Synthetic Experimental Design for a UBI pilot study[1]

Jaume Vives-i-Bastida (MIT)

1. Introduction

This report provides a guide for practitioners wanting to use synthetic experimental designs to evaluate policy interventions. It focuses on the Catalan universal basic income pilot study that aims to treat two towns in 2023 with a substantial universal basic income for a period of two years. The main goal of the report is to show how inference on various outcomes of interest can be achieved by choosing the towns to treat using the synthetic experimental design framework of Abadie and Zhao (2021). We show that approximate inference can be achieved despite the small number of treated units. This guide, however, expands beyond the standard synthetic experimental design framework by considering inference on multiple outcomes (see Trejo et al. (2021) and Amjad et al. (2019) for related methods) and by providing a point-by-point rubric to dealing with practical concerns such as choosing exclusion constraints or thinking about allocation fairness.

The main conclusion of the report is that in the Catalan UBI pilot study setting there exists a valid synthetic experimental design. The weighted pair of towns A and B is able to replicate the population values for a variety of outcomes in the years leading to the experiment (names redacted until the experiment is official). The report ties theory and practical concerns to show how we would do inference on the outcomes of interests if a pair such as A and B were chosen for treatment. Furthermore, it details all the design choices that led to this pair being chosen. These include, amongst others, the choice of the target population, the set of outcome variables and covariates considered, the exclusion restrictions imposed to avoid interference and the specific form of the target loss function. In particular, each section in the report is devoted to a step in building the synthetic experimental design.

- **Section 2 (Theory)** considers the theory behind using synthetic experimental design to perform inference on the outcomes of the policy intervention. It shows how one might expand Abadie and Zhao (2021) to include multiple outcomes and provides a fast algorithm for implementation when the number of treated units is small relative to the pool of units.

[1]This report has been anonymized to avoid revealing the identity of the porposed treatment and control villages. Throughout the report the preferred pair is referred as A and B.

– **Section 3 (Design choices)** details the criteria used to choose the outcome variables, the potentially treated units, the set of controls units and the target population. These set of criteria could be useful to researchers thinking about similar designs.

– **Section 4 (UBI application)** applies the methodology to the Catalan UBI pilot study and evaluates the goodness of fit of the preferred pair in order to check if the theoretical assumptions for inference are likely to hold. The conclusion is that the A and B pair likely satisfy the theoretical assumptions for most outcomes. Furthermore, a minimum detectable effect analysis suggests that for most outcomes of interest the experiment will be able to detect moderate effects with sizes larger than 0.1 standard deviations.

– **Section 5 (Lotteries and fairness)** evaluates whether a fairer allocation could be achieved through the use of a lottery while maintaining good pre-treatment fit. The conclusion is that in the Catalan UBI setting a lottery would not help resolve the tension between fairness and efficiency. However, the section may provide guidance for other settings in which a lottery may be suitable.

Overall, the sections of this report are thought as a step by step guide on how to use synthetic experimental design for experimental policy evaluation. From the theory behind inference, to how to think about the design choices and target parameters and how to get and evaluate the synthetic treated and control units. Finally, section 5 also considers the related problem of how to trade-off fairness and efficiency when treatment is not chosen completely at random.

2. The theory of multi-experimental design using synthetic controls

We are interested in the experimental design setting of Abadie and Zhao (2021) with multiple outcome variables. The framework we use is based on the standard synthetic control method (Abadie and Gardeazabal (2003), Abadie et al. (2010), Abadie et al. (2015)), but alternative designs have also been proposed by Doudchenko et al. (2021). Consider the problem of designing an experiment for a small number of aggregate treated units. The target parameter is the average treatment effect with respect to a weighted average of the overall population of units for various outcomes of interest. We aim to treat a small number m of units at time T_0 from a pool of J units indexed $j = 1, \ldots, J$ observed for T periods of time. We denote the $k = 1, \ldots, K$ outcome of interest by Y_{it}^k with potential outcomes $Y_{it}^k(D_{it})$ indexed by

treatment status $D_{it} \in \{0, 1\}$. The unit-level treatment effect of interest for outcome k is then defined as

$$\tau_{it}^k = Y_{it}^k(1) - Y_{it}^k(0),$$

for $j = 1, \ldots, J$, $t = T_0 + 1, \ldots, T$ and $k = 1, \ldots, K$. As in Abadie and Zhao (2021), our target parameter of interest is a weighted average treatment effect

$$\tau_t^k = \sum_{j=1}^{J} f_j(Y_{it}^k(1) - Y_{it}^k(0)),$$

where $f_1, \ldots, f_j$ represent known population weights such that $f_j \geq 0$ and $\sum_j f_j = 1$. In the simplest case, each unit is weighted uniformly and $f_j = 1/J$.

Experimental design In order to estimate τ_t for $t = T_0 + 1, \ldots, T$ the researcher chooses two sets of weights:

$$\text{Treatment weights}: \quad (w_1, \ldots, w_J) \quad \text{s.t} \sum_{j=1}^{J} w_j = 1,$$

$$\text{Control weights}: \quad (v_1, \ldots, v_J) \quad \text{s.t} \sum_{j=1}^{J} v_j = 1,$$

$$w_j \geq 0, v_j \geq 0, \text{ and } w_j v_j = 0, \text{ for all } j = 1, \ldots, J.$$

The treated units are the ones with $w_j > 0$ and will be assigned to the intervention at time T_0, while the control units are the ones with $v_j > 0$ and will be used to estimate the counterfactual in absence of intervention for the treated units for $t > T_0$. Overall, we can separate the experimenter problem in two main goals:

1. **Representative experiment**: we want to choose $w_1, \ldots, w_J$ such that in absence of the intervention the treated units are representative of the population. That is, for each outcome k,
$$\sum_{j=1}^{J} w_j Y_{jt}^k(0) = \sum_{j=1}^{J} f_j Y_{jt}^k(0).$$

2. **Valid control group**: we want to choose $v_1, 0 \ldots, v_J$ such that in absence of the intervention the outcomes for the treated units and the controls units are very close.

