?
3. Interpretation: SPS minimizes the prediction error in generalization
   - Without (often infeasible) random sampling, no method can guarantee unbiasedness
   - Transparency + Optimal diversification of observed variables ($\leftrightarrow$ the current practice)

Framework for External Validity

Egami & Hartman. 2023. "Elements of External Validity." *American Political Science Review.*

Four Dimensions of External Validity

1. X-**validity: Population**
   - Convenience samples, Non-representative samples, ...
2. T-**validity: Treatments**
   - Realistic treatments, Topics of information treatments, Differences in implementations
3. Y-**validity: Outcome measures**
   - Behavioral or survey outcomes, Short- vs Long-term outcomes, ...
4. C-**validity: Contexts**
   - Mechanisms differ across geography, time, political institutions, ...

External Validity as Public Goods

- Generalizable knowledge is a public good: Scientists rely on each other's findings!
- **A Collective Action Problem** (a model of scientists; joint with Kishishita)
  - Common to leave external validity for "a topic of future studies"
  - Hope that others evaluate external validity of existing findings
  - Academia rewards the "first-mover" $\rightarrow$ Others also work on their new questions

$\rightsquigarrow$ Most studies lack evaluation of external validation ($\rightsquigarrow$ Few meta analyses)

- **Multi-Site Experiment Platform** (with Lee & Weiss) inspired by TESS and Metaketa
  - Open calls $\rightarrow$ Field 15 survey experiments across 10 countries for free
  - Theoretical innovations & Contribution to external validity
  - Purposive + Randomized selection of sites and topics

Multi-Site Experiment Platform

- **15 Multi-Site Experimental Studies**
  - Each of the 15 research teams will have a 10-country survey experiment
  - Directly solve the collective action dilemma
- **Meta-Science Paper**: Fast and Credible Accumulation of Knowledge

(1) Document cross-site heterogeneity over 15 diverse topics

(2) What could have been findings if experiments had been done only in one or a few sites?

(3) Test whether and how well cross-site heterogeneity is predicted by experts a priori

- **Pre-registered Empirical Evaluation of Generalization Estimators**
  - Evaluate whether popular generalization estimators can predict causal effects in as-if-unobserved sites $\rightsquigarrow$ First, pre-registered evaluation of generalization estimators

Subgroup Average-Site ATE

Setup

- **Subgroup Average-Site ATE** (a.k.a, the conditional average-site ATE)
  - Test the implications of a theoretical mechanism
  - e.g., preferences toward highly skilled migrants could be stronger in countries with higer fiscal exposure to migration

$$\theta^g_{sub} := \frac{1}{N_g} \sum_{k:G_k=g} \theta_k \quad \text{where variable } G \text{ defines subgroups}$$

- SPS Estimator

$$\widehat{\theta}^g_{sub} := \frac{1}{N_g} \Big(\sum_{j \in \mathcal{S}, G_j = g} \widehat{\theta}_j + \sum_{k \in \mathcal{R}, G_k = g} \widehat{\theta}^W_k \Big)$$

where $\mathcal{S}$ and $\mathcal{R}$ are selected and non-selected sites.

- Note: Each of the subgroup average-site ATE is a causal effect, but the difference between them is descriptive (as in usual experiments)

Site-Level Cross-Validation

Unobserved

- For each iteration b, randomly split selected sites into two equally sized sets, as-if-non-selected sites $\mathcal{S}_0^b$ and as-if-selected sites $\mathcal{S}_1^b$.
- Then, we test the actual average-site ATE estimates in the as-if-non-selected sites $=$ the predicted average-site ATE based on the as-if-selected sites?

$$\widehat{\delta}_b := \underbrace{\frac{1}{N_{b,0}} \sum_{k \in \mathcal{S}_0^b} \widehat{\theta}_k}_{\substack{\text{ASATE in} \\ \text{As-If-Non-Selected Sites}}} - \underbrace{\frac{1}{N_{b,0}} \sum_{k \in \mathcal{S}_0^b} \widehat{\theta}_k^{W_b}}_{\substack{\text{ASATE estimated from} \\ \text{As-If-Selected Sites}}}$$

where $\widehat{\theta}_k^{W_b} = \sum_{j \in \mathcal{S}_1^b} \widehat{W}_{jk}^b \widehat{\theta}_j$ is the SPS estimator for the ATE in site k based on the as-if-selected sites $\mathcal{S}_1^b$.

