When Estimation Becomes the Intervention: Multiple Systems Estimation and Causal Inference Under Structural Change

Albert A-N Scott Moser

22, July 2025

Table of contents

As discussed in the this footnote the results are significant.[1]

0.1 Introduction

0.1.1 1.1 Background

Multiple Systems Estimation (MSE) is widely used to estimate hidden population sizes—such as victims of modern slavery—using overlap in capture across different administrative or NGO lists. In anti-slavery efforts by organizations like IJM, population estimates inform donor decisions, intervention targeting, and country-level impact evaluations.

However, interventions often change not just victimization but the **way people are recorded**, e.g. by increasing list overlap through improved coordination. These measurement artifacts threaten both MSE's validity and any downstream causal inferences.

0.1.2 1.2 Motivating Concern

If a post-treatment increase in list overlap leads to lower MSE estimates—despite no true reduction in the population—MSE may spuriously indicate impact. Conversely, real reductions in prevalence may be masked by offsetting structural changes in the data.

> "Without controlling for measurement artifacts (overlap changes), estimated prevalence changes can mislead causal inference."

0.1.3 1.3 Key Assumption Under Threat

Causal inference via Difference-in-Differences (DiD) or related methods assumes **SUTVA**—that the treatment affects only outcomes, not how outcomes are measured. When treatment affects **list formation, overlap, or coverage**, this assumption fails.

[1] Your detailed footnote text goes here.

0.1.4 1.4 Prior Work

- Lum et al. (2013) and Binette & Steorts (2022) show that violations in list independence or inclusion assumptions cause serious bias in MSE estimates.
- Far et al. (2021) show this empirically in Romanian anti-slavery MSE work.
- Boesche (2022) and Kainou (2017) critique the fragility of SUTVA in quasi-experimental designs, urging alternative assumptions or diagnostics.

0.1.5 1.5 Contributions

- We introduce simulation scenarios showing how structural changes—independent of true prevalence—can bias MSE.
- We evaluate four multivariate binary models to simulate capture processes with controllable overlap and dependence.
- We relate MSE errors to causal inference threats and offer guidelines for quasi-experimental MSE under structural evolution.

0.2 Methodology

0.2.1 2.1 Simulation Framework

- **Population (N):** 1000 hidden individuals
- **Lists (K):** 3 or 4 (e.g., Police, NGO, Immigration, Medical)
- **Prevalence Control:** Set true prevalence (e.g., 5%) and simulate list inclusion for those individuals.

Each individual is represented by a binary vector (e.g., `[1, 0, 1]` means captured on Lists A and C).

0.2.2 2.2 Capture Models

We consider four distinct modeling strategies for the joint distribution over binary list-inclusion vectors $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK}) \in \{0,1\}^K$, capturing various forms of dependence among the K lists. These include both exchangeable and non-exchangeable approaches and range from parametric to nonparametric Bayesian models.

0.2.3 Log-linear Models (Frequentist & Bayesian)

0.2.3.1 Model 1: Log-linear Model

- Full model with interaction terms (e.g., 2-list, 3-list overlaps).

- Strength: interpretable, common in MSE literature.
- Limitation: requires large sample sizes; unstable under sparse data.

Log-linear models provide a flexible parametric framework for modeling dependence structures among binary vectors. Let $\mathcal{Z} = \{0,1\}^K$ denote the space of binary vectors of length K. The probability mass function of $\mathbf{Z}_i \in \mathcal{Z}$ is modeled as:

$$\mathbb{P}(\mathbf{Z}_i = \mathbf{z}) \propto \exp\left(\sum_{s \subseteq \{1,\ldots,K\}} \theta_s \prod_{k \in s} z_k\right),$$

where the sum is over all subsets s of $\{1,\ldots,K\}$, and $\theta_s \in \mathbb{R}$ are interaction parameters. For example, $\theta_{\{k\}}$ captures the marginal effect of list k, while $\theta_{\{k,\ell\}}$ quantifies the pairwise interaction between lists k and ℓ.

In the Bayesian setting, Gaussian priors $\theta_s \sim \mathcal{N}(0, \sigma^2)$ can be placed on the interaction terms to induce regularization. Due to the exponential growth in the number of parameters (2^K), these models are practical only for small K. Computational inference can proceed via Markov Chain Monte Carlo (MCMC), often using data augmentation or Metropolis-Hastings steps.