This is the "good pre-treatment fit" condition that is standard in synthetic control studies. For each outcome k it boils down to having that

$$\sum_{j=1}^{J} w_j Y_{jt}^k(0) = \sum_{j=1}^{J} v_j Y_{jt}^k(0).$$

Both conditions, the representative condition and the valid control group condition, may only hold approximately in practice. However, the validity of our method, both in terms of identification of our parameter of interest τ_t^k and inference, relies on our model being able to satisfy both conditions. Furthermore, the same guiding principles as in synthetic controls apply (see Abadie and Vives-i-Bastida 2021 for a detailed guide). In this case, our estimated treatment effect for weights $(\boldsymbol{w}, \boldsymbol{v})$ recuperates the weighted average treatment effect for $t > T_0$

$$\tau_t^k(\boldsymbol{w}, \boldsymbol{v}) = \sum_{j=1}^{J} w_j Y_{jt}^k(1) - \sum_{j=1}^{J} v_j Y_{jt}^k(0) = \sum_{j=1}^{J} f_j (Y_{jt}^k(1) - Y_{jt}^k(0)) = \tau_t^k.$$

Finding $\boldsymbol{w}$ and $\boldsymbol{v}$ As in the standard synthetic control method, we estimate the weights by using predictors of the pre-intervention outcomes. Given that in our setting we have k outcomes of interest, as in Trejo et al. (2021) we construct our design matrix by concatenating all outcomes Y_{it}^k for the pre-intervention periods. Let

$$\boldsymbol{X}_j = (Y_{j1}^1, \ldots, Y_{jT_0}^1, \ldots, Y_{j1}^K, \ldots, Y_{jT_0}^K, Z_j)' \quad \text{and} \quad \bar{\boldsymbol{X}} = \sum_{j=1}^{J} f_j \boldsymbol{X}_j$$

be the set of predictors and target predictor of interest respectively, where Z_j denotes time-invariate observed covariates we want to match in addition to the outcomes. Then, we estimate the treatment and control weights by solving a program similar to the one proposed

by Abadie and Zhao (2021),

$$\begin{aligned}
&\min_{\boldsymbol{w},\boldsymbol{v}} \left\| \bar{\boldsymbol{X}} - \sum_{j=1}^{J} w_j \boldsymbol{X}_j \right\|_H^2 + \xi \left\| \sum_{j=1}^{J} w_j \boldsymbol{X}_j - \sum_{j=1}^{J} v_j \boldsymbol{X}_j \right\|_H^2 + \xi\lambda \sum_{j=1}^{J} v_j \left\| \sum_{i=1}^{J} w_i \boldsymbol{X}_i - \boldsymbol{X}_j \right\|_H^2, \\
&\text{s.t.} \quad \sum_{j=1}^{J} w_j = 1, \\
&\qquad \sum_{j=1}^{J} v_j = 1, \\
&\qquad w_j, v_j \geq 0, w_j v_j = 0, \text{ for all } j = 1, \ldots, J, \\
&\qquad \|\boldsymbol{w}\|_0 = m,
\end{aligned} \tag{1}$$

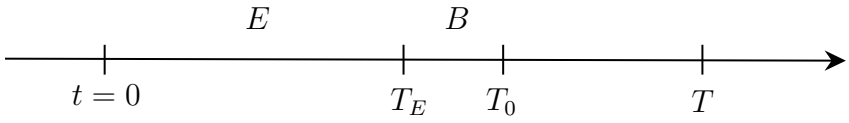
where $\|A\|_H^2$ denotes the weighted norm $A'HA$ for a diagonal matrix H with diagonal elements $h_1, \ldots, h_{dim(A)}$. The h coefficients can be chosen by researchers to encode preferences over which predictors/outcomes are more important.

The first term in the program represents the first objective of the experimenter, achieving a representative treated unit, and the second term measures the discrepancy between the treated and the synthetic control. The preference of the researcher for each loss is parameterized by ξ, if $\xi = 1$ then the loss function puts equal weight on the discrepancies between the treated and the population average, and the control and the population average. The λ-term is the penalty term from Abadie and L'Hour (2021). The penalty term may be particularly useful in cases in which the synthetic treated falls in the convex hull of the control units. This is likely to happen when there are many potential controls available (large J) and, in these cases, the standard synthetic control will be dense and include many donor units. Setting $\lambda > 0$ will make the choice of $\boldsymbol{v}$ sparse and interpretable. Finally, the constraint $\|\boldsymbol{w}\|_0 = m$ ensures that the number of treated is exactly m and $w_j v_j = 0$ that the treated and control groups do not overlap.

Inference Valid inference can be performed on estimators of the weighted average treatment effect given by weights $\boldsymbol{w}^*$ and $\boldsymbol{v}^*$ that solve the above minimization program and are approximately optimal in the sense that

$$\bar{\boldsymbol{X}} - \sum_{j=1}^{J} w_j^* \boldsymbol{X}_j \approx \boldsymbol{0} \quad \text{and} \quad \bar{\boldsymbol{X}} - \sum_{j=1}^{J} v_j^* \boldsymbol{X}_j \approx \boldsymbol{0},$$

Figure 1: Experimental design timeline



ensuring that the two goals of the experimenter are achieved. Furthermore, as proposed by Abadie and Zhao (2021), we divide the pre-intervention period in a *fitting period* $E = 1, \ldots, T_E$ and a *blank period* $B = T_E + 1, \ldots T_0$. Figure 2 depicts the timeline of the experimental design. In general, it is not necessary to separate sequentially the fitting and the blank periods, however, in practice it facilitates the pre-treatment fit check. After T_0, once the intervention has been assigned to the treatment units, we define the synthetic control estimator of the average treatment effect for outcome k as

$$\hat{\tau}_t^k(\boldsymbol{w}^*, \boldsymbol{v}^*) = \sum_{j=1}^{J} w_j^* Y_{jt}^k - \sum_{j=1}^{J} v_j^* Y_{jt}^k, \quad \text{for } t > T_0.$$

The main idea behind the inference procedure in Abadie and Zhao 2022 is that under the null hypothesis $H_0 : Y_{jt}^k(0) \sim Y_{jt}^k(1)$ for $t > T_0$, that is that the outcomes under intervention coincide with the outcomes in absence of the intervention, the distribution of the "placebo" treatment effects in the blank period B and the distribution of the real treatment effects after T_0 should coincide. The inferential procedure then consists of a permutation test in which we compare the estimated treatment effects in $t > T_0$ with permuted treatment effects including the placebo effects from the blank periods.