- We repeat the same procedure many times by randomly splitting selected sites and then combine p-values by the Holm–Bonferroni correction.

How to Incorporate Additional Knowledge into SPS

Flexible

Modifying the Objective Function

- SPS can naturally include any flexible functions of site-level variables, e.g., interaction and higher order terms
- SPS can assign differential weights to different sites and different site-level variables

$$\min_{(\mathbf{s},\ \mathbf{w})} \quad \frac{1}{N - N_S} \sum_{k=1}^{N} \xi_k (1 - S_k) \left(\frac{1}{L} \sum_{\ell=1}^{L} \eta_\ell B_{k\ell}(\mathbf{W}, \mathbf{S}) \right)$$

How to Incorporate Additional Knowledge into SPS

Adding Constraints

- Include practical constraints and domain knowledge
  - Add $S_k = 0$ for infeasible sites and $S_k = 1$ for site k we always select
- Stratification
  - e.g., at least 2 democracies and autocracies; $\sum_k S_k \text{Dem}_k \geq 2$ and $\sum_k S_k \text{Auto}_k \geq 2$
  - e.g., select sites from different quantiles of GDP

    $\sum_k S_k \mathbf{1}\{\text{GDP}_k \leq 20 \text{ percentile}\} \geq 1$, $\sum_k S_k \mathbf{1}\{\text{GDP}_k \in (40 \text{ and } 60 \text{ percentiles})\} \geq 1$,
    $\sum_k S_k \mathbf{1}\{\text{GDP}_k \geq 80 \text{ percentile}\} \geq 1$
- Budget Constraints: $\sum_k S_k \text{cost}_k \leq \text{Total Budget}$

SPS Minimizes the Worst-Case MSE

- SPS optimization problem

$$\min_{(\mathbf{S}\in\{0,1\},\ \mathbf{W})} \quad \lambda_1 \times \frac{1}{N-N_S}\sum_{k=1}^{N}(1-S_k)\left(\frac{1}{L}\sum_{\ell=1}^{L} B_{k\ell}(\mathbf{W},\mathbf{S})\right)$$

$$+\ \lambda_2 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N}\underbrace{W_{jk}S_j(1-S_k)\frac{1}{L}\sum_{\ell=1}^{L}(X_{j\ell}-X_{k\ell})^2}_{\text{Weighted Distance between Selected and Non-selected Sites}}$$

$$+\ \lambda_3 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N}\underbrace{S_j(1-S_k)W_{jk}^2}_{\text{Deviation from Uniform Weights}}$$

$$\text{such that} \quad \sum_{k=1}^{N} S_k = N_S, \quad \mathbf{W}\geq 0, \quad \sum_{j} S_j W_{jk} = 1 \ \text{ for all sites } k \text{ with } S_k = 0.$$

SPS Minimizes the Worst-Case MSE

- MSE of any weighted-average estimators are bounded

$$
\begin{aligned}
\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left\{\left(\theta_k - \widehat{\theta}_k(\mathbf{W})\right)^2\right\} \;\leq\; & \lambda_1 \times \frac{1}{N-N_S}\sum_{k=1}^{N}(1-S_k)\left(\frac{1}{L}\sum_{\ell=1}^{L}B_{k\ell}(\mathbf{W},\mathbf{S})\right) \\
& + \lambda_2 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N}W_{jk}S_j(1-S_k)\frac{1}{L}\sum_{\ell=1}^{L}(X_{j\ell}-X_{k\ell})^2 \\
& + \lambda_3 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N}S_j(1-S_k)W_{jk}^2 + C
\end{aligned}
$$

- If a linear model of X can explain a larger amount of across-site heterogeneity, λ_1 should be larger
- If the DGP is highly non-linear, λ_2 should be larger
- If unmeasured moderators have larger effect, λ_3 should be larger

Variance Estimation

- SPS is a weighted average estimator

$$\widehat{\theta}_{AS} \;=\; \sum_{j \in \mathcal{S}} \widetilde{W}_j \widehat{\theta}_j$$

where $\widetilde{W}_j = (1 + \sum_{k \in \mathcal{R}} \widehat{W}_{jk})/N$, and $\sum_{j \in \mathcal{S}} \widetilde{W}_j = 1$.

