

Interpolation Bias in Synthetic Controls: Evidence from Carbon Tax Evaluations

Sachintha Fernando

Martin Luther University Halle-Wittenberg, *sachintha.fernando@wiwi.uni-halle.de*

Version: Work in progress, September 2026

Abstract

The synthetic control method requires researchers to specify a donor pool from which the synthetic control is constructed. However, the existing literature provides few objective criteria for doing so. The guidance offered is that the donors should resemble the treated unit with respect to the predictors of the outcome and that smaller pools made up of more similar units are preferable. In practice, however, published ex-post evaluations of carbon taxes use donor pools that vary substantially in size and composition, even between studies of the same policy or with near-identical designs applied to comparable countries. I replicate three evaluations from this literature: two on carbon taxes on transport emissions in Finland and the UK, and one on the 2001 increase in the Swedish carbon tax and its effect on residential emissions. In all three cases, the estimated effect depends on a single donor country that is arguably structurally dissimilar to the treated unit yet receives a positive weight. I identify this donor by re-estimating the synthetic control on many subsets of the donor pool, a procedure that reveals influential donors regardless of their weight. Excluding this donor substantially attenuates the estimated effect in the two transport evaluations, rendering it indistinguishable from placebo effects, and in the residential evaluation leaves no convex combination of the remaining donors that reproduces the treated unit's pre-treatment path. This pattern is consistent with interpolation bias and is masked by a good pre-treatment fit, which applied work treats as a central prerequisite for applying the synthetic control method. These findings serve as a caution to applied researchers when using the synthetic control method. While the evidence here comes from carbon tax evaluations, this mechanism is not specific to that setting and may arise wherever a donor pool contains units that are structurally dissimilar to the treated unit.

Keywords: synthetic control method, interpolation bias, replications, carbon taxes, policy evaluation

Acknowledgments

I am grateful to Christoph Wunder, Amelie Wuppermann, Jörg Ankel-Peters, Hans Manner, Stefan Borsky, Jörn Kleinert, Tobias Eibinger, Thilina Fernando, and Markus Niedobitek, as well as seminar participants at the Leibniz Open Science Day 2024 (Berlin), the Applied Econometrics Workshop 2025 (Bamberg), the University of Graz Seminar Series 2025 (Graz), and the EMCC conference 2026 (Aalborg) for their valuable feedback on this project. I thank Jaume Vives-i-Bastida for sharing code from Abadie and Vives-i-Bastida (2025), and Torben Mideksa and Elisa Grieg for sharing their analytical datasets. I also thank Anton Cronet for excellent research assistance.

1 Introduction

The pricing of carbon dioxide (CO₂) is widely considered to be one of the most efficient methods of internalizing the negative externalities of CO₂ emitting activities (Pizer, 2002; Stern, 2008; Nordhaus, 2019). However, until recently, there had been a lack of rigorous ex-post evaluations to determine their real-world effectiveness (Green, 2021). The scarcity of ex-post evaluations was largely attributed to the methodological difficulty of identifying a baseline scenario against which to measure causal effects (Baranzini et al., 2000). However, recent advances in causal inference methods, particularly the development of the synthetic control method (SCM), have provided a solution to this challenge by offering a data-driven approach for constructing aggregate counterfactuals (Abadie & Gardeazabal, 2003; Abadie, Diamond, & Hainmueller, 2010, 2015; Abadie, 2021).

The SCM is considered a major innovation in the modern causal inference toolkit and is described by Athey and Imbens (2017) as “arguably the most important innovation in the policy evaluation literature in the last 15 years.” Developed in a series of seminal papers by Abadie and Gardeazabal (2003), and Abadie, Diamond, and Hainmueller (2010, 2015), the method is designed to estimate causal effects of interventions that occur at an aggregate level, affecting a small number (typically one) of large units, such as cities, regions, or countries. By constructing a convex weighted combination of untreated units, the SCM approximates the counterfactual trajectory that the treated unit would have followed in the absence of the intervention. The method’s core insight is that a data-driven, convex combination of comparison units offers a more credible counterfactual than any single unexposed unit alone.

In this paper, I demonstrate how the lack of objective guidance on selecting the donor pool, the set of units from which the synthetic control is constructed, manifests in empirical work. Drawing on the CO₂ tax evaluation literature, I show that studies examining closely comparable policies in similar settings arrive at donor pools that differ substantially in both size and composition. I then examine the implications of this discretion by replicating three empirical evaluations of CO₂ taxes.

The adoption of SCM in the ex-post evaluation of CO₂ taxes is no coincidence; rather, it reflects how closely the design of the estimator matches the jurisdictional level at which CO₂ taxation policies are implemented. As CO₂ taxation policies are inherently macro-level interventions applied across entire states, regions, or countries, they typically leave only a small number of aggregate comparison units, which limits traditional matching or difference-in-differences designs. This has made it difficult to construct a baseline against which to measure causal effects, and helps explain the historical scarcity of rigorous ex-post evaluations. However, since SCM was explicitly designed to estimate causal effects in precisely these aggregate, single-treated-unit settings, it has effectively unlocked the field, rapidly becoming the dominant method for evaluating CO₂ and energy taxes.