Formally, let the set of $(T - T_0)-$combinations of the time indices $T_E, \ldots, T$ be denoted by Π. Then, for each combination $\pi \in \Pi$ define the estimated residual as

$$\hat{\boldsymbol{e}}_\pi^k = (\hat{u}_{\pi(1)}^k, \ldots, \hat{u}_{\pi(T-T_0)}^k),$$

where $\hat{u}_t = \sum_{j=1}^{J} w_j^* Y_{jt}^k - \sum_{j=1}^{J} v_j^* Y_{jt}^k$ for $t \in \{T_E, \ldots, T\}$. Similarly, the estimated treatment effect vector is

$$\hat{\boldsymbol{\tau}}^k = (\hat{\tau}_{T_0+1}^k, \ldots, \hat{\tau}_T^k).$$

The permutation test statistic is defined as an average of the absolute residuals, given a

$(T - T_0)$ vector $\boldsymbol{e}$,

$$S(\boldsymbol{e}) = \frac{1}{T - T_0} \sum_{t=1}^{T-T_0} |e_t|,$$

analogously, the permutation $p-$value for outcome k is given by

$$\hat{p}^k = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbf{1}\{S(\hat{\boldsymbol{e}}^k) \geq S(\hat{\boldsymbol{\tau}})\}.$$

Under technical conditions Abadie and Zhao (2021) show that this p-value is approximately valid as $T_E \to \infty$ when the outcomes of interest are generated by linear factor models (the standard assumption in the synthetic control literature). Their results are conditional on a goodness of fit assumption that can be re-stated as follows for each outcome of interest k

Assumption 1 *Given outcomes Y_{jt} that depend on observed covariates Z_j, with probability one, for all $t \in E$*

$$\sum_{j=1}^{J} w_j^* Y_{jt}^k = \sum_{j=1}^{J} f_j Y_{jt}^k \quad \textit{and} \quad \sum_{j=1}^{J} v_j^* Y_{jt}^k = \sum_{j=1}^{J} f_j Y_{jt}^k$$
$$\sum_{j=1}^{J} w_j^* Z_j^k = \sum_{j=1}^{J} f_j Z_j^k \quad \textit{and} \quad \sum_{j=1}^{J} v_j^* Z_j^k = \sum_{j=1}^{J} f_j Z_j^k$$

The addition of covariates Z_j highlights the importance to also match observed variables that affect the outcomes of interest. Conditional on this assumption holding for all k, and an appropriate multiple hypothesis testing correction (which is beyond the scope of this report), valid inference can be performed to evaluate the treatment effects on various potential outcomes of our experimental intervention.

Implementation In practice, for problems with small m but large control pools, it may be computationally attractive to solve the program lexicographically, first by finding the set of best synthetic treated units and then for the best treated units finding their respective synthetic controls. This is equivalent to solving program (1) jointly with $\xi \to 0$, ensuring we prioritize the fit of the treated units. The following algorithm summarizes the procedure.

Algorithm 1: Fast Synthetic Experimental Design

Result: $\hat{\boldsymbol{w}}, \hat{\boldsymbol{v}}$ for top N candidate tuples.

Data: design matrix $\boldsymbol{X}$, outcome/covariate weights $\boldsymbol{H}$, population weights $\boldsymbol{f}$, candidate set $\mathcal{M}$, number of treated m, time periods E, B, penalty λ.

generate all m-tuples from candidate set $\mathcal{M}$;

get $\tilde{\boldsymbol{X}} = \left[\frac{\boldsymbol{X}'_k - \mu_k}{\sigma_k}\right]_{k=1,\ldots,K}$, i.e. for each predictor k subtract the $\boldsymbol{f}$-weighted mean and divide by the s.d.;

get $\tilde{\boldsymbol{X}}^E$ by sub-setting $\tilde{\boldsymbol{X}}$ to experimental periods E;

for *each candidate m-tuple* **do**

  find $\hat{\boldsymbol{w}}$ by minimizing $\| \sum_{j \in \text{tuple}} w_j \tilde{\boldsymbol{X}}_j^E \|_H^2$ s.t. weight constraints;

  store loss associated with $\hat{\boldsymbol{w}}$;

end

for *the top N m-tuples with lowest loss* **do**

  given $\hat{\boldsymbol{w}}$ find $\hat{\boldsymbol{v}}$ by minimizing $\left\| \sum_j \hat{w}_j \tilde{\boldsymbol{X}}_j^E - \sum_j v_j \tilde{\boldsymbol{X}}_j^E \right\|_H^2 + \lambda \sum_j v_j \left\| \sum_i \hat{w}_i \tilde{\boldsymbol{X}}_i^E - \tilde{\boldsymbol{X}}_j^E \right\|_H^2$ s.t. weight and exclusion constraints;

end

collect the top N candidate tuples $(\hat{\boldsymbol{w}}, \hat{\boldsymbol{v}})$;

report summary statistics over the E and B time periods;

3. Choosing the outcomes, the data and the exclusion criteria

The main goal of the UBI pilot study using synthetic experimental design is to estimate the general equilibrium effects of the UBI on outcomes related to education, health, the labor market and the use of public services subject to implementation constraints (i.e. a government spending constraint). The main challenge with choosing the set of outcomes Y^k for $k = 1, \ldots, K$ is that the dimension of the matching problem increases with the number of outcomes. Considering more outcomes of interest will come at the cost of the fit for each outcome. With this in mind, our preferred approach is to choose one outcome for each vertical. An assumption justifying this approach is that the chosen outcomes are generated by a very similar factor model to the other outcomes in the vertical. Under this assumption our synthetic model will also be approximately valid for all the other outcomes in the vertical. The set of criteria to choose our outcomes of interest can then be summarized as follows

Outcome criteria:

1. **Representation**: for each vertical, the outcome of interest should be representative of the overall trends in the vertical.

2. **Strength of treatment**: each outcome of interest should be affected by the UBI treatment in a non-mechanical way. For example, income (while an important covariate) is mechanically affected by the UBI. On the other hand, the graduation rate for mandatory schooling is unlikely to be affected by the treatment.

3. **Data availability**: outcomes of interest should be available for all candidate treatment units, for a number of pre-treatment periods and have limited measurement error given enforcement.