- Variance

$$\mathrm{Var}(\widehat{\theta}_{AS}) \;=\; \sum_{j \in \mathcal{S}} \widetilde{W}_j^2 \mathrm{Var}(\widehat{\theta}_j) \;=\; \sum_{j \in \mathcal{S}} \widetilde{W}_j^2 (\sigma_j^2 + \tau^2)$$

where

  - σ_j^2 is the within-site variance of the site-specific ATE estimator
  - τ^2 is the across-site variance.

Variance Estimation

$$\widehat{\tau}^2 := \frac{\sum_{k \in \mathcal{S}} \widehat{e}_k^2 - (\sum_{k \in \mathcal{S}} \widehat{\sigma}_k^2 + \sum_{k \in \mathcal{S}} \sum_{j \in \mathcal{S}_k} \overline{W}_{jk}^2 \widehat{\sigma}_j^2)}{\sum_{k \in \mathcal{S}} (1 + \sum_{j \in \mathcal{S}_k} \overline{W}_{jk}^2)},$$

where

$$\widehat{e}_k \quad := \quad \widehat{\theta}_k - \sum_{j \in \mathcal{S}_k} \overline{W}_{jk} \widehat{\theta}_j,$$

- $\widehat{\sigma}_k^2$ is an (asymptotically) unbiased estimate of the within-site variance σ_k^2
- $\mathcal{S}_k$ is a set of selected sites after removing site k
- $\overline{W}_{jk}$ is the SPS weight we estimate to approximate site k only using sites in $\mathcal{S}_k$
- This variance estimator is valid even when the SPS estimator is biased.

How to Specify the Target Population of Sites

- Target population of sites: a target against which external validity is evaluated

  $\approx$ Scope condition of a substantive theory

- Explicit specification of the target population guards against over-generalization
- Explicit specification of the target population is essential for *any systematic* site selection

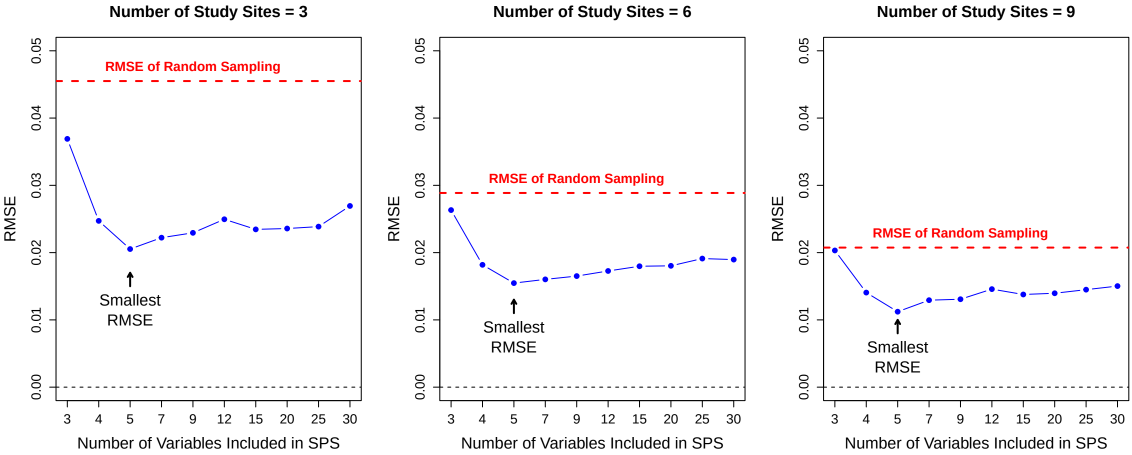
- Users can specify multiple sub-populations (e.g., democracies and autocracies)