A growing body of empirical work using the SCM illustrates this trend. Andersson (2019) is one of the first to use the SCM to evaluate the impact of the Swedish CO₂ tax, implemented in 1991, on transport CO₂ emissions, attributing a 6% reduction in emissions to the tax. Pretis (2022) studies the CO₂ tax in British Columbia and finds no effect on aggregate emissions, but a 5% reduction in transport emissions, comparable to the estimate of Andersson (2019). Similarly, Runst and Höhle (2022) use the SCM to evaluate the German eco tax¹ introduced in 1999. Their baseline estimates imply that per capita transport CO₂ emissions in 2005 were about 12% below the synthetic counterfactual, and averaged over the post-treatment period 1999–2005, the reduction amounts to about 6%.² Mideksa (2024) applies the design of Andersson (2019) to Finland, which introduced a CO₂ tax in 1990, and finds an average reduction in transport CO₂ emissions of 19% over the post-treatment period, which, to my knowledge, is the largest reduction recorded in the transport sector. Bretschger and Grieg (2024) evaluate the fuel tax escalator, an implicit CO₂ tax introduced in the UK in 1993 and conclude that it reduced transport CO₂ emissions by about 13% on average. Moving to the power sector, Leroutier (2022) examines the carbon tax introduced in the UK power sector in 2013 and finds that emissions declined by 20% to 26% per year on average between 2013 and 2017. Xiang and Lawley (2019) estimate the effect of British Columbia’s CO₂ tax on per capita residential natural gas consumption and find a reduction of approximately 7%. Also studying the residential sector, Runst and Thonipara (2020) find that the 2001 increase in the Swedish CO₂ tax reduced residential CO₂ emissions by up to 800 kg per capita per year relative to the counterfactual built from countries with no or low CO₂ taxes. Averaged over the post-treatment period 2001–2016, the reduction amounts to about 0.61 tonnes, or about 58% below the synthetic counterfactual.³ Aus dem Moore et al. (2025) study the effect of the Swedish CO₂ tax on innovation and find an average annual increase of 7.96 clean transport patents between 1991 and 2005, an increase of approximately 59% relative to the synthetic control. Taken together, these studies share a common empirical structure: a single treated unit, an aggregate outcome, and a CO₂ tax (or an implicit CO₂ tax) as the policy intervention, which is precisely the setting for which the SCM was designed.

One of the key advantages of the SCM is that the weights assigned to donor units are chosen by a data-driven optimization procedure rather than by the researcher. This makes the construction of the counterfactual transparent, and it removes a substantial source of researcher degrees of freedom. However, discretion enters the procedure through two prior choices: the set of predictors used in the optimization and the composition of the donor pool. The consequences of the first choice are well documented, with many studies focusing on this issue (Ferman et al., 2020; Kaul et al., 2022; Vives-i-Bastida, 2022). Empirical practice is also beginning to converge on how to

¹Although officially an eco tax on fuels and electricity, its effect on motor fuel prices corresponds to an implicit CO₂ price (Runst & Höhle, 2022); see also Basaglia et al. (2025).

²Runst and Höhle (2022) report a reduction in transport CO₂ emissions of 0.2 to 0.35 tonnes per capita attributable to the tax. The corresponding percentage changes relative to the synthetic counterfactual are my own calculations.

³Own calculation based on the replication data provided by the authors.

handle it.⁴ The latter has received considerably less attention, and it is the focus of this paper. According to Abadie (2021), the donor pool should consist of units similar to the treated unit with respect to the predictors of the outcome, but these units must not have experienced a similar shock or policy intervention during the study period. The latter condition can usually be verified through expert knowledge and the institutional and policy context. In contrast, what constitutes a similar unit is considerably more open to interpretation.

The literature offers few objective criteria here, leaving researchers with substantial discretion over the initial donor pool from which the synthetic control is ultimately constructed. This discretion is consequential given how the SCM constructs its counterfactual. The convexity of the weights rules out extrapolation, but it offers no comparable protection against interpolation. The optimization rewards aggregate pre-treatment fit, which can be achieved by a combination of donors that are individually dissimilar to the treated unit. Individual similarity is not itself a requirement. The method relies on convex combinations precisely because no single donor needs to resemble the treated unit. However, structurally different donors can achieve this fit if their deviations from the treated unit offset one another in the weighted average. A good pre-treatment fit then masks their dissimilarity. These structural dissimilarities can then affect the trajectory of the counterfactual in the post-treatment period. Since the estimated effect is the difference between the treated unit and its synthetic counterpart, this divergence is interpreted as the effect of treatment, which is then overstated or understated. What the literature does offer is a robustness test in which each donor unit that receives positive weight is dropped in turn and the synthetic control re-estimated. If the estimated effect is largely unchanged across these leave-one-out estimates, the result is deemed robust to the exclusion of any single donor. This test is, however, inconsistently applied in empirical work.