Our experimental design is constrained by treatment criteria that restrict the set of candidate units. In particular,

Treatment criteria:

1. **Number of units**: two units are treated, i.e. $m = 2$.

2. **Unit population**: each unit should have a population between 1200 and 1400 inhabitants.

3. **No interference**: treated units can not belong to the same province.

4. **No known shocks**: treated units can not have received an unexpected shock (e.g. large lottery win) during the pre-treatment periods.

5. **No dis-aggregated units**: units consisting of multiple population nuclei should have one nuclei with at least 1000 inhabitants.

The outcome criteria are derived from budget constraints regarding the maximum expense of the UBI pilot study and statistical constraints. The number of units to be treated is set to 2 instead of 1 to minimize the risk of unexpected shocks affecting the only treated unit and improve the odds of adequately reproducing the population average. The statistical constraints ensure that the treatment effects can be identified and estimated. These statistical criteria are closely related to the criteria for the population and control groups.

Population criteria:

1. **Target population restrictions**: the set of population units consists of all villages and semi-dense towns in Catalonia (according to Eurostat (2021)).

The population criteria define the *scope* of the experiment. The target parameter is limited to be the average treatment effect for villages and semi-dense towns, excluding big cities, for which we have data for all of our outcome variables. This means our target population includes 852 of the 947 towns in Catalonia at the time of this report[2] and 40% of the population (3 million people). The reason why we restrict the scope of the trial is that (1) it would be impossible to replicate big cities like Barcelona with a convex combination of two towns of less than 1500 inhabitants and (2) the general equilibrium effects of the UBI are likely very different between small towns and large dense cities.

Control criteria:

1. **Exclusion restriction**: control units can not be within the same "labor market", defined by being in a different region (comarca) as the treated units.

2. **Sparsity of controls**: the generated synthetic controls must be sparse and supported on a few units.

The control exclusion criteria limits the interference between control units and treated units. By limiting controls to be outside the labor market of the treated units we ensure that the UBI treatment does not spill over to the control units and contaminate our estimated treatment effect. The sparsity criteria requires the synthetic controls to only include a few units with positive weight. The rationale behind this constraint is to allow for pre and post intervention surveys to be collected also in the control units. Given the budget constraints involved in running surveys for a large amount of units, we limit the set of potential synthetic control units to be sparse.

Given our set of *outcome*, *treatment*, *population* and *control* criteria we choose the following outcomes and data sources

The data:

1. **Outcome variables**

   - **Unemployment rate**: constructed at the town level with data from Idescat and the *Departement d'Empresa i Treball*. Available from 2012 to 2021.

[2]The number of villages and semi-dense towns is 905, but after removing villages with missing values for our main outcomes of interest we are left with 852.

- **Share of hospitalizations**: number of hospitalizations relative to town population. Constructed with data from the *Agencia de Qualitat i Avaluacio Sanitaries de Catalunya (AQuAS)*. Available from 2012 to 2021.
- **Share of users of social services**: number of users of social services relative to town population. Constructed with data from the *Departament de Drets Socials de la Generalitat de Catalunya*. Available from 2012-2014, 2016-2017 and 2019-2020.[3]
- **Entrance rate into *Batxillerat***: share of population that after mandatory formal education (until 16 years old), enrol in the optional *Batxillerat* high school program in public schools. Constructed with data from the *Departament d'Educacio de la Generalitat de Catalunya*. Available from 2017 to 2021.

2. **Covariates**

- Average age at the town level. Constructed with data from the *Estadistica del Padro continu* of the *INE*. Available from 2012 to 2021.
- Percentage of foreign population at the town level. Constructed with data from the *Padro municipal d'habitants de l'Idescat*. Available from 2012 to 2021.
- Average net income at the town level. Constructed with data from the *Atlas d'estadistica experimental de l'INE*. Available from 2015 to 2019.
- Gini coefficient. Constructed with data from the *Atlas d'estadistica experimental de l'INE*. Available from 2015 to 2019.

Given the set of *treatment* and *population* criteria, we have 15 candidate units for treatment.

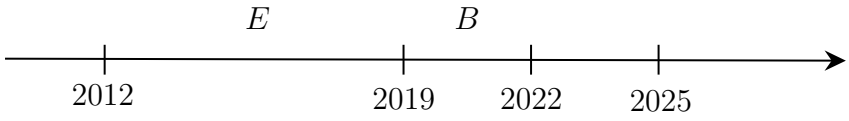
The treatment candidates:

- Almoster, Anglesola, Bellcaire d'Urgell, Corçà, La Fuliola, Miralcamp, Perafort, El Pla del Penedès, Puigverd de Lleida, Sant Julià del Llor i Bonmatí, Sant Martí de Tous, Térmens, Torà, Vallbona d'Anoia and Vila-rodona.

Finally, we consider the timeline of the Catalan UBI pilot study experimental design. Figure 2 details the timeline. In short, the treatment will happen in 2023, years 2012 to 2018 (avoiding the covid years) are used to design the experiment (periods E) and 2019 -

[3] 2018 removed as data may suffer from severe measurement error/corruption, as several units show extreme values.

Figure 2: Timeline Catalan UBI Pilot



2022 will be used as blank periods (periods B) to evaluate the design and perform valid inference (currently only data from 2019-2021 is available). The experiment will then be evaluated during the treatment years 2023-2024 and after the treatment periods in 2025 and beyond. The number of blank periods is chosen to be 4 to ensure that the lower bound of the approximate p-value $\hat{p}^k \geq 1/|\Pi|$ is below 0.05. A smaller number of blank periods risks not being able to do inference. Indeed, having only 3 blank periods will give us a lower bound of exactly 0.05 when only 3 post-treatment periods are available. On the other hand, given that we don't have many pre-treatment periods choosing a number of blank periods greater than 4 will increase the risk that we over-fit to the E periods and that our blank period fit is bad making inference more difficult.

4. Designing the Catalan UBI Synthetic Experiment

To find the best synthetic experiment for the Catalan UBI pilot study, we solve program (1) by running algorithm 1 as described in section 2. We proceed in this manner to simplify the computational problem and because given that we have 15 candidates but 852 potential controls the hard part of program (1) is minimizing the first term. In particular, we use Algorithm 1 with the following parameters:

Parameters:

1. **Data**: outcome variables defined in section 2 and with covariates values averaged over 2012-2018.

