Synthetic Difference in Differences for Repeated Cross-Sectional Data

Yoann Morin[1]

[1] *CESAER UMR1041, INRAE, Institut Agro, Université Bourgogne Franche-Comté*

Abstract

The synthetic difference-in-differences method provides an efficient method to estimate a causal effect with a latent factor model. However, it relies on the use of panel data. This paper presents an adaptation of the synthetic difference-in-differences method for repeated cross-sectional data. The treatment is considered to be at the group level. Thus, it is possible to aggregate data by group to compute the two types of synthetic difference-in-differences weights on these aggregated data. Then, I compute a third type of weight that account for the different number of observations for each cross-section. I also provide simulation results showing the performance of the synthetic difference-in-differences estimator is improved when using the third type of weights on repeated cross-sectional data.

JEL Classification: C29, C15

Keywords: causal inference, difference-in-difference, Synthetic difference-in-differences, latent factor models, repeated cross-sections

1 Introduction

The synthetic difference-in-differences (SDiD) method (Arkhangelsky et al., 2021) combines features from difference-in-differences and synthetic control methods. Like synthetic controls methods, it computes weights that match pretreatment outcomes between the treated and the control groups. This reduces the need for the parallel trend assumption, often hard to verify in practice (Roth, 2022). Like difference-in-differences methods, it can control for unit-specific shifts and include covariates.

This method is based on panel data, yet, repeated cross-sectional data, where each unit is observed only once, is often the only available data. For example, the evaluation of rent control policies often relies on online listing portals data that are repeated cross-sections.

In this paper, I adapt the SDiD method to repeated cross-sectional data. I focus on the case where each unit belongs to the same group across time and where the treatment is applied to one or more group. Once data are aggregated, unit and time weights are computed for each group using the SDID method. Then, a third type of weight is computed to account for the different number of observations in each group-period. Using a simulation study, I show that this estimator (RC-SDiD) performs better in terms of bias, standard deviation and RMSE that the SDiD one when the number of observations differs in each group-period.

In section 2, I present the synthetic difference-in-differences estimator for repeated cross-sectional data. Then, I demonstrate the properties of the RC-SDiD estimator in a simulation study in section 3. Section 4 concludes.

2 The RC-SDID estimator

Consider a dataset with T independent cross-sections, $t = 1, ..., T$. Each cross-section t is a random sample from the underlying population. There are K groups: the first K^{co} are the control groups and the last $K^{\text{tr}} = K - K^{\text{co}}$ the treated groups. Each group is observed for every period in the sample, but individuals are not observed for each t. Thus, the individual index depends on the cross-section, such as $i(k,t)$ is an individual observed at time t in the group k, and its outcome is $Y_{i(k,t),t}$. The number of individuals in each cross-section may vary over time. $N_{k,t}$ is the number of observations in group k at time t. The treatment is assigned to groups K^{tr} at time $t \geq T_{\text{post}}$.

I apply the SDiD method on aggregated data. $Y_{i(k,t),t}$ is aggregated as:

$$\bar{Y}_{k,t} = \frac{1}{N_{k,t}} \sum_{i} Y_{i(k,t),t} \tag{1}$$

where $\bar{Y}_{k,t}$ is the average of the outcome in group k at time t.
Those data are used to compute the weights presented in Arkhangelsky et al. (2021). The first weights to estimate are similar to those used in synthetic control methods (Abadie et al., 2010). They match the outcome of the treated group with a combination of the outcome in the control groups in the pre-treatment period. They also include a group-specific shift that is accounted for in the last step in a weighted DiD regression (equation 6). Thus, those weights only need to make the outcomes from the treated and control groups parallel and not an exact match. Those weights $\hat{\omega}^{sdid}$ are computed as:

$$\begin{aligned}
&\left(\hat{\omega}_0, \hat{\omega}^{\text{sdid}}\right) = \underset{\omega_0 \in \mathbb{R}, \omega \in \Omega}{\arg\min}\ \ell_{unit}(\omega_0, \omega) \quad \text{where} \\
&\ell_{unit}(\omega_0, \omega) = \sum_{t=1}^{T_{\text{pre}}} \left(\omega_0 + \sum_{k=1}^{K_{\text{co}}} \omega_k \bar{Y}_{k,t} - \frac{1}{K_{\text{tr}}} \sum_{k=K_{\text{co}}+1}^{N} \bar{Y}_{k,t} \right)^2 + \zeta^2 T_{\text{pre}} \left\| \omega \right\|_2^2, \\
&\Omega = \left\{ \omega \in \mathbb{R}_+^N : \sum_{k=1}^{K_{\text{co}}} \omega_k = 1,\ \omega_k = K_{\text{tr}}^{-1} \text{ for all } k = K_{\text{co}}+1, \ldots, K \right\},
\end{aligned} \tag{2}$$