In this paper, I first conduct comprehensive replications of Mideksa (2024) and Bretschger and Grieg (2024), who use the SCM to evaluate the effects of CO₂ taxes on transport emissions in Finland and the UK, respectively. Both studies report large effects. To assess whether these estimates depend on particular donors, I re-estimate the synthetic control on subsets of the donor pool. I define a donor as influential if, across these subsets, its inclusion consistently improves the pre-treatment fit and shifts the estimated effect, irrespective of which other donors are present. Influence differs from weight. A donor with a small weight can be influential if it is essential for achieving a good pre-treatment fit, while a donor with a large weight may have little influence if others can serve as substitutes. In each case, the resampling identifies a single influential donor, Luxembourg for Finland and New Zealand for the UK, while no other donor has a comparable influence. Both donors are structurally unsuitable comparisons: Luxembourg's recorded transport emissions are inflated by fuel tourism, and New Zealand's trajectory reflects idiosyncratic domestic developments. Excluding the influential donor substantially attenuates the estimated effect, which

⁴Ferman et al. (2020) show that synthetic control estimates can be sensitive to the choice of predictors and recommend a specification using all pre-treatment outcome lags, which serves as a baseline benchmark free of researcher discretion.

becomes indistinguishable from the placebo distribution.

Second, I replicate selected results from Runst and Thonipara (2020), who evaluate the impact of the Swedish CO₂ tax on residential CO₂ emissions. Here the mechanism is more direct. Spain receives between 49% and 87% of the weight, depending on the specification, because for most of the pre-treatment period it is the only donor whose emissions lie below Sweden's: its inclusion is necessary to pull the synthetic control down to Sweden's level. Excluding Spain shows that no convex combination of the remaining donors can reproduce Sweden's pre-treatment trajectory. Spain is nevertheless a poor comparison unit. Its residential emissions rise steadily over the matching window while Sweden's fall, and the two series cross before treatment, so the pre-treatment fit is a match in levels between units moving in opposite directions. Spain's upward trend continues after 2001, pulling the counterfactual upward and inflating the estimated treatment effect. In all three cases, the pre-treatment fit, the prerequisite for applying the SCM, is improved by exactly the donor that undermines the counterfactual.

Finally, I turn to the proposal of Abadie and Vives-i-Bastida (2025) to trim the donor pool based on Euclidean distance to the treated unit in order to reduce interpolation bias. I evaluate this proposal by adapting their simulation design to the small donor pools typical of country-level policy evaluation, and show that its applicability in such settings is limited. Moreover, Euclidean distance does not reliably identify the influential donor: in none of the three empirical examples is the influential donor the most distant, and in Runst and Thonipara (2020), Spain is the closest donor of all.

This paper makes three contributions. First, it contributes to the methodological literature on donor pool selection in the SCM. Existing guidance is largely qualitative, and I show that it leaves room for widely divergent choices in practice. More importantly, I show that a good pre-treatment fit can mask interpolation bias, when fit is improved by the very donors that undermine the counterfactual, and that Euclidean distance does not identify these donors. The benefits of trimming on Euclidean distance also do not carry over to the small donor pools typical of country-level policy evaluation. I then show how influential donors can be identified by re-estimating the synthetic control on many subsets of the donor pool. This extends the standard leave-one-out test, which both Mideksa (2024) and Bretschger and Grieg (2024) report. The leave-one-out test drops each donor once while holding the rest of the pool fixed, yielding a single re-estimate that is easily dismissed as one scenario among several. Resampling varies the remaining pool, producing distributions of pre-treatment fit and estimated effects both with and without each donor. This reveals each donor's influence on both, regardless of which other donors are included. Once an influential donor is identified, the researcher can scrutinize it and assess whether it is a suitable comparison for the treated unit. In this way, the procedure extends the transparency of the synthetic control from the weights each donor receives to the influence each donor has on the estimated effect. The procedure

is straightforward to apply in any SCM study with a moderately sized donor pool.

Second, I contribute to the empirical literature on the effectiveness of CO₂ pricing. Ex-post evidence on single-jurisdiction policies such as national CO₂ taxes rests largely on the SCM, so the robustness of SCM estimates is central to what this literature can claim. A recent meta-analysis concludes that most carbon pricing schemes have produced substantial emission reductions (Döbbeling-Hildebrandt et al., 2024), but the strength of the underlying evidence has since been questioned (Piseddu et al., 2026). My findings point to a distinct source of fragility: interpolation bias arising from the composition of the donor pool. In three published, policy-relevant evaluations, the estimated effect rests on a single structurally unsuitable donor. To my knowledge, these are among the first documented cases in the CO₂ pricing literature in which donor pool composition overturns the conclusions of published SCM applications. The findings underline the need for practitioners and referees to scrutinize the robustness of SCM results to donor pool composition.

Finally, the paper contributes to the literature on replication in economics. CO₂ pricing remains the subject of active policy debate,⁵ and ex-post evaluations provide direct evidence on whether it reduces emissions. That evidence carries weight only if it is robust. All three original studies are computationally reproducible with the authors' data, yet their estimates do not survive a defensible change to the donor pool. These findings exemplify what Stevenson and Fischman (2026) refer to as "humans in the loop": estimates depend not only on the methodological approach but also on the specific decisions made by researchers during implementation. In the SCM, the composition of the initial donor pool is among the most consequential of these choices, yet it is often overlooked. Replication that revisits such choices is therefore important, especially where published estimates inform policy.

