Performance of Synthetic Diff-in-Diff models for Geo-Randomized Experiments

Paula Meloni
Amazon.com and UC Berkeley

Stefan Hut
Amazon.com

Mahnaz Islam
Amazon.com

There are different reasons why experimenters may want to randomize their experiment at a region level. In some cases, treatments cannot be turned on or off at the individual level, therefore requiring randomization at a group level, for which regions can be a good candidate. In other cases, experimenters may worry about network effects or other types of spillovers within a geographic area, and opt to randomize at the region level to address these.

These types of experiments oftentimes call for randomization and analysis methods that can account for a relatively small set of randomization units (geo-locations or other clusters of individual ids) that are highly heterogeneous. The use of stratification within the randomization procedure and Synthetic Diff-in-Diffs in the analysis can help ensure that treated and control groups are comparable, a criteria which is not necessarily met when using traditional analyses like ANCOVA or covariate adjusted regressions in these settings. However, the types of geo-randomized experiments vary substantially both in terms of heterogeneity and numbers of clusters.

To this end, we build a simulation approach to compare the performance of Diff-in-Diff (DID) and Synthetic-Diff-in-Diff (SDID) estimators across experiment settings in terms of bias, mean-squared error and standard errors. We construct an imbalance metric based using mean deviations from parallel trends between the control and treatment groups across time. This allows us to gauge how the estimators perform across the metrics of interest as the observed imbalance in trends increases while maintaining all other experimental features constant (i.e., number of units, outcome metric, experimental dates).

Our findings are suggestive that although SDID exhibits lower variance, whereby coefficients are associated with lower standard errors (and correspondingly higher power), it produces biased results for a set of the experiment settings both in A/A tests as well as under homogenous 1% and 5% treatment effects. Under other data settings, SDID outperforms DID by resulting in similar coefficients centered around the 'true' estimates across the observed imbalance spectrum, while the DID estimates exhibit higher variance that grows as the observed imbalance increases. The simulation framework can be used by experimenters to understand whether a SDID exhibits bias in their specific experiment setting. In these cases, we suggest a data-driven approach for defining the regularization parameter for time-weights in the SDiD as a potential solution.

1 Method Comparison

Two of the most commonly used estimators to analyze region-level randomized experiments are SDID and DID. Under the DID estimator, the outcomes of the treated units are compared with those of the control units whilst "partialing" out any changes in outcomes that are common across treatment and control units across time, as well as those that are common within the treatment and the control group across time:

$$(\hat{\tau}, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = \underset{\tau,\mu,\alpha,\beta}{\operatorname{argmin}} \sum_{g=1}^{G} \sum_{t=1}^{T} (Y_{gt} - \mu - \alpha_g - \beta_t - D_{gt}\tau)^2 \tag{1}$$

The SDID estimator solves the same two-way fixed effects regression problem but additionally selects weights $\hat{\omega}_g^{sdid}, \hat{\lambda}_t^{sdid}$ that align the trends of the treated and untreated units in the pre-experimental period, and between the pre- and post-experimental periods as in Arkhangelsky et al. (2021) (details in appendix):

$$(\hat{\tau}, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = \underset{\tau,\mu,\alpha,\beta}{\operatorname{argmin}} \sum_{g=1}^{G} \sum_{t=1}^{T} (Y_{gt} - \mu - \alpha_g - \beta_t - D_{gt}\tau)^2 \hat{\omega}_g^{sdid} \hat{\lambda}_t^{sdid} \tag{2}$$

By putting more weight on control units that are more similar to those of the treated units and time periods that are more similar to the treated periods, the SDID estimator can accommodate non-parallel pre-treatment trends. SDID removes bias by down

weighting units that are very dissimilar to treated units and improves precision by reducing the weight of time periods that are different from the post-treatment periods.

However, if there is little systematic heterogeneity across time or between units, then the unequal weighting of time periods and units will yield less precise outcomes as compared to the traditional DID estimator. Furthermore, the method is not applicable in settings with a large number of units, or treatment units. Intuitively, we would expect relatively better performance of SDID as compared to DID when the number of units or clusters is small and outcomes are heterogeneous.

2 Simulation Framework

In order to assess the performance of the two estimators across settings with different degrees of observed imbalance, we adopt a Monte Carlo approach as outlined below:

Step 1: For each simulation s=1,...,S, where S=10,000, we randomly split the geographical units into a treatment and a control group.