  $\rightsquigarrow$ Implement SPS separately within each sub-population

Practical Guides

1. Avoid Including Irrelevant Variables
   - Include too many irrelevant variables $\rightarrow$ Decrease the diversity in key site-level variables
   - See Simulation Simulation
2. Agnostic Use of SPS
   - Users might not be interested in the average-site ATE
   - An agnostic site selection approach to systematically diversify observed site-level covariates, while accommodating logistical and ethical constraints
3. SPS should not be used for "site-hacking"
   - Recommended to transparently report practical constraints and optimally diversify site selection within such constraints
   - The risk exists even for random sampling

Simulation Study



Connection to Synthetic Controls and Synthetic Control Design

- Our basic optimization problem is similar to SCM when sites are already selected and only weights are estimated (i.e., $\mathbf{S}$ is fixed)

- Our basic optimization problem is similar to the one of the synthetic control design (Abadie and Zhao, 2021; Doudchenko et al., 2021).
  - Synthetic Control Design is for internal validity and uses pre-treatment outcomes
  - SPS is for external validity and does not require pre-treatment outcomes

$\rightsquigarrow$ Different causal estimands, constraints, and estimators + Different theoretical guarantees

Connection to and Differences from Qualitative Case Selection

- Huge Literature: Lieberman (2005), Gerring (2006), Fearon and Laitin (2008), Seawright & Gerring (2008), Glynn and Ichino (2016), Nielsen (2016), ...

- **Connection**
  - SPS: Synthetic Controls (Quantitative) + Purposive Sampling (Qualitative)
  - The most common version of purposive sampling: Diverse case selection
  - SPS can be used to select within-case observations

- **Differences**

(1) SPS: Confirmatory analysis (testing hypotheses or estimating effects)

Case Selection: Exploratory analysis (generate new hypotheses or theories)

(2) SPS: External validity (becase causal inference is done within each site)

Case Selection: Causal inference

Connections to and Differences from Meta-Analysis Estimators

- **<u>Connection</u>**
  - Both the SPS estimator and the traditional meta-analysis estimators (fixed-effect and random-effect) are weighted-average estimators

- **<u>Differences</u>**
  - Meta-analysis estimators: Assume across-site differences in the ATEs are random

    $\rightsquigarrow$ Weights are valid only under random sampling
  - SPS estimator: Explicitly takes into account site-level differences in X

    $\rightsquigarrow$ Weights are estimated such that covariates of non-selected sites are well approximated by the weighted average of covariates of selected sites

Limitation: Not Optimized for Meta-Regression

- Some might be interested in running meta-regression to explain cross-site heterogeneity with site-level variables
- Even more difficult problem than estimating the average-site ATE
  - e.g., estimating the effect of 5 site-level variables with 6 study sites
- Meta-regression is more common and reliable in areas where the number of study sites is much larger
  - e.g., psychology, education, and medicine (65 on average)
  - In political science, only 13% of multi-site studies have more than 10 sites
- Researchers can still estimate the subgroup average-site ATEs

Literature Review

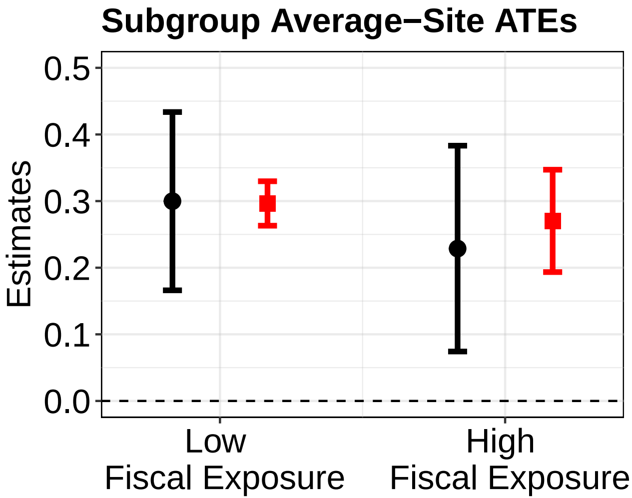
- Top 10 Political Science Journals
  American Political Science Review (APSR), American Journal of Political Science (AJPS), Journal of Politics (JOP), Political Behavior (PB), Quarterly Journal of Political Science (QJPS), British Journal of Political Science (BJPS), Comparative Political Studies (CPS), World Politics (WP), International Organization (IO), and Journal of Experimental Political Science (JEPS).
- Year: 2000 - 2022
- Multi-Site Experimental and Observational Studies: 111 papers and 22 papers