The regularization parameter ζ is:

$$\begin{aligned}
&\zeta = (K_{\text{tr}} T_{\text{post}})^{1/4}\ \hat{\sigma} \quad \text{with} \quad \hat{\sigma}^2 = \frac{1}{K_{\text{co}}(T_{\text{pre}}-1)} \sum_{k=1}^{K_{\text{co}}} \sum_{t=1}^{T_{\text{pre}}-1} \left(\Delta_{k,t} - \overline{\Delta} \right)^2, \\
&\text{where} \quad \Delta_{k,t} = \bar{Y}_{k,(t+1)} - \bar{Y}_{k,t}, \quad \text{and} \quad \overline{\Delta} = \frac{1}{K_{\text{co}}(T_{\text{pre}}-1)} \sum_{k=1}^{K_{\text{co}}} \sum_{t=1}^{T_{\text{pre}}-1} \Delta_{k,t}.
\end{aligned} \tag{3}$$

ζ allows increasing the dispersion of the weights and ensuring their uniqueness.

The second type of weights, $\hat{\lambda}^{\text{sdid}}$ match pretreatment periods with post-treatment periods for the control group. They give more weight to pre-treatment periods that are similar to post-treatment periods. They are computed as:

$$
\begin{aligned}
\left(\hat{\lambda}_0,\, \hat{\lambda}^{\mathrm{sdid}}\right) &= \underset{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda}{\arg\min}\; \ell_{time}(\lambda_0, \lambda) \quad \text{where} \\
\ell_{time}(\lambda_0, \lambda) &= \sum_{k=1}^{K_{\mathrm{co}}} \left(\lambda_0 + \sum_{t=1}^{T_{\mathrm{pre}}} \lambda_t \bar{Y}_{k,t} - \frac{1}{T_{\mathrm{post}}} \sum_{t=T_{\mathrm{pre}}+1}^{T} \bar{Y}_{k,t} \right)^2 \zeta^2 K_{\mathrm{co}} \|\lambda\|^2, \\
\Lambda &= \left\{ \lambda \in \mathbb{R}_+^T : \sum_{t=1}^{T_{\mathrm{pre}}} \lambda_t = 1,\, \lambda_t = T_{\mathrm{post}}^{-1} \text{ for all } t = T_{\mathrm{pre}}+1, \ldots, T \right\}.
\end{aligned}
\tag{4}
$$

where the regularization parameter ζ is set to $\zeta = 10^{-6}\ \hat{\sigma}$, $\hat{\sigma}$ taking the same value than in equation 3.

Because the number of observations in each group may be different for each group-period, $\hat{\omega}_k^{\mathrm{sdid}}$ do not assure a similar trend between the treated and control groups. I compute a third type of weight that accounts for the different number of observations in each group-period. The cross-sectional weights $\nu_{k,t}^{RC}$ are computed as:

$$
\nu_{k,t}^{RC} = \frac{1}{N_{k,t}} \tag{5}
$$

The weights $\hat{\omega}_k^{\mathrm{sdid}} \times \nu_{k,t}^{RC}$ allow matching the outcome of the treated and control groups for repeated cross-sectional data. For each individual $i(k,t), t$, the weights $\nu_{k,t}^{RC}$ sum to 1 in each group-period. It allows making each period equally weighted, as it was the case when computing the weights $\hat{\lambda}_t^{\mathrm{sdid}}$ on aggregated data.

Once the weights $\nu_{k,t}^{RC}$ are obtained, the treatment effect is estimated in a weighted regression as:

$$
\left(\hat{\tau}^{rc-sdid},\, \hat{\mu},\, \hat{\alpha},\, \hat{\beta}\right) = \underset{\tau,\mu,\alpha,\beta}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{i(k,t),t} - \mu - \alpha_k - \beta_t - W_{kt}\tau \right)^2 \hat{\omega}_k^{\mathrm{sdid}} \hat{\lambda}_t^{\mathrm{sdid}} \nu_{k,t}^{RC} \right\} \tag{6}
$$

where μ is an intercept, α_k a group fixed effect and β_t a time fixed effect. $W_{k,t}$ is a binary variable representing treatment exposure, and τ is the treatment effect.