2. **Outcome weights**: all outcomes receive the same weight, with covariates receiving 1/10th of the weight. This is done to ensure we choose the weights that best reproduce all outcomes, and amongst the weights that have good outcome fit the ones that also balance the covariates. The results are robust to changing the covariate weightings.

3. **Population weights**: given our *population criteria* in section 2 we choose a uniform weighting such that the target parameter is a simple average treatment effect.

4. **Candidate set**: given in section 2, including 15 units and 105 candidate pairs.

5. **Treatment number**: $m = 2$.

6. **Time periods**: experimental design time periods E are 2012 to 2018, and the blank time periods B are 2019 to 2021.

7. **Penalty**: $\lambda = 0.1$. As shown in Abadie and L'Hour 2021, $\lambda \to \epsilon > 0$ is sufficient to select a sparse solution. The results are robust to different choices of λ.

Results and evaluation We run the procedure described in Algorithm 1 for the given parameters. This allows us to identify the preferred candidate pair to be treated as the candidate pair with the lowest mean squared error in the experimental periods 2012-2018 (i.e. the best fit).

- **The preferred pair**: A and B with weights of w_A and w_B respectively.
- **The preferred synthetic control**: a convex combination of five towns.[4]

To evaluate how good the fit is of the preferred pair, and check whether assumption 1 is likely to hold, we consider a normalized version of the mean squared error (NMSE) for a set of time periods T,

$$\text{NMSE}_T\left(\bar{\boldsymbol{Y}}^k, \hat{\boldsymbol{Y}}^{k,synth}\right) = \frac{1}{|T|}\sum_{t\in T}\left(\frac{\bar{Y}_t^k - \hat{Y}_t^{k,synth}}{\sigma_t^k}\right)^2,$$

where $\bar{Y}$ denotes the population average, $\hat{Y}^{synth}$ denotes the weighted average for the synthetic treated units and σ_t^k is the standard deviation over units for the outcome of interest each year. We use the NMSE because we want to be able to compare the MSE for different outcomes in the same scale.

Table 1 displays normalized mean squared errors averaged for the time periods 2012 to 2018 for the outcome variables[5]. The fit of the synthetic treated unit (A-B) is very good for all outcome variables. In the cases of the unemployment rate and the hospitalizations rate it is of the order of 0.01 (with 1 being the standard deviation of each variable), with the NMSE being close to 0.03 for the social service usage and 0.09 for the *Batxillerat* enrollment rate.

[4]Table 7 has the list of the top 10 units with most weight in the synthetic control, almost all weight is concentrated in the top 5.

[5]This is the target loss function for $\tilde{\boldsymbol{X}}$ used in Algorithm 1.

Table 1: Synthetic Experiment: pre-treatment fit

Outcome fit	NMSE (ST, Pop)	NMSE (ST, SC)
Unemployment rate	0.008	0.004
Batxillerat enrollment rate	0.088	0.077
Hospitalizations rate	0.010	0.014
Social service usage rate	0.030	0.012

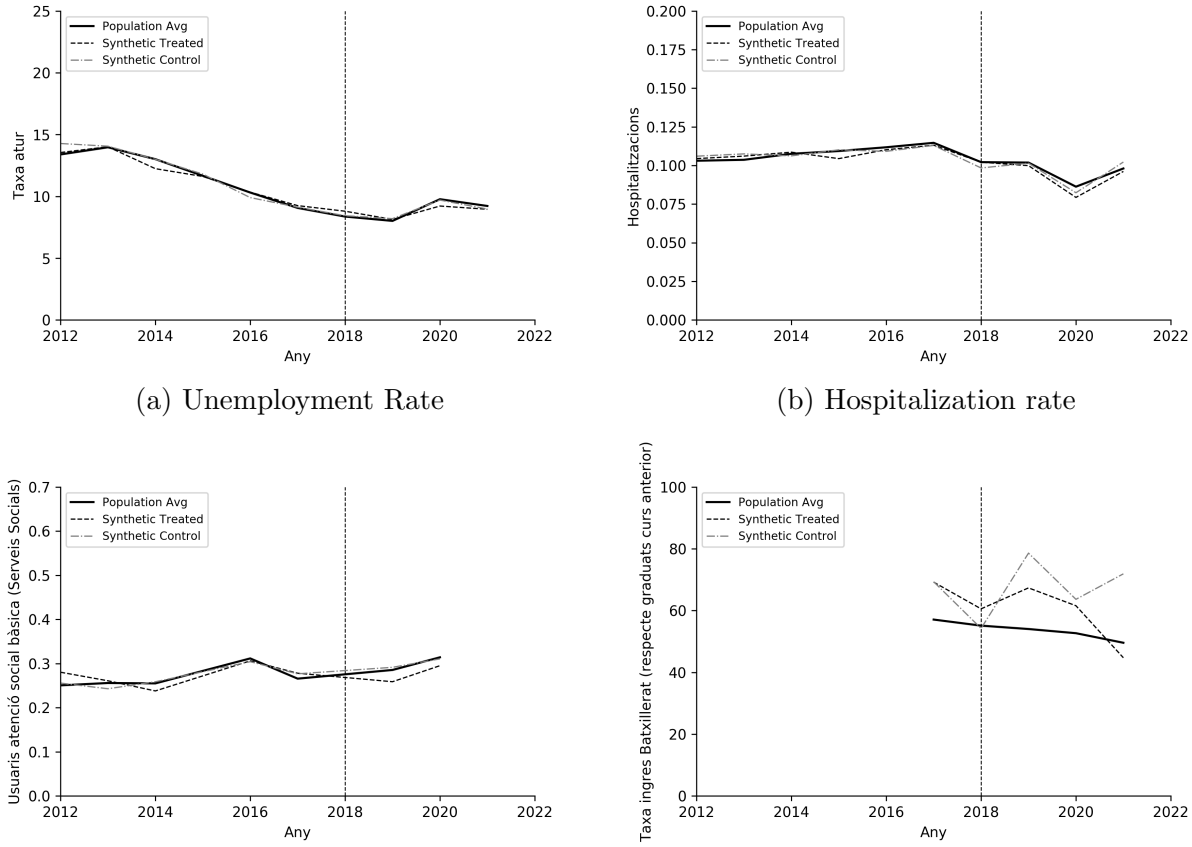
Notes: Normalized mean squared error between the synthetic treated unit (ST) and the population (Pop) and synthetic control (SC) for the four outcomes of interest.