Naumann et al. (2018): Multi-Country Survey Experiment

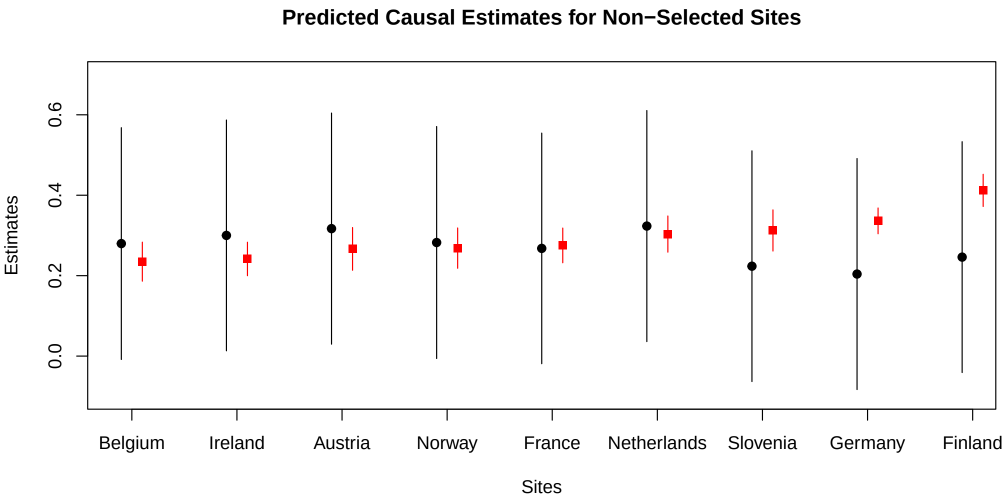
Application

- Target population: all 15 European countries in the original paper
- Site-level variables (7): GDP, Size of migrant population, Unemployment rates, Support for immigration, fiscal exposure, age, education
- Stratification to improve diversity:
  - For each continuous variable, we make sure to select at least one site below the 20th percentile, at least one site between the 40th and 60th percentile, and at least one site above the 80th percentile.
  - For a binary variable (i.e., fiscal exposure), we make sure to select at least one site with high exposure and at least one site with low exposure
- Result: Selected 6 sites {Sweden, Denmark, Spain, Switzerland, Czechia, UK}

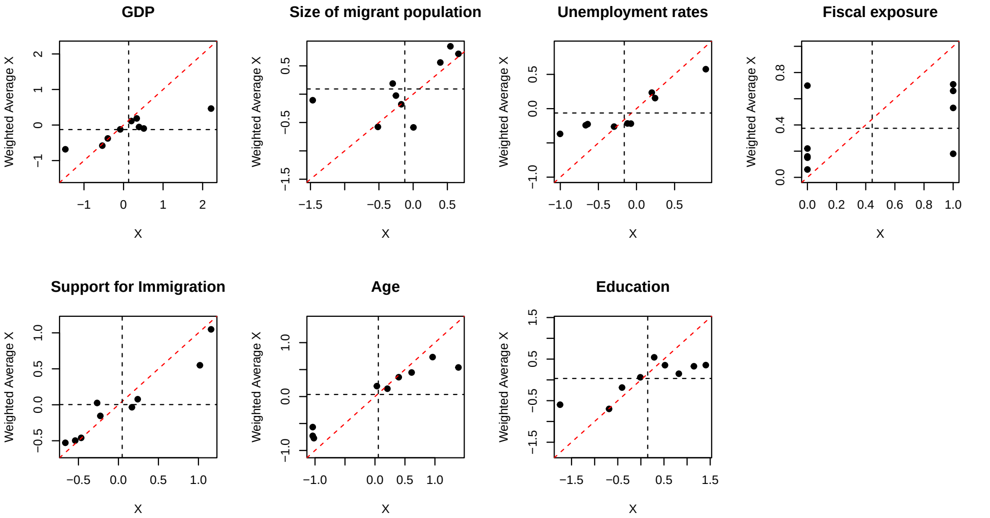
Naumann et al. (2018): Subgroup Average-Site ATEs