The remainder of the paper is structured as follows. Section 2 introduces the SCM, its inference procedure, and the leave-one-out test. Section 3 reviews the literature's guidance on donor pool selection and compares it with empirical practice in CO₂ tax evaluations. Sections 4 and 5 replicate Mideksa (2024) and Bretschger and Grieg (2024), use the resampling procedure to identify the influential donor in each study, and show how the results change when that donor is excluded. Section 6 replicates Runst and Thonipara (2020), where the mechanism of interpolation bias is especially transparent. Section 7 evaluates whether trimming the donor pool by Euclidean distance helps in small donor pool settings. Section 8 discusses why the leave-one-out tests did not detect the influential donors, how influence differs from weight, and the implications for replication and policy. Section 9 concludes.

⁵For example, Australia introduced a CO₂ pricing mechanism in 2012 and repealed it in 2014, after its removal had been a central platform of the winning coalition in the 2013 federal election (Commonwealth of Australia, 2014). In France, the government suspended the scheduled 2019 increase in the CO₂ tax in response to the yellow vest protests (Douenne & Fabre, 2022).

2 The synthetic control method

2.1 Estimation

This section provides an overview of the mechanics of the SCM. Consider a panel of $j = 1, \dots, J+1$ aggregate units observed over time periods $t = 1, \dots, T$, with outcome Y_{jt} . Unit $j = 1$ is exposed to a policy intervention after period T_0 , where $T_0 < T$. The remaining units $j = 2, \dots, J+1$ are never exposed and form the donor pool. The goal is to estimate the effect of the intervention on the treated unit during the post-treatment periods $t = T_0 + 1, \dots, T$.

Let Y_{jt}^I and Y_{jt}^N denote the potential outcomes for unit j at time t under treatment and in the absence of treatment, respectively. For the treated unit, the treatment effect is

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N, \quad t = T_0 + 1, \dots, T. \quad (1)$$

Because $Y_{1t}^I = Y_{1t}$ is observed after treatment, estimating τ_{1t} reduces to estimating the counterfactual Y_{1t}^N , which the SCM constructs as a weighted average of donor units:

$$\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} W_j^* Y_{jt}, \quad \text{where } W_j \geq 0 \quad \text{and} \quad \sum_{j=2}^{J+1} W_j = 1. \quad (2)$$

These constraints imply that the synthetic control is a convex combination of donor units. The estimated treatment effect is then

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} W_j^* Y_{jt}, \quad t = T_0 + 1, \dots, T. \quad (3)$$

The choice of weights is therefore consequential. For each unit j we observe k predictors of the outcome, $X_{1j}, \dots, X_{kj}$, which may include pre-intervention values of Y_{jt} . Let X_j denote the $(k \times 1)$ vector of predictors for unit j , and let the $(k \times J)$ matrix $X_0 = \begin{bmatrix} X_2 & \cdots & X_{J+1} \end{bmatrix}$ collect the predictors of the donor units. Given a set of nonnegative constants $V = (v_1, \dots, v_k)$, the weights W^* are chosen to minimize

$$\|X_1 - X_0 W\| = \left(\sum_{h=1}^k v_h \left(X_{h1} - \sum_{j=2}^{J+1} W_j X_{hj} \right)^2 \right)^{1/2} \quad (4)$$

subject to the constraints above, so that the synthetic control resembles the treated unit as closely as possible on the pre-intervention values of the predictors. The constants $v_1, \dots, v_k$ reflect the relative importance of each predictor for Y_{1t}^N , and are typically chosen by a data-driven procedure: V is selected such that the resulting $W^*(V)$ minimizes the mean squared prediction error (MSPE) of the outcome over the pre-intervention periods. For a more comprehensive account of the method, see Abadie (2021).

2.2 Inference

Abadie, Diamond, and Hainmueller (2010) propose a method of inference for the SCM based on permutation methods. Because the method is typically applied to a single treated unit, conventional large-sample inference is unavailable. Instead, the intervention is reassigned in turn to each unit in the donor pool, and a placebo effect is estimated for each such reassignment. This yields a permutation distribution of estimated effects under the null hypothesis of no intervention effect, against which the effect estimated for the treated unit can be compared. The estimated effect is judged significant when its magnitude is extreme relative to this distribution.

Comparing effects directly across units is complicated by the fact that the pre-intervention fit on the outcome variable may be of different quality across units. A placebo unit that is poorly fitted by its own synthetic control will tend to produce a large post-intervention gap regardless of whether any intervention occurred. Abadie, Diamond, and Hainmueller (2010, 2015) suggest using the ratio between the post-treatment and pre-treatment root mean squared prediction error (RMSPE) as the test statistic. Let:

$$R_j(t_1, t_2) = \left(\frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} (Y_{jt} - \hat{Y}_{jt}^N)^2 \right)^{1/2}, \quad (5)$$

where $\hat{Y}_{jt}^N$ is the outcome produced at period t by a synthetic control constructed for unit j , treating j as though it were the treated unit and using the remaining J units as its donor pool. Inference is then based on the permutation distribution of the ratio

$$r_j = \frac{R_j(T_0 + 1, T)}{R_j(1, T_0)}, \quad (6)$$

which measures the post-intervention gap relative to the fit achieved before the intervention.