Step 2: For each simulation we introduce the treatment effect to the treatment group. We start with an A/A test where the treatment effect is zero.

Step 3: For each s, we calculate an imbalance metric between treatment and control units. The imbalance metric is the de-meaned norm between the mean of the control and treatment groups across time:

$$\sqrt{\sum_{t=1}^{T}((\bar{Y_t^1} - \bar{Y^1}) - (\bar{Y_t^0} - \bar{Y^0}))^2} \tag{3}$$

Step 4: We sort simulations into 100 j distinct buckets in order of imbalance.

Step 5: We estimate the treatment effect as specified and calculate our outcomes of interest for each of the j buckets under each respective method A:

(a) Bias: We compute the bias as the difference between our estimate and the "true effect":

$$Bias(\widehat{ATE_{A,j}}) = \widehat{ATE_A} - ATE \tag{4}$$

(b) MSE:

$$MSE(\widehat{ATE}_{A,j}) = |S_j|^{-1} \sum_{s \subset S_j} (\widehat{ATE}_A^s - ATE)^2 \tag{5}$$

(c) Standard errors:

$$\widehat{SE}_{A,j} = \sqrt{\hat{V}_A} \tag{6}$$

$$\hat{V_A} = (G-1)N^{-1} \sum_{i=1}^{N} (\hat{\tau}^{-i} - \hat{\tau})^2 \tag{7}$$

Step 6: We plot our metric of interest versus the imbalance metric under the DiD and SDiD approach and compare performance.

Step 7: We repeat this procedure introducing treatment effects of 1%, 3% and 5% to assess how bias and MSE scale with the treatment effect. We then additionally repeat this procedure using stratification in the randomization process.

3 Simulation Results

The pre-and post experiment outcome variable trends, as well as the simulation results for two of the experiments are outlined below. The simulation plots display the mean square error as well as the median coefficients and standard errors for each imbalance quintile. The imbalance quintiles are based on percentiles.

We show results for two different regional experiments, using real experiment data. The simulation results for the first experiment are intuitive as we expect that the SDID estimator will have better performance than the DID estimator in the form of lower bias

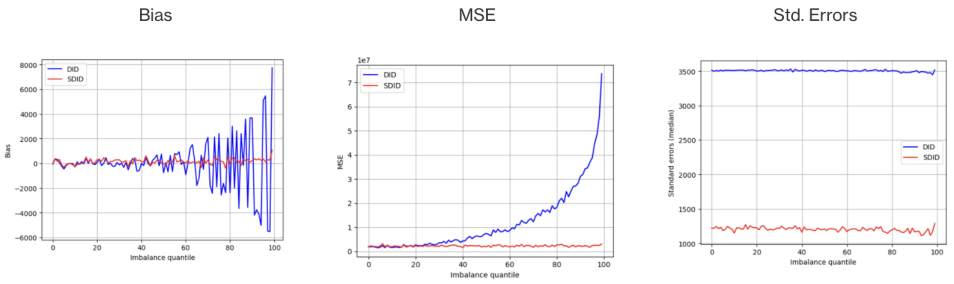

Figure 1: Simulation results for Experiment 1

as the imbalance worsens. We also find that the SDID estimator has correspondingly lower mean squared errors. In addition, standard errors are lower under SDID than DID. We also find that the bias of the DID estimator does not scale with the size of the treatment effect. Results for 1% and 5% treatment effects are displayed here.

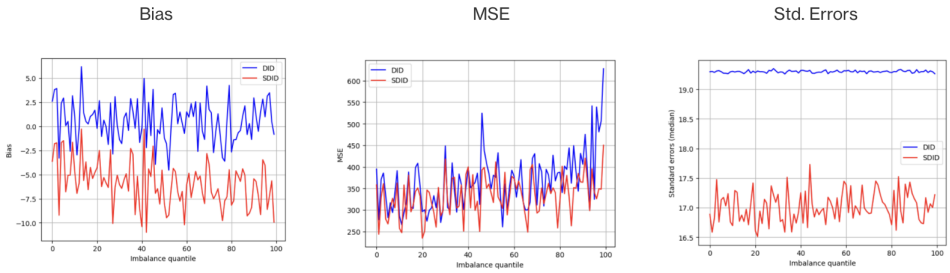

Figure 2: Simulation results for Experiment 2

In the second experiment, however, we see that the DID estimator performs relatively better overall as compared to SDID in terms of bias and the size of standard errors, although the MSE is similar. The relative performance is very similar across the A/A test, the 1% ATE simulation, and the 5% ATE simulation. While the standard errors are smaller for SDID as compared to DID, the variance is very similar, and bias is smaller for DID vs. SDID. Results for the 5% ATE simulations range within 0.125% from the 'true estimates' for DID and 0.15% for SDID, although the mean of the estimates is 0.033% off from the 'true estimates' as compared to 0.1% for SDID. Results for the 1% ATE simulation are on average within 0.07% of the 'true estimates' for SDID while they are within 0.01% of the 'true estimates' for DID.