0.2.4 Exchangeable Binary Arrays with Finite Sufficient Statistics

0.2.4.1 Model 2: Exchangeable Binary Arrays

- Captures average dependence between lists without needing all high-order terms.
- Strength: lower complexity, works with fewer observations.
- Limitation: doesn't capture specific pairwise list behavior.

To mitigate the combinatorial burden of full log-linear modeling, we consider models where the probability of $\mathbf{Z}_i$ depends only on low-dimensional sufficient statistics. Specifically, let:

$$\mathbb{P}(\mathbf{Z}_i = \mathbf{z}) = f\left(\sum_{k=1}^{K} \alpha_k z_k + \sum_{k<\ell} \beta_{k\ell} z_k z_\ell\right),$$

for some function $f : \mathbb{R} \to \mathbb{R}_+$, where α_k are main effect parameters and $\beta_{k\ell}$ encode symmetric pairwise interactions. This model retains parsimony by focusing only on low-order interactions and is naturally exchangeable across individuals i, although not across dimensions k.

Inference proceeds by choosing a parametric or semiparametric form for $f(\cdot)$, such as exponential or logistic link functions, and fitting via MCMC or variational Bayes. This class of models can be viewed as generalizations of Ising models and belongs to the family of finite exchangeable binary arrays as discussed in Diaconis & Freedman (1980).

0.2.5 Pólya-Gamma Augmented Multivariate Probit

0.2.5.1 Model 3: Pólya-Gamma Augmented Multivariate Probit

- Latent Gaussian structure; correlation between lists through shared latent trait.
- Strength: smooths estimates, handles moderate K (lists).
- Limitation: requires more complex inference (e.g., Gibbs sampler).

The multivariate probit model introduces correlation across binary indicators via latent Gaussian variables. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK}) \sim \mathcal{N}(\mu, \Sigma)$ denote a latent Gaussian vector, and define:

$$Z_{ik} = \mathbb{I}(X_{ik} > 0), \quad k = 1, \ldots, K.$$

This induces a joint distribution over $\mathbf{Z}_i \in \{0,1\}^K$ where dependence among the Z_{ik} is encoded entirely in the covariance matrix Σ. A Pólya-Gamma augmentation (Polson et al. (2013)) enables efficient Bayesian inference even for the multivariate probit case by transforming the probit link into a conditionally Gaussian form, making Gibbs sampling tractable.

Priors over Σ can be specified using the LKJ distribution for correlation matrices or inverse-Wishart priors for full covariance matrices. This model is parsimonious, interpretable, and suitable for moderate to large K.

0.2.6 Canonical Bayesian Nonparametric Model: Dirichlet Process Mixture of Product Bernoullis

0.2.6.1 Model 4: Dirichlet Process Mixture of Product Bernoullis

- Allows for heterogeneous "types" of individuals with different capture profiles.
- Strength: flexible, nonparametric.
- Limitation: harder to interpret; computationally expensive.

To allow flexible modeling without specifying a fixed number of latent classes, we employ a Dirichlet Process (DP) mixture model. Let each individual's binary vector $\mathbf{Z}_i$ be generated conditionally independently given a latent parameter vector θ_i:

$$Z_{ik} \mid \theta_i \sim \text{Bernoulli}(\theta_{ik}), \quad \theta_i \sim G, \quad G \sim \text{DP}(\alpha, G_0),$$

where G_0 is a base measure over $[0,1]^K$, often taken as a product of independent Beta distributions: $G_0 = \prod_{k=1}^{K} \text{Beta}(a, b)$.

Marginalizing over G induces clustering of individuals with similar list-inclusion profiles. Dependencies among list indicators are induced through the shared latent parameters θ_i, even though the Z_{ik} are conditionally independent given θ_i. Inference is typically performed using Chinese Restaurant Process (CRP) representations or stick-breaking constructions.

This model is exchangeable over individuals and highly flexible, adapting the complexity of the model to the observed data without requiring a fixed number of latent components.

0.2.7 2.3 Scenario Definitions

0.2.7.1 Scenario 1: False Positive

- **True N:** 1000 pre, 1000 post.
- **Overlap Change:** pairwise correlation between lists rises from 0.0 to 0.6.
- **Expected Result:** MSE shows drop in estimated N (e.g., from 980 $\rightarrow$ 700) despite no true change.