The p-value from this procedure is simply the proportion of units in the sample whose fit ratio is at least as large as that of the treated unit (Abadie, 2021). One feature of this procedure warrants attention in small samples. Because the treated unit is counted in its own permutation distribution, the smallest p-value attainable is one divided by the number of units in the sample. Conventional significance levels may therefore be out of reach. For a setting with ten donor units, even the most extreme possible result yields a p-value above 0.09⁶. Therefore, reporting the permutation distribution is often more informative than the p-value alone, since it shows how far the estimated effect lies from the placebo effects.⁷

⁶ $\frac{1}{J+1} = 1/11 = 0.091$

⁷ More recent work develops confidence intervals for synthetic control estimates, including approaches based on test inversion (Firpo & Possebom, 2018) and conformal inference (Chernozhukov et al., 2021). I restrict attention to the canonical permutation-based procedure here.

2.3 The leave-one-out robustness test

Abadie (2021) recommends the leave-one-out test to check whether the estimated effect is driven by a single donor unit. The test involves re-estimating the synthetic control repeatedly, each time omitting one of the donors that received a positive weight and allowing the optimization to select new weights from the units that remain. The effect is judged robust if it does not depend on any single donor. Where it does, this should prompt further scrutiny of whether that donor is a necessary and appropriate comparison unit. Abadie (2021) further notes that the sparsity and transparency of the synthetic control estimator are what make such scrutiny possible. While the guidance on how to construct the donor pool remains subjective, this is the diagnostic the literature offers for assessing the consequences of that choice after the fact.

3 The donor pool and interpolation bias

3.1 How to select the donor pool?

The synthetic control for the treated unit is a weighted average of the donor pool. The weights are nonnegative, sum to one, and are typically sparse. The optimization can only draw on the units it is given. The composition of the initial donor pool is therefore consequential. The convexity constraint rules out extrapolation but not interpolation. Interpolation bias arises when units that are dissimilar to the treated unit in predictor space nonetheless combine to produce a close pre-treatment fit. The consequences appear after treatment, when those dissimilarities drive the counterfactual. This section summarizes the guidance the literature offers on donor pool selection. To avoid interpolation bias, the literature consistently recommends restricting the donor pool to units similar to the treated unit, and limiting the size of the donor pool to those units.

Abadie, Diamond, and Hainmueller (2010, p. 495) state:

Researchers trying to minimize biases caused by interpolating across regions with very different characteristics may restrict the donor pool to regions with similar characteristics to the region exposed to the event or intervention of interest.

The same recommendation appears in Abadie, Diamond, and Hainmueller (2015, p. 500):

Finally, to avoid interpolation biases, it is important to restrict the donor pool to units with characteristics similar to the treated unit.

And again, in Abadie (2021, p. 409):

Moreover, it is important to restrict the donor pool to units with characteristics that are similar to the affected unit. The reason is that, while the restrictions placed on the weights, W , do not allow extrapolation, interpolation biases may still be important if the synthetic control matches the characteristics of the affected unit by averaging away large discrepancies between the characteristics of the affected unit and the characteristics of the units in the synthetic control.

Finally, Abadie and Vives-i-Bastida (2025, p. 7) list seven guiding principles for empirical practice when using the SCM. Principle 3, related to interpolation, states:

A larger donor pool is not necessarily better than a smaller one. Adopting a small donor pool of untreated units that are close to the treated unit in the space of the predictors helps reduce over-fitting and interpolation biases.

3.2 How the guidance is applied in empirical research

The guidance has remained remarkably stable and consistent for over a decade. How it translates into empirical practice has not. Table 1 summarizes a selection of published studies that use the SCM to evaluate CO₂ taxes. The donor pools reported in these studies range from 9 to 45 units. Some of this variation reflects genuine differences in setting, outcome variable, and data availability. But the variation persists within jurisdictions. Consider Pretis (2022) and Xiang and Lawley (2019), both of which evaluate the CO₂ tax introduced in British Columbia in 2008. The two studies arrive at very different donor pools: Pretis (2022) includes only other Canadian provinces, yielding 9 units, whereas Xiang and Lawley (2019) also include the states of the United States of America (USA), yielding 45. The last column of Table 1 records whether each study reports a leave-one-out test. Four of the nine do not, highlighting the inconsistency with which this test is applied.

A comparison of the donor pool of Andersson (2019) and Mideksa (2024) further highlights the discretion involved in constructing the donor pool. Andersson (2019) and Mideksa (2024) both evaluate CO₂ taxes on transport emissions in Sweden and Finland, respectively. The empirical setup in Mideksa (2024) closely follows Andersson (2019), a pioneering application of the SCM to CO₂ pricing evaluation. Mideksa adopts the same outcome variable, the same set of predictors, and a similar sample period, in effect applying Andersson’s Swedish design almost directly to the Finnish case. Table 2 reports the composition of the two donor pools: Andersson uses 14 units, Mideksa 20. The two studies evaluate near-identical policies, in the same sector, in neighboring Nordic economies, over a similar period, but diverge on their decision on which countries are similar enough to enter their respective donor pools.