The fit is also good between the synthetic control and the synthetic treated unit with similar magnitudes. This suggests that assumption 1 for the outcomes of interest is likely satisfied for three outcome verticals. For the *Batxillerat* enrollment rate it might be necessary to use a bias correction such as the one proposed in Abadie and L'Hour 2021 or use a different inference procedure.

While covariate fit is a secondary goal of our analysis, Table 4 in the appendix shows the mean covariate values for the population, synthetic treated, and synthetic control over the 2012-2018 period. The synthetic treated unit has covariate values close to the population. In all cases, in underestimates the population average by a small amount when compared to the respective standard deviation of each covariate. In other words, while the pair (A-B) almost perfectly replicates the outcomes of the population, it does so while having a slightly younger population, with a lower net income and Gini coefficient and with slightly less foreign inhabitants.

Figure 3 shows the outcome fit for every time period. This is a more intuitive way to evaluate the fit of our model related directly to Assumption 1, akin to the checking the fit in synthetic control studies or the parallel trends assumption in Diff-in-Diff designs. Each panel in Figure 3 plots five lines: the dark black line is the population average (our target), the dashed black line the synthetic treated unit, the dashed gray line the synthetic control and the other two lines are the outcome values for each unit in the preferred pair (A and B). As can be seen in panels (a)-(c) the fit is good, with both the synthetic treated and synthetic control being closed to the population. Furthermore, each unit in the preferred pair has outcomes also close to the population, meaning that our model does not need to interpolate significantly to achieve a good fit. This is good because it means that were there

Figure 3: Pre-treatment outcome fit



(a) Unemployment Rate

(b) Hospitalization rate

(c) Social service usage rate

(d) *Batxillerat* enrollment rate

Notes: Each panel plots the outcome values for 2012-2021 for the synthetic treated unit, the synthetic control and the population average. The vertical dashed line at 2018 indicates the end of the experimental period (E).

to be a shock/anomaly in one of the units in the pair, we would be able to use the other unit to perform our statistical analysis.

We run our analysis for the top 15 pairs, in Table 8 in the appendix we provide the total normalized NMSE across all outcomes and the covariates for the top 15 pairs. The total NMSE fit between the synthetic treated and the population for the preferred pair is 0.0440 while the average for the top 15 pairs is 0.0694, the standard deviation is 0.0173 and the maximum value is 0.0988, more than double the NMSE of the best pair. Given the results reported in Table 1 and Figure 3 the clear recommendation is to use the preferred pair (A and B) in the design of the UBI pilot study, other choices risk using pairs that won't satisfy Assumption 1 for the main outcomes and will difficult inference. Furthermore, all pairs in the top 10 include A or B (Table 8), with the fit decreasing significantly for pairs in the top 15 without either unit. Despite this, in the next section we discuss guidelines on how a lottery could be used to select the pair given our model results.

Blank period fit and detectable effects While we can't decide the outcome pair on the basis of the fit in the blank period (otherwise we would bias our statistical analysis). We can use the blank periods in 2019-2021 to evaluate whether our model is likely to have over-fitted to the experimental periods. If the fit were bad in the blank periods it could be that our inference procedure would yield biased results. Figure 3 shows that this is likely not the case as the fit is also good in the blank periods (after 2018, the dashed line). Furthermore, the NMSE for the blank periods are similar in magnitude to the ones reported in Table 1, as can be seen in Table 2, and the total NMSE is 0.0559, also close to the value for the experimental periods. The only exception is the SC fit for the *Batxillerat* enrollment rate which indicates we may have to compute a separate synthetic control for this variable. This implies that (1) our method likely did not over-fit to the experimental periods 2012-2018 and (2) that we may be able to detect reasonable treatment effects.

The blank period fit allows us to simulate potential approximate p-values assuming different levels of treatment effects. As in the case of the Catalan UBI pilot study we consider a setting in which we only have three post-experiment periods, and we assume a constant average treatment effect over time, $\tau_t^k = \tau^k$ for $t \in \{T_0+1, T_0+2, T_0+3\}$. Then, we compute the $\hat{p}^k$ approximate p-value described in section 2 using the blank period $\hat{u}_t$s we obtain from our model and the simulated treatment effects τ_t^k. Given that we don't observe at the time of the experimental design the values for 2022 we consider the p-value when $\hat{u}_{2022}$ is imputed with the mean, max and twice the max of the $\hat{u}_t$s observed in the blank periods 2019-2021.

Table 2: Synthetic Experiment: blank period fit

Outcome fit	NMSE (ST, Pop)	NMSE (ST, SC)
Unemployment rate	0.009	0.003
Batxillerat enrollment rate	0.091	0.384
Hospitalizations rate	0.072	0.031
Social services usage rate	0.049	0.004

Notes: Normalized mean squared error between the synthetic treated unit (ST) and the population (Pop) and synthetic control (SC) for the four outcomes of interest in the blank periods 2019-2021.

Table 3: Synthetic Experiment: minimum detectable effects

Outcome	mean $\hat{u}_t$	max $\hat{u}_t$	2×max $\hat{u}_t$
Unemployment rate	0.1	0.1	0.2
Batxillerat enrollment rate	0.87	0.87	None
Hospitalizations rate	0.26	0.26	0.53
Social service usage rate	0.16	0.16	0.32

Notes: Minimum τ^k relative to one standard deviation needed for $\hat{p}^k \leq 0.05$ for different assumptions on how to impute the $\hat{u}_t$ for $t = 2022$. 'mean' and 'max' are taken over the values of $\hat{u}_t$ for the periods 2019-2021.

Table 3 shows the ATE levels (τ^k) needed for the approximate p-value to be smaller or equal to 0.05. It is important to note that the approximate p-value is bounded below by the $1/|\Pi| = 0.0285$ in the case with 4 blank periods and 3 post-treatment periods. Overall Table 3 shows that our experimental design should be able to detect effects of moderate size, with minimum effects between 0.1 and 0.3 standard deviations for the unemployment rate, the hospitalizations rate and the social service usage rate. It should not be surprising that given the small number of time periods and high variance of the education outcome, the treatment effect has to be at least 0.87 standard deviations. While these numbers might seem high, putting them in context reveals that these effects are associated with moderate level changes (except in the case of *Batxillerat* enrollment). Indeed, a 0.1 s.d. increase in the unemployment rate is equivalent to a 0.5 point increase (s.d. is 4.8), a 0.26 s.d. increase in the hospitalizations rate is also equivalent to a 0.5 point increase (s.d. is 2), a 0.16 s.d.

increase in the social service usage is equivalent to a 3.2 point increase (s.d. is 20) and a 0.87 increase in the *Batxillerat* enrollment rate is equivalent to a 27.5 point increase (s.d. is 31.6).