Naumann et al. (2018): Estimates in Non-Selected Sites



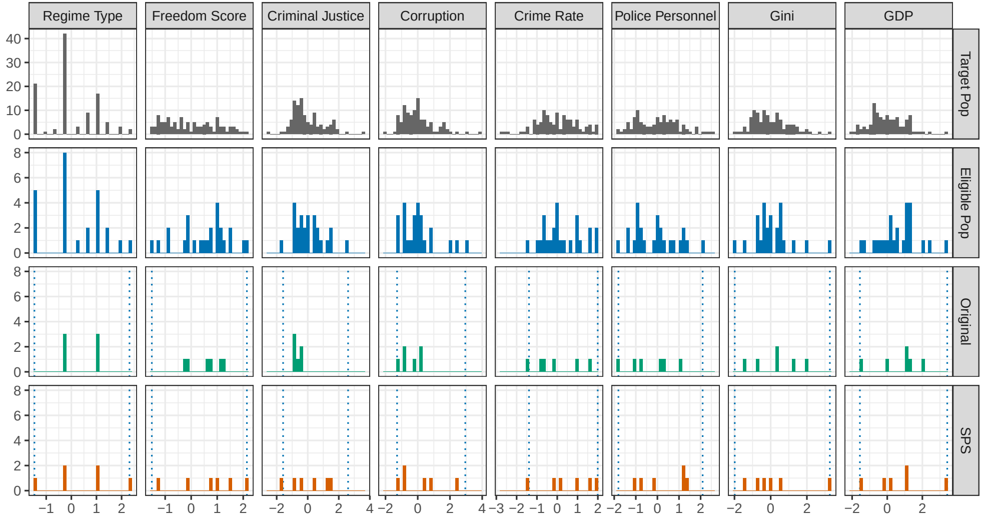
Naumann et al. (2018): Approximation of Site-level Variables



Blair et al. (2021): Multi-Country Field Experiment by Metaketa

- We follow the discussions in the original paper
- Target population: Global South countries with populations of size at least 1 million in Africa, Asia, and South America
- Site-level variables (8): regime type, freedom score, corruption score, criminal justice score, crime rate, the number of police personnel, Gini index, and GDP.
- Constraints: Only EGAP member countries are eligible
- Stratification to improve diversity:
  - Choose two countries from each of the three regions
  - we select at least two democracies and at least two autocracies
  - For the remaining variables, we select sites such that at least one selected site is above 1 standard deviation and at least one selected site is below -1 standard deviation
- Result: Selected 6 sites {Bolivia, China, Liberia, Pakistan, South Africa, Uruguay}