Taken together, these examples suggest that donor pool selection in practice is not governed by any shared standard. Studies evaluating the same policy, and studies applying near-identical designs to comparable settings, arrive at donor pools that differ in both size and composition. In the following sections, I replicate three studies, Mideksa (2024), Bretschger and Grieg (2024), and Runst and Thonipara (2020), and show how discretion over the donor pool can drive the estimated effect. In all three replications, I use the analytical data provided by the authors.⁸

⁸I refer to this exercise as a replication rather than a reproduction because, although I use the analytical data provided by the authors, I do not use their code.

Table 1: Published articles using SCM to evaluate CO₂ pricing instruments

Authors	Journal	Jurisdiction	Instrument	Outcome variable	Donors	Year	LOO
Andersson (2019)	AEJ:EP	Sweden	CO ₂ tax (+ VAT on fuels)	Transport CO ₂ p.c.	14	1991	Yes
Xiang and Lawley (2019)	EE	BC	CO ₂ tax	Residential natural gas consumption p.c.	45	2008	No
Runst and Thonipara (2020)	EE	Sweden	CO ₂ tax increase	Residential CO ₂ p.c.	17	2001	No
Leroutier (2022)	JEEM	UK	CO ₂ price floor (top-up on EU ETS)	Power sector CO ₂ p.c.	15	2013	Yes
Runst and Höhle (2022)	EP	Germany	Energy tax (Ökosteuern)	Transport CO ₂ p.c.	23	1999	No
Pretis (2022)	ERE	BC	CO ₂ tax	Sectoral CO ₂ p.c.	9	2008	No
Bretschger and Grieg (2024)	EP	UK	Fuel duty escalator	Transport CO ₂ p.c.	13	1993	Yes
Mideksa (2024)	JEEM	Finland	CO ₂ tax	Transport CO ₂ p.c.	20	1990	Yes
Aus dem Moore et al. (2025)	JPUB	Sweden	CO ₂ tax (+ VAT on fuels)	Clean transport innovation patents	16	1991	Yes

Notes: The table lists published studies that use the SCM with a single treated unit to evaluate instruments that raise the price of CO₂-emitting fossil fuels. Most are explicit carbon taxes; the others (an energy tax, a fuel duty escalator, and a carbon price floor) price CO₂ implicitly or as a top-up to an existing scheme, as indicated in *Instrument*. *Donors* is the number of units from which the synthetic control is constructed; *Year* is the treatment year in the SCM; *LOO* indicates whether the study reports a leave-one-out test. p.c. = per capita; BC = British Columbia; EU ETS = EU Emissions Trading System. Journal abbreviations: AEJ:EP (American Economic Journal: Economic Policy), EE (Energy Economics), EP (Energy Policy), JEEM (Journal of Environmental Economics and Management), ERE (Environmental and Resource Economics), JPUB (Journal of Public Economics). The list aims to cover all published applications of this kind, though it may not be exhaustive.

Table 2: Comparison of donor pool composition across two empirical SCM applications

Andersson (2019)	Mideksa (2024)
1. Australia	1. Australia
2. Belgium	2. Austria
3. Canada	3. Belgium
4. Denmark	4. Canada
5. France	5. France
6. Greece	6. Germany
7. Iceland	7. Greece
8. Japan	8. Iceland
9. New Zealand	9. Ireland
10. Poland	10. Italy
11. Portugal	11. Japan
12. Spain	12. Luxembourg
13. Switzerland	13. Netherlands
14. USA	14. New Zealand
	15. Portugal
	16. Spain
	17. Switzerland
	18. Turkey
	19. UK
	20. USA

Notes: The table reports the donor pools used by Andersson (2019) and Mideksa (2024) to estimate the effect of introducing CO₂ taxes on transport CO₂ emissions in Sweden and Finland, respectively. Countries shaded in gray appear in both donor pools; unshaded countries are unique to each study.

4 Replication of Mideksa (2024, JEEM)

In 1990, Finland became the first country in the world to introduce a CO₂ tax. Mideksa (2024) leverages this quasi-experimental setting to evaluate the impact of introducing the tax on transport CO₂ emissions using the SCM. To construct synthetic Finland, the study utilizes a donor pool of 20 OECD countries: Australia, Austria, Belgium, Canada, France, Germany, Greece, Iceland, Ireland, Italy, Japan, Luxembourg, the Netherlands, New Zealand, Portugal, Spain, Switzerland, Turkey, the UK, and the USA. The data, which span the period from 1970 to 2005, are divided into a 20-year pre-treatment window (1970–1989) and a 16-year post-treatment window (1990–2005). Regarding the donor pool composition, Mideksa (2024) argues that relying on an objective criterion, specifically, OECD membership, “prevents hidden bias from contaminating the choice,” as opposed to a handpicked selection of countries. The donor pool weights are determined by minimizing the pre-treatment discrepancy between Finland and a weighted average of the donor units across a set of key predictors. The synthetic control is therefore constructed by matching Finland on GDP per capita, gasoline consumption per capita, motor vehicles per 1,000 persons, the percentage of urban population, and pre-treatment outcome values for 1975, 1980, 1985, and 1989. This specification

uses the same set of predictors as Andersson (2019).⁹ The study reports substantial treatment effects: relative to the counterfactual, Finnish emissions are estimated to be 16% lower in 1995, 25% lower in 2000, and 30% lower in 2004. Notably, these effects are among the largest reported in the literature for the transport sector.