Disentangling all potential sources of bias while using real experimental data is likely to be infeasible. However, as a first step to diagnose the root of the problem, we re-run the simulations for Experiments 3 as described in Section 2 including two additional approaches: (a) DID with time weights as determined by the SDID estimator, and (b) DID with unit weights as determined by the SDID estimator. This will allow us to determine whether the issue might be related to the unit weights, the time weights, or both.

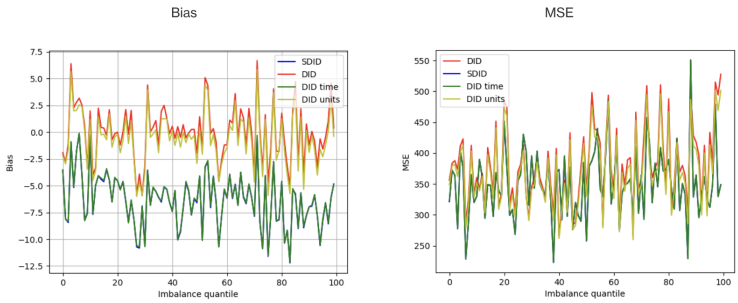

Figure 3: Simulation results with/without weights for Experiment 2

The results under DID with no unit weights are centered at zero while those with time weights are not. The time weights in SDID are subject to a regularization parameter set to $\zeta = 10^{-6}\hat{\sigma}$ where $\hat{\sigma}$ is defined based on the control period. Defining this parameter via a data-driven approach might be a solution to addressing bias in some contexts.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The review of economic studies*, 72(1):1–19.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.

Bottmer, L., Imbens, G. W., Spiess, J., and Warnick, M. (2024). A design-based perspective on synthetic control methods. *Journal of Business & Economic Statistics*, 42(2):762–773.

A Appendix:

A.1 Unit and time weight selection in SDID:

Specifically, time weights ($\hat{\lambda}^{sdid}$) are obtained by solving:

$$(\hat{\lambda}_0, \hat{\lambda}^{sdid}) = \underset{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda}{\operatorname{argmin}} \ \ell_{time}(\lambda_0, \lambda) \tag{8}$$

where

$$\ell_{time}(\lambda_0, \lambda) = \sum_{g=1}^{G_{co}} (\lambda_0 + \sum_{t=1}^{T_{pre}} \lambda_t Y_{gt} - \frac{1}{T_{post}} \sum_{t=T_{pre+1}}^{T} Y_{gt})^2 \tag{9}$$

$$\Lambda = \left\{ \lambda \in \mathbb{R}_+^T : \sum_{t=1}^{T^{pre}} \lambda_t = 1, \ \lambda_t = T_{post}^{-1} \text{ for all } t = T^{pre} + 1, \ldots, T \right\} \tag{10}$$

Unit weights ($\hat{\omega}^{sdid}$) are obtained by solving the optimization problem:

$$(\hat{\omega}_0, \hat{\omega}^{sdid}) = \underset{\omega_0 \in \mathbb{R}, \omega \in \Omega}{\operatorname{argmin}} \ \ell_{unit}(\omega_0, \omega) \tag{11}$$

$$\ell_{unit}(\omega_0, \omega) = \sum_{t=1}^{T_{pre}} (\omega_0 + \sum_{g=1}^{G_{co}} \omega_i Y_{gt} - \frac{1}{G_{tr}} \sum_{g=G_{co}+1}^{G} Y_{gt})^2 + \zeta^2 T_{pre} \|\omega\|_2^2, \tag{12}$$

$$\Omega = \omega \in \mathbb{R}_+^G : \sum_{g=1}^{G_{co}} \omega_g = 1, \omega_g = G_{tr}^{-1} \text{ for all } g = G_{co} + 1, \ldots, G \tag{13}$$

where ζ is a regularization parameter as described in Arkhangelsky et al. (2021).