0.2.7.2 Scenario 2: False Negative

- **True N:** 1000 $\rightarrow$ 600 (real reduction).
- **Overlap also increases** (e.g., $= 0 \rightarrow 0.6$).
- **Expected Result:** MSE estimate shows smaller decline (e.g., 980 $\rightarrow$ 720), or possibly none.

0.2.7.3 Optional Scenario 3: Structural Perturbation

- One list is removed, or a subnational list is treated as national.
- Tests effect of **coverage misspecification**.

0.2.7.4 Optional Scenario 4: Spillover

- Agency coordination in one country causes better overlap in another.
- Tests **network-based SUTVA violation**.

0.2.8 2.4 Metrics

- **Bias:** $\hat{N} - N$
- **RMSE:** Across replications
- **False positive/negative rate:** In detecting real change
- **Coverage:** % of simulations where CI includes true N

0.3 Results

0.3.1 3.1 Scenario 1

- MSE underestimates N when overlap rises.
- Error scales with degree of list correlation.
- Example: At $\rho = 0.6$, estimate is 25% too low on average.

0.3.2 3.2 Scenario 2

- True prevalence drops, but MSE misses effect due to offsetting overlap.
- Estimated declines are ~50% of true magnitude.

0.3.3 3.3 Model Comparisons

- Log-linear models perform worst under sparse overlap shifts.
- Pólya-Gamma models more stable, but coverage still poor.
- Dirichlet Process shows best robustness, but least interpretability.

0.4 Discussion

0.4.1 4.1 Causal Inference with Structural Change

- MSE can't be used for causal inference unless structural list changes are controlled.
- SUTVA violations here are **systematic, not random**—not merely noise, but caused by the treatment itself.

0.4.2 4.2 Relation to Literature

- Far et al. (2021) observe this empirically in Romania.
- **Boesche (2020)** proposes SMUTVA (weaker assumptions).
- **Kazunari (2017)** emphasizes need for **sensitivity analyses**—our simulations fulfill this call.

0.4.3 4.3 Practical Implications

- Never interpret MSE declines as treatment effects without evaluating structural list shifts.
- Develop diagnostics for overlap structure (e.g., correlation matrix, list dependency graphs).
- Encourage multi-method approaches (e.g., triangulating with survey, NSUM).

0.5 Conclusion

- MSE estimates are highly sensitive to structural features of list overlap.
- Without accounting for SUTVA violations, causal claims from MSE are often invalid.
- We call for careful modeling, simulation-based diagnostics, and new methodological tools at the intersection of MSE and causal inference.

---

1 Appendices

- Formal model definitions
- Code snippets in R/Python
- Convergence diagnostics
- Real-world IJM country scenarios (simulated data)

References

Binette, O., & Steorts, R. C. (2022). On the reliability of multiple systems estimation for the quantification of modern slavery. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(2), 640–676. https://doi.org/10.1111/rssa.12803

Boesche, T. (2022). Reassessing Quasi-experiments: Policy Evaluation, Induction, and SUTVA. *The British Journal for the Philosophy of Science*, *73*(1), 1–22. https://doi.org/10.1093/bjps/axz006

Diaconis, P., & Freedman, D. (1980). Finite Exchangeable Sequences. *The Annals of Probability*, *8*(4), 745–764.

Far, S. S., King, R., Bird, S., Overstall, A., Worthington, H., & Jewell, N. (2021). Multiple Systems Estimation for Modern Slavery: Robustness of List Omission and Combination Special Issue: Applying Multiple Systems Estimation to Measure Modern Slavery: Methodological Challenges and Innovations. *Crime and Delinquency*, *67*(13–14), 2213–2236.

Kainou, K. (2017). *Review of Necessary Assumptions for Difference-In-Difference (DID) Estimation and Development of Bias Correction Methods for DID where Spillover Effects of Treatment/Causal Effects to the Control Group are not Ignorable and SUTVA Violated (Japanese)* (17075). Research Institute of Economy, Trade and Industry (RIETI).

Lum, K., Price, M. E., & Banks, D. (2013). Applications of Multiple Systems Estimation in Human Rights Research. *The American Statistician*, *67*(4), 191–200. https://doi.org/10.1080/00031305.2013.821093

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349. https://doi.org/10.1080/01621459.2013.829001