4.1 Replication of main findings

Figure 1 presents the replication of Mideksa’s main results using the original predictor specification and the analytical data provided by the author. The replication is carried out using the `synth2` package in Stata.¹⁰ Figure 1 (a) shows the per-capita transport emission trajectory of Finland alongside its synthetic counterpart. Following the introduction of the CO₂ tax in 1990, a clear divergence emerges between actual Finnish emissions and the estimated counterfactual trajectory, i.e., the emissions path Finland would have followed in the absence of the policy. Figure 1 (b) plots the gap between the synthetic control and actual Finnish emissions, providing a direct visual representation of the estimated treatment effect over time.

Figure 1: Replication results

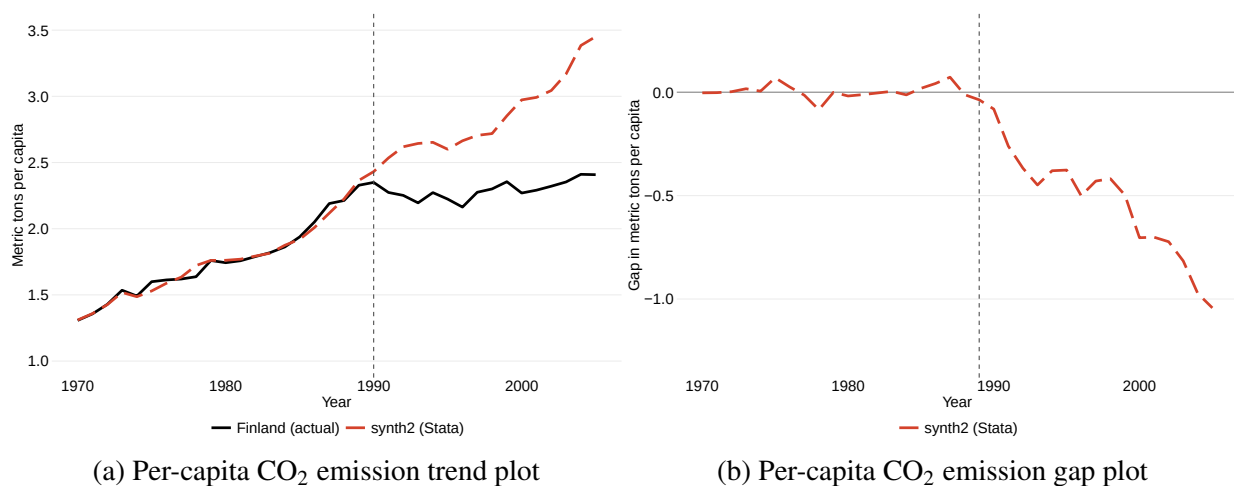


Table 3 compares the donor weights used to construct the synthetic control in the replication with those reported in the original study (shaded in gray). The replicated weights resemble the original ones with the same countries receiving weights of similar magnitude.¹¹

⁹Mideksa (2024) and Andersson (2019) use the same set of predictors; however, the pre-treatment outcome lags differ. Andersson uses outcome lags for the years 1970, 1980 and 1989.

¹⁰I also replicate the results using the software packages `synth` in Stata and `Synth` in R. The cross-software comparison serves to establish that the replication results hold across implementations. The results of this exercise are presented in Appendix A.

¹¹In addition to the donor weights, I replicate the predictor balance table and find some discrepancies in the reported results. This issue is separate from the main argument of this paper. Nonetheless, I detail the discrepancies between my results and those reported by Mideksa (2024) in Appendix B.

Table 3: Donor countries and synthetic control weights

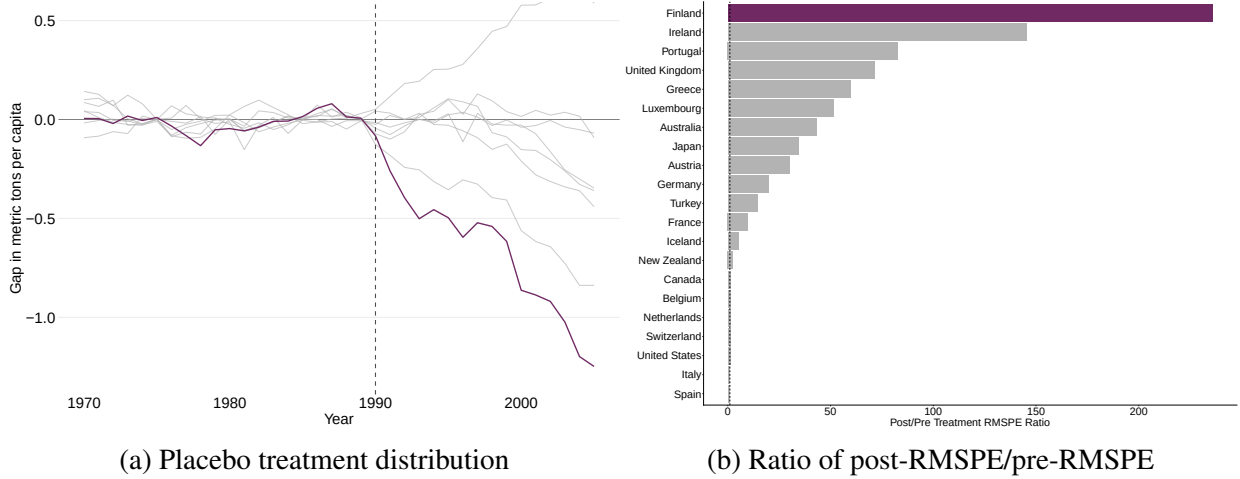
Donor Country	(Mideksa, 2024)	synth2 (Stata)
Australia	0	0
Austria	0	0
Belgium	0	0
Canada	0	0
Switzerland	0.094	0.095
Germany	0	0
Spain	0	0
France	0	0
UK	0.432	0.44
Greece	0	0
Ireland	0	0
Iceland	0	0
Italy	0	0
Japan	0	0
Luxembourg	0.102	0.099
Netherlands	0	0
New Zealand	0.159	0.164
Portugal	0	0
Turkey	0.184	0.175
USA	0.029	0.026
$\Sigma \approx$	1	1