The RC-SDiD method can be adapted to include covariates (Arkhangelsky et al., 2021; Kranz, 2022) and to staggered treatment timing (Arkhangelsky et al., 2021; Porreca, 2022).

3 Monte Carlo simulations

I use Monte Carlo simulations to study the properties of the RC-SDiD estimator compared to the SDiD and DiD estimators. I generate the outcome using a data generating process (DGP) similar to Xu (2017) but without covariates. The treatment being at the group level, I also consider individual effects to be at this level. The unobservable dimension is generated using interactive fixed effects (Bai, 2009). The outcome for observation i in group k at time t is simulated as:

$$Y_{i(k,t),t} = \tau W_{k,t} + \alpha_k + \beta_t + \Lambda_k^{'} f_t + \varepsilon_{i(k,t),t} \quad (7)$$

Only the first group is considered to be treated to keep a reasonable number of observations. Λ_k is a vector of group factor loadings and f_t a vector of time factors. Both are of size r such as: $\Lambda_k = (\Lambda_1, \Lambda_2, ..., \Lambda_r)^{'}$ and $f_t = (f_1, f_2, ..., f_r)^{'}$. The error term $\varepsilon_{i(k,t),t}$, time fixed effects β_t, and f_t are i.i.d. $N(0,1)$. The values of group fixed effects α_k and factor loadings Λ_k are drawn from uniform distributions. For the control groups, I use $U[-\sqrt{3}, \sqrt{3}]$ and $U[\sqrt{3}-2w\sqrt{3}, 3\sqrt{3}-2w\sqrt{3}]$ for the treated group, where $w \in [0,1]$. As highlighted in Xu (2017), it allows treatment status and group-specific effects to be correlated and to have a variance of 1 for the random variables (when $0 \leq w < 1$).

The varying number of observations in each group-period is computed as:

$$N_{k,t} = \begin{cases} S_k \times Base_{RC} + S_k \times E_{k,t} & \text{if t=1} \\ N_{k,t-1} + S_k \times E_{k,t} & \text{if t>1} \end{cases} \quad (8)$$

where S_k is a scale parameter for each group k. It allows the number of observations to differ between each group. It is drawn from a discrete uniform distribution and I test different range for this parameter. The draw of S_k is allowed to be correlated with group fixed effects for different correlation levels ρ. $Base_{RC}$ is the baseline number of observations in each cross-section. To simulate the evolution of the number of observations, the variable $E_{k,t}$ is defined as:

$$E_{k,t} = N(Base_{RC} \times 0.02, \frac{\sqrt{Base_{RC}}}{2}) \tag{9}$$

On average, the number of observations in each cross-section will slightly grow, while keeping a reasonable number of observations.

I draw new independent values of $\varepsilon_{i(k,t),t}$ for each of the 1000 repetitions. The other parameters are fixed, with $Base_{RC} = 100$, $S_k \in [1, 10]$ is correlated with group fixed effects for $\rho = 0.2$, and $w = 0.2$. I use 30 control groups and 30 periods where half are pre-treatment periods. The treatment effect is set to $\tau = 0.3$ and the number of factors and loadinds is $r = 1$. I compare the performance of the RC-SDiD estimator with DiD and SDiD by computing the mean bias, the standard deviation and the RMSE.

	Mean bias			SD			RMSE		
	DiD	RC-SDiD	SDiD	DiD	RC-SDiD	SDiD	DiD	RC-SDiD	SDiD
$S_k = 1$	0.4468	0.0013	-0.0373	0.0320	0.0411	0.0436	0.4480	0.0411	0.0574
$S_k \in [1, 2]$	0.4472	0.0005	-0.0437	0.0231	0.0309	0.0332	0.4478	0.0308	0.0549
$S_k \in [1, 4]$	0.4409	0.0007	-0.0422	0.0162	0.0213	0.0238	0.4412	0.0213	0.0484
$S_k \in [1, 6]$	0.4350	-0.0013	-0.0446	0.0128	0.0173	0.0201	0.4352	0.0173	0.0489
$S_k \in [1, 8]$	0.4397	-0.0008	-0.0443	0.0116	0.0158	0.0186	0.4399	0.0158	0.0480
$S_k \in [1, 10]$	0.4380	0.0010	-0.0434	0.0103	0.0130	0.0166	0.4381	0.0130	0.0464
$S_k \in [1, 15]$	0.4366	0.0005	-0.0438	0.0086	0.0112	0.0145	0.4367	0.0112	0.0461
$S_k \in [1, 20]$	0.4375	0.0002	-0.0445	0.0079	0.0099	0.0129	0.4376	0.0099	0.0463