5. Want to do a lottery?

Often in public policy evaluation allocation fairness is an important consideration. Lotteries may be an appealing instrument to ensure allocation outcomes are ex-ante fair. However, there is no free lunch. Lotteries and fairness considerations come at the cost of efficiency. In our setting, we can model the trade-off between fairness and efficiency as a simple constrained optimization problem.

Let an allocation rule $x : \mathcal{M} \to [0, 1]$, for the set of candidate units $\mathcal{M}$, with $J = |\mathcal{M}|$ such that $\sum_{m\in\mathcal{M}} x(m) = 1$, denote distribution of selection probabilities into the pilot study. We say that an allocation is δ-*fair* if its Bhattacharyya distance to the uniform distribution, denoted U_J, is δ

$$D_B(x, U_J) = -\log\left(\frac{1}{\sqrt{J}}\sum_{m\in\mathcal{M}}\sqrt{x(m)}\right) = \delta(x).$$

In the case in which the allocation x is uniform the distance becomes 0 and we say the allocation is *fair*. On the other hand, we also have a statistical model $f : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ that given a pair of units gives us the statistical fit (mean squared error) of the model. Therefore, given an allocation rule x and model f we say that the expected fit is

$$\mathbb{E}_f[x] = \sum_{(m,n)\in\mathcal{P}(\mathcal{M})} \frac{x(m)x(n)}{\sum_{(m,n)\in\mathcal{P}(\mathcal{M})} x(m)x(n)} f(m, n),$$

where $\mathcal{P}(\mathcal{M})$ denotes the set of unique pairs given units $\mathcal{M}$, with $|\mathcal{P}(\mathcal{M})| = J(J-1)/2$. Observe that $\mathbb{E}_f[x]$ is just a weighted average of the fits of each pair. Another candidate evaluation function is the worst case fit

$$\mathcal{W}_f[x] = \max_{m,n,\text{ s.t. } m\neq n} x(m)x_{-m}(n)f(m, n).$$

In practice, it is useful to normalize the worst case fit relative to the best fit amongst all pairs:

$$r_f(x) = \frac{\mathcal{W}_f[x]}{\min_{m,n} f(m, n)}.$$

A relative worst case fit of 1 means that the allocation will yield the pair that minimizes

the fit. Similarly, a relative worst case fit of 2 means that the allocation x has a chance of increasing the mean squared error by a factor of 2. An intuitive objective function for policy makers is one that achieves the best *fairness* conditional on a maximum relative worst case fit:

$$X^W(\beta, f) = \{\arg\min_{x \in \mathcal{X}} \delta(x) \quad \text{s.t. } r_f(x) \leq \beta\}.$$

Our recommendation would be to choose and allocation in $X^W(\beta, f)$, however a simple theoretical result shows that this set might be empty in practice (proof in the appendix).

Lemma 1 *The optimal allocations in $X^W(\beta, f)$ are uniform distributions supported in all units that are in a pair below the worst-fit threshold. That is,*

$$X^W(\beta, f) = \{U(\{x_1, \ldots, x_b\}) | \text{ for any } x_i, x_j, f(x_i, x_j) < \beta \min_{m,n} f(m, n)\}$$

where $U(\{x_1, \ldots, x_b\})$ denotes the uniform distribution supported on a set of units $\{x_1, \ldots, x_b\}$.

This result implies that in many applications $X(\beta, f) = \emptyset$ when β is small or the fit depends heavily on a few units. Unfortunately, this is our case.

– **Guideline 1 − no lottery will be fair and ensure a good fit**: In our setting $X^W(2, f) = \emptyset$. We cannot simultaneously ensure a fair allocation and a mean squared error that is at most 2 times greater than the lowest mean squared error.

Given the negative result implied by Guideline 1, we consider next a weaker constraint on the trade-off between fairness and efficiency. Let the set of allocations that have best possible fit conditional on being at least $\delta-$fair

$$X^E(\beta, f) = \{\arg\min_{x \in \mathcal{X}} \mathbb{E}_f[x] \quad \text{s.t. } D_B(x, U_J) \leq \delta\}.$$

This set of distributions is harder to characterize. However, if we consider the fairness at the *pair* level and if the following condition holds

$$-\log\left(\frac{1}{\sqrt{J(J-1)/2}} \frac{1}{\sqrt{\sum_{(m,n)\in\mathcal{P}(\mathcal{M})} 1/f(m,n)}} \sum_{(m,n)\in\mathcal{P}(\mathcal{M})} \sqrt{f(m,n)}\right) \leq \delta,$$

then, an allocation that minimizes risk while being *pair* δ-fair is the one that inversely weights each pair according to fit. That is,

$$X^{\text{weighted fit}} = \left\{ x : \text{ for any } a, b \; x(a)x(b) = \frac{1/f(a,b)}{\sum_{(m,n)\in\mathcal{P}(\mathcal{M})} 1/f(m,n)} \right\}.$$

In our setting, given that the NMSE drops uniformly in Table 8, we might consider using an allocation in $X^{\text{weighted fit}}$ to trade-off fit and fairness. However, the problem is that this allocation will be far from *fair* in the Bhattacharyya distance sense unless we consider all pairs, and the expected fit will be large unless we restrict to top pairs. This tension can not be resolved easily.

– **Guideline 2 − second best**: using a weighted fit allocation ($X^{\text{weighted fit}}$) for the top 15 pairs will yield an expected total NMSE in the experimental period of 0.0656 (50% larger than for the preferred pair), in the blank period of 0.0865 (54% larger than for the preferred pair) and will be at least 0.98-*fair*. Considering fairness with respect to only the top 15 pairs means the allocation is 0.007-*fair*, closer to the uniform distribution.