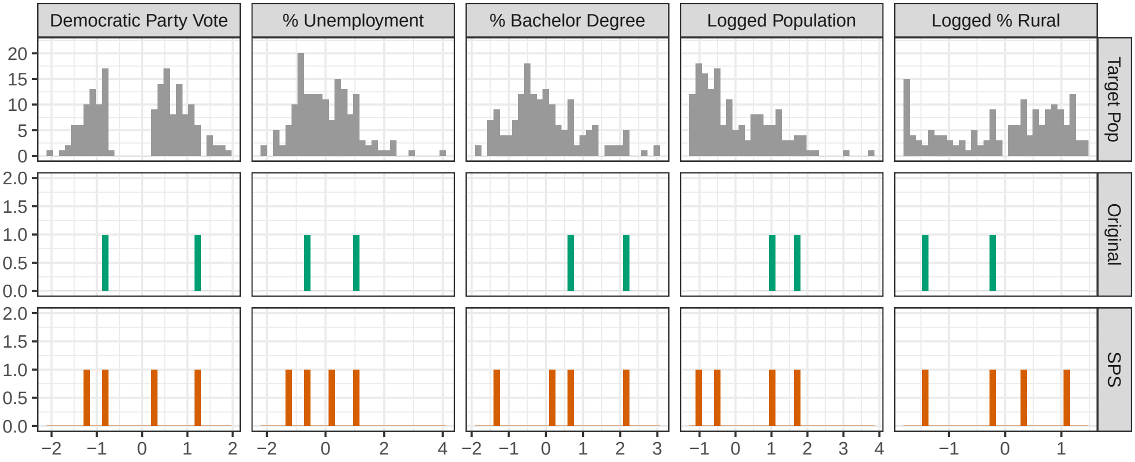
Blair et al. (2021): Multi-Country Field Experiment by Metaketa



Gift and Gift (2015): Multi-County Field Experiment in the US

- Target population:
  - The US counties that are either highly liberal or highly conservative (counties whose vote share for a single party was greater than 60%, following the discussion in the original paper)
  - We also limit the target population to counties with relatively large populations (at least 100K) and relatively high proportions of entry-level jobs for college graduates (above the median).
- Site-level variables (5): Democrat vote share, unemployment rate, education level, population size, and rural population size.
- Constraints: Given the two sites selected by the original authors, select additional two sites
- Stratification to improve diversity:
  - For each variable, we make sure that at least one selected site is above the 80th percentile and at least one selected site is below the 20th percentile of each variable
- Result: Selected 2 sites {Blount County, TN, and Linn County, IA}

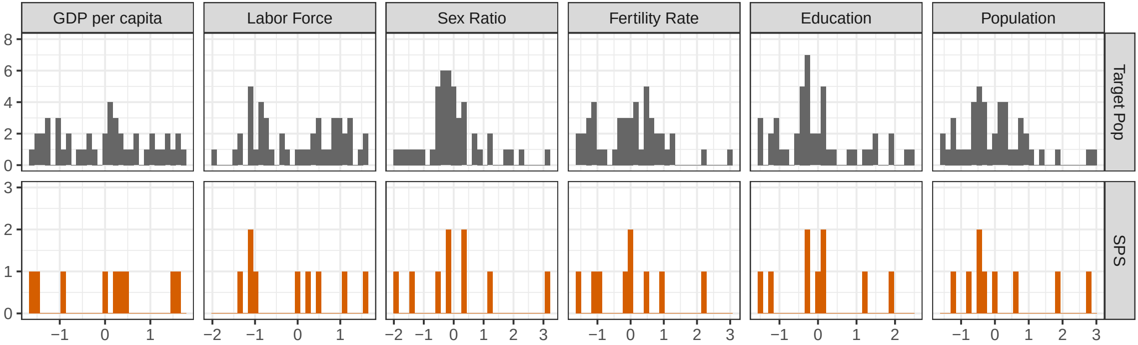
Gift and Gift (2015): Multi-County Field Experiment in the US



Bisbee et al (2017): Multi-Context Observational Study

- Target population: 45 unique countries in the original paper
- Site-level variables (6): GDP per capita, female labor force participation, the sex ratio imbalance, the total fertility rate, population size, and education attainment
- Constraint: we first select the United States based on substantive importance and then select the remaining sites using SPS
- Stratification to improve diversity:
  - at least two sites are selected from each of the four regions: Africa, Americas, Asia, and Europe.
  - For each variable, we include stratification conditions such that we select at least two sites below the 20th percentile, at least two sites between the 40th and 60th percentiles, and at least two sites above the 80th percentile
- Result: Selected 10 sites {Belarus, Chile, Costa Rica, India, Jordan, Rwanda, Spain, Uganda, US}

Bisbee et al (2017): Multi-Context Observational Study



Bisbee et al (2017): Multi-Context Observational Study