Table 1: Variation of the scale parameter

In table 1, the effect of the scale parameter on the performance of the model is studied. When the value of this parameter increases, the number of observations in the dataset also increases. Thus, an efficient estimate should perform better when S_k grows. The DiD estimate is the most biased estimate because of interactive fixed effects, while the RC-SDiD is the least biased one. Its bias is reduced when S_k grows. The SD of all models gets lower as S_k grows because of the higher number of observations. However, the RMSE of the DiD and SDiD estimators are only slightly affected by the increase of S_k, while the RMSE of RC-SDiD is greatly improved. Thus, adding the third type of weight presented in section 2 significantly improves the performance of the estimator when the number of observations differs in each cross-section.

	Mean bias			SD			RMSE		
	DiD	RC-SDiD	SDiD	DiD	RC-SDiD	SDiD	DiD	RC-SDiD	SDiD
$r = 0$	0.0000	-0.0002	-0.0001	0.0106	0.0128	0.0128	0.0106	0.0128	0.0128
$r = 1$	0.4375	0.0001	-0.0446	0.0104	0.0133	0.0163	0.4376	0.0133	0.0474
$r = 2$	-0.0750	0.0002	-0.0254	0.0103	0.0161	0.0263	0.0757	0.0161	0.0366
$r = 3$	0.9439	-0.0001	-0.0472	0.0104	0.0184	0.0269	0.9439	0.0184	0.0543
$r = 4$	1.6735	-0.0004	-0.0973	0.0103	0.0188	0.0289	1.6735	0.0188	0.1015

Table 2: Variation of the number of factors and loadings

Table 2 focuses on the effects of the number of factors and loadings on the performance of the estimators. When $r = 0$, the DGP only includes group and time fixed effect, and the DiD estimator performs the best across all measures. However, when r increases, it becomes the estimator that performs the worst because it cannot account for interactive fixed effects. The RC-SDiD estimator has the best performance across all values of $r > 0$. The bias, SD and RMSE of the SDiD estimator are all highly increasing with r, while they stay moderate for the RC-SDiD one.

In appendix A, B and C the results are reported for respectively: different values of w, different values of ρ and for different sample sizes. In all cases, RC-SDiD performs better than DiD and SDiD, confirming our previous results.

4 Conclusion

In this paper, I adapted the SDiD estimator (Arkhangelsky et al., 2021) for repeated cross-sectional data, with a simple implementation that consists in adding a third type of weight that depends on the number of observations in each pair of group-period. Using Monte Carlo simulations, I demonstrated that the RC-SDiD estimator significantly improves the performance of the SDiD estimator when using repeated cross-sectional data.

Acknowledgements

I am grateful to the comments and suggestions from Marie Breuillé, Capucine Chapel, Julie Le Gallo and Morgan Ubeda. Financial support from the Atelier Parisien d'Urbanisme is gratefully acknowledged.

References

Abadie, A., Diamond, A., & 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.

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

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, *77*(4), 1229–1279.

Kranz, S. (2022). Synthetic difference-in-differences with time-varying covariates.

Porreca, Z. (2022). Synthetic difference-in-differences estimation with staggered treatment timing. *Economics Letters*, *220*, 110874.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, *4*(3), 305–322.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, *25*(1), 57–76.

Appendices

Appendix A

Table A1 presents the results for different values of w. If $w = 1$, treatment assignment is random, while when it decreases toward 0, individual fixed effects and factor loadings are more shifted, reducing the overlap between the control and treated groups. As in the previous table, the RC-SDiD estimator performs the best in all cases. The more the treatment status is correlated with individual effects, the more the SDiD and DiD estimators are biased. The RC-SDiD estimator performs well for all values of w, and its SD and RMSE are only slightly higher for the highest values of w compared to the lowest ones.