Guideline 2 highlights the trade-off between fairness and efficiency for a reasonable choice of allocation. Its main takeaway is that doing a lottery requires an important relaxation of the notion of fairness, and even then, may come at cost of a decrease in fit of, on average, 50%. Together with Guideline 1, this suggests that this particular setting may be ill suited for a design that includes a lottery step, as fairness can not be reconciled with statistical fit.

6. Conclusion

This report provides a guide on how to design an experimental policy intervention using synthetic designs. It explains the theory of synthetic experimental design and details the practical steps researchers should take to bring the theory to practice. The report is concerned with the Catalan UBI pilot study design in which two towns are to receive a UBI treatment at the start of 2023. The research question is how to choose these two towns and how to do inference on a variety of outcomes after the UBI experiment finishes. The report advocates for the use of synthetic experimental design and shows that there exist a pair of towns (A and B) that replicate the population outcomes of interest in the years leading to treatment and have good associated synthetic controls. Finally, the report concludes with

a discussion on why a lottery in the Catalan UBI setting might not be suitable to trade-off fairness and efficiency.

References

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.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132.

Abadie, A. and L'Hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*. Forthcoming.

Abadie, A. and Zhao, J. (2021). Synthetic controls for experimental design. Papers, arXiv.org.

Amjad, M., Misra, V., Shah, D., and Shen, D. (2019). Mrsc: Multi-dimensional robust synthetic control. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(2).

Doudchenko, N., Khosravi, K., Pouget-Abadie, J., Lahaie, S., Lubin, M., Mirrokni, V., Spiess, J., and Imbens, G. (2021). Synthetic design: An optimization approach to experimental design with synthetic controls.

Trejo, S., Yeomans-Maldonado, G., and Jacob, B. (2021). The psychosocial effects of the flint water crisis on school-age children. Working Paper 29341, National Bureau of Economic Research.

Table 4: Synthetic Experiment: covariate fit

Covariate fit	Mean (Pop)	Mean (ST)	Mean (SC)
Net income (euros)	12320.0	11188.7	14679.4
Percentage foreign	21.2	20.5	17.5
Average age	43.9	42.9	41.9
Gini coefficient	30.3	28.9	29.7

Notes: Means of the covariates over the 2012-2018 period for the synthetic treated unit, the synthetic control and the population.

Table 5: Synthetic Experiment: blank period fit with two blank periods

Outcome fit	NMSE (ST, Pop)	NMSE (ST, SC)
Unemployment rate	0.0085	0.0023
Hospitalizations rate	0.0945	0.0125
Social services usage rate	0.0327	0.0000
Batxillerat enrollment rate	0.0421	0.1009

Notes: Normalized mean squared error between the synthetic treated unit (ST) and the population (Pop) and synthetic control (SC) for the four outcomes of interest in the blank periods 2020 and 2021.

A.1. Additional Tables

A.2. Model results for top 15 pairs

A.3. Lottery results proofs

Lemma 1 Proof

If we were in a continuous setting, at the optimum the constraint would always bind because otherwise there would exist a profitable deviation that lowers $\delta(x)$ while satisfying the constraint. This implies that in the discrete case we need only consider the set of allocations that give zero probability to units that only appear in pairs with fit worse than $\beta \min_{m,n} f(m,n)$. Amongst these allocations we can't include any distribution that assigns positive probability to two units that have

Table 6: Synthetic Experiment: pre-treatment fit with two blank periods

Outcome fit	NMSE (ST, Pop)	NMSE (ST, SC)
Unemployment rate	0.0113	0.0152
Hospitalizations rate	0.0079	0.0180
Social services usage rate	0.0314	0.0318
Batxillerat enrollment rate	0.0669	0.1561

Covariate fit	Mean (ST)	Mean (SC)	Mean (Pop)
Age	42.9	44.5	43.9
Percentage foreign	20.7	23.4	21.3
Net income (euros)	11325	11530	12558
Gini coefficient	29.0	28.0	30.0

Notes: The upper panel describes the normalized mean squared error between the synthetic treated unit (ST) and the population (Pop) and synthetic control (SC) for the four outcomes of interest. The lower panel reports the means of the covariates over the 2012-2019 period for the synthetic treated unit, the synthetic control and the population.

a pair with fit greater than $\beta \min_{m,n} f(m,n)$, i.e.

$$x \in X^W(\beta, f) \implies \text{ for any a,b with } x(a), x(b) > 0, f(a,b) < \beta \min_{m,n} f(m,n).$$

In other words, any units that have positive probability under the allocation rule need to have *all* pairs satisfy the worst case fit constraint. Within this set, by the definition of D_B it follows that the optimal allocations are

$$X^W(\beta, f) = \{U(\{x_1, \ldots, x_b\}) | \text{ for any } x_i, x_j, f(x_i, x_j) < \beta \min_{m,n} f(m,n)\}$$

where $U(\{x_1, \ldots, x_b\})$ denotes the uniform distribution supported on a set of units $\{x_1, \ldots, x_b\}$. This result implies that in many applications $X(\beta, f) = \emptyset$ for small β parameters.

Table 7: Synthetic Control: top 10 units

Name	Weight
1	0.296
2	0.285
3	0.237
4	0.089
5	0.085
6	0.007
7	0.002
8	0.000
9	0.000
10	0.000

Notes: Synthetic control for the preferred pair. Includes the weights of the top 10 units with weights rounded to 3 decimal places.

Table 8: Synthetic Experiment: top 15 pairs

Pair	Weights	Total NMSE
(A, B)	(w_A, w_B)	0.044
2	(0.13, 0.87)	0.051
3	(0.34, 0.66)	0.054
4	(0.15, 0.85)	0.056
5	(0.11, 0.89)	0.058
6	(0.09, 0.91)	0.061
7	(0.05, 0.95)	0.062
8	(0.06, 0.94)	0.062
9	(0.01, 0.99)	0.064
10	(0.26, 0.74)	0.080
11	(0.57, 0.43)	0.081
12	(0.69, 0.31)	0.084
13	(0.47, 0.53)	0.088
14	(0.29, 0.71)	0.097
15	(0.48, 0.52)	0.099

Notes: For each of the top 15 pairs from Algorithm 1, this table reports the weight each unit receives in the synthetic treated pair and the total normalized mean squared error between the population average and the synthetic treated unit in the experimental period.