	Mean bias			SD			RMSE		
	DiD	RC-SDiD	SDiD	DiD	RC-SDiD	SDiD	DiD	RC-SDiD	SDiD
$w = 1$	-0.1408	-0.0002	0.0108	0.0112	0.0129	0.0132	0.1413	0.0129	0.0170
$w = 0.8$	0.0039	0.0000	0.0025	0.0111	0.0123	0.0124	0.0117	0.0123	0.0126
$w = 0.6$	0.1486	0.0005	-0.0071	0.0103	0.0121	0.0126	0.1490	0.0121	0.0145
$w = 0.4$	0.2931	0.0006	-0.0201	0.0103	0.0122	0.0138	0.2933	0.0123	0.0244
$w = 0.2$	0.4372	0.0000	-0.0443	0.0107	0.0139	0.0169	0.4373	0.0139	0.0474
$w = 0$	0.5819	-0.0001	-0.0705	0.0104	0.0147	0.0204	0.5820	0.0147	0.0734

Table A1: Variation of treatment assignment

Appendix B

The results for different values of ρ are presented in table B1. This parameter controls the correlation between the scale parameter S_k and the individual fixed effects. This parameter has overall a pretty low effect on the simulation results. The mean bias, standard deviation and RMSE all have similar values for the different values of ρ that are tested. The RC-SDID estimates are on average the least biased and have the lowest RMSE. Its standard deviation is only slightly higher than the DID one (which is highly biased) and is always lower than for SDID estimates.

	Mean bias			SD			RMSE		
	DiD	RC-SDiD	SDiD	DiD	RC-SDiD	SDiD	DiD	RC-SDiD	SDiD
$\rho = 0$	0.4254	-0.0005	-0.0379	0.0120	0.0156	0.0196	0.4256	0.0156	0.0427
$\rho = 0.2$	0.4373	-0.0001	-0.0447	0.0100	0.0139	0.0166	0.4374	0.0139	0.0477
$\rho = 0.5$	0.4465	-0.0002	-0.0421	0.0159	0.0196	0.0248	0.4468	0.0196	0.0489
$\rho = 0.8$	0.4579	-0.0002	-0.0424	0.0126	0.0163	0.0204	0.4581	0.0163	0.0470
$\rho = 1$	0.4523	-0.0009	-0.0420	0.0133	0.0171	0.0209	0.4525	0.0171	0.0469

Table B1: Variation of the correlation between individual fixed effects and the scale parameter

Appendix C

The results reported in table C1 study the effects of sample size on the performance of the three estimators. Samples are simulated with a different number of baseline observations in each cross-section (50 vs 100) and a different number of control groups and periods (30 vs 15 for both). For each sample size considered, RC-SDiD still performs better than DiD and SDiD estimates. The performance of the RC-SDiD estimator is more impacted by reducing the number of time periods than the number of groups. Reducing the number of observations in each cross-section also slightly raise its RMSE and standard deviation. But it performs better than DiD and SDiD in all cases.

	Mean bias			SD			RMSE		
	DiD	RC-SDiD	SDiD	DiD	RC-SDiD	SDiD	DiD	RC-SDiD	SDiD
$Base_{RC} = 100$									
$K^{co} = 30, T = 30$	0.4368	-0.0003	-0.0444	0.0105	0.0135	0.0165	0.4370	0.0135	0.0473
$K^{co} = 15, T = 30$	0.4542	-0.0003	-0.0691	0.0103	0.0149	0.0225	0.4543	0.0149	0.0727
$K^{co} = 30, T = 15$	-0.1763	-0.0001	-0.0408	0.0154	0.0186	0.0196	0.1770	0.0186	0.0453
$K^{co} = 15, T = 15$	-0.1818	0.0006	-0.0456	0.0158	0.0207	0.0226	0.1825	0.0207	0.0508
$Base_{RC} = 50$									
$K^{co} = 30, T = 30$	0.4304	-0.0011	-0.0469	0.0143	0.0193	0.0217	0.4307	0.0194	0.0516
$K^{co} = 15, T = 30$	0.4450	0.0009	-0.0745	0.0143	0.0219	0.0305	0.4452	0.0219	0.0805
$K^{co} = 30, T = 15$	-0.1668	-0.0004	-0.0415	0.0234	0.0284	0.0288	0.1685	0.0284	0.0505
$K^{co} = 15, T = 15$	-0.1729	-0.0018	-0.0480	0.0225	0.0317	0.0339	0.1743	0.0318	0.0587

Table C1: Variation of the number of control groups, time periods, and baseline number of observation by group