Figure 2 presents the results of the in-space placebo (permutation) test used to assess the statistical significance of the estimated effect, as detailed in Section 2.2. Figure 2(a) shows the gap trajectories for all placebo units with a pre-treatment fit comparable to that of Finland, defined, following Mideksa (2024), as those whose pre-treatment RMSPE does not exceed Finland's by more than 25%. Figure 2(b) plots the ratio of post-treatment to pre-treatment RMSPE for Finland and all placebo units. Finland's post-treatment gap is the most negative among the placebo gaps in panel (a), and its RMSPE ratio is the largest in the sample in panel (b), implying a permutation p -value of $1/21 \approx 0.048$.¹² Both results are consistent with Mideksa (2024).¹³

¹²Mideksa (2024) reports a p -value of $1/20 = 0.05$. However, the permutation p -value should include the treated unit in the denominator (Abadie, Diamond, & Hainmueller, 2015); with 20 donor countries and Finland, it is $1/21 \approx 0.048$.

¹³The corresponding results estimated using `synth` in Stata and `Synth` in R are presented in Figures C1 and C2 in Appendix C, and yield qualitatively similar results.

Figure 2: Placebo distributions



4.2 Sensitivity to donor pool composition

Next, I examine how the estimated treatment effect in Mideksa (2024) varies with the composition of the donor pool. I draw 300 unique random subsamples of 14 donors from the pool of 20 countries.¹⁴ This donor pool size was chosen to maintain comparability with Andersson (2019), which used 14 donor countries. For each subsample, I append the treated unit, Finland, and re-estimate the synthetic control. Because each donor country is included in a given subsample with probability $14/20 = 0.7$, it appears in expectation in 210 and is absent from 90 of the 300 subsamples. This allows me to assess how sensitive the estimated treatment effect is to the inclusion of each donor unit that received a positive weight in the original specification. This resampling exercise was carried out in R using the `Synth` package.¹⁵

I conduct this resampling exercise for each of the six donors that receive a positive weight in Mideksa (2024): the United Kingdom, New Zealand, Turkey, Luxembourg, Switzerland, and the United States. I additionally include Italy, which receives a positive weight when the synthetic control is estimated with `Synth` in R, the package used for the resampling exercise; see Table A1 in Appendix A. For each donor, I compare the subsets that include it with those that exclude it, in terms of both pre-treatment RMSPE and the average post-treatment gap. For most donors, inclusion makes little difference. This holds even for the UK, which receives the largest weight in the original study (43.2%): whether or not the UK is included, the pre-treatment RMSPE distributions overlap substantially and the average post-treatment gaps are qualitatively indistinguishable (Figure 3). Luxembourg is the only exception. Despite receiving a weight of only 10.2%, its inclusion sharply separates the subsets (Figure 4). The pre-treatment RMSPE distribution of subsets

¹⁴There are $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \binom{20}{14} = \frac{20!}{14!6!} = 38\,760$ unique combinations, of which I choose a random sample of 300.

¹⁵The resampling exercise is carried out in R using the `Synth` package, which is better suited to repeated estimation. The main analysis uses the `synth2` command in Stata. Appendix A reports results from `synth` and `synth2` in Stata and `Synth` in R, which are qualitatively similar, so the choice of software does not affect the conclusions drawn from the resampling.

containing Luxembourg lies to the left of, and closer to zero than, the distribution of subsets excluding it, with virtually no overlap between the two. Including Luxembourg thus consistently improves the pre-treatment fit. The same bifurcation is evident in the estimated treatment effects: subsets including Luxembourg systematically yield a large post-treatment divergence, while those excluding it do not. Similar to the UK, there is no clear difference between subsets with and without the remaining donors in terms of either the pre-treatment RMSPE or the estimated treatment effect. These results are detailed in [Appendix E](#). [Appendix F](#) and [Appendix G](#) illustrate four draws from subsets with and without Luxembourg, respectively, making the role of Luxembourg's inclusion and exclusion clearly visible in individual draws.

Figure 3: Samples with and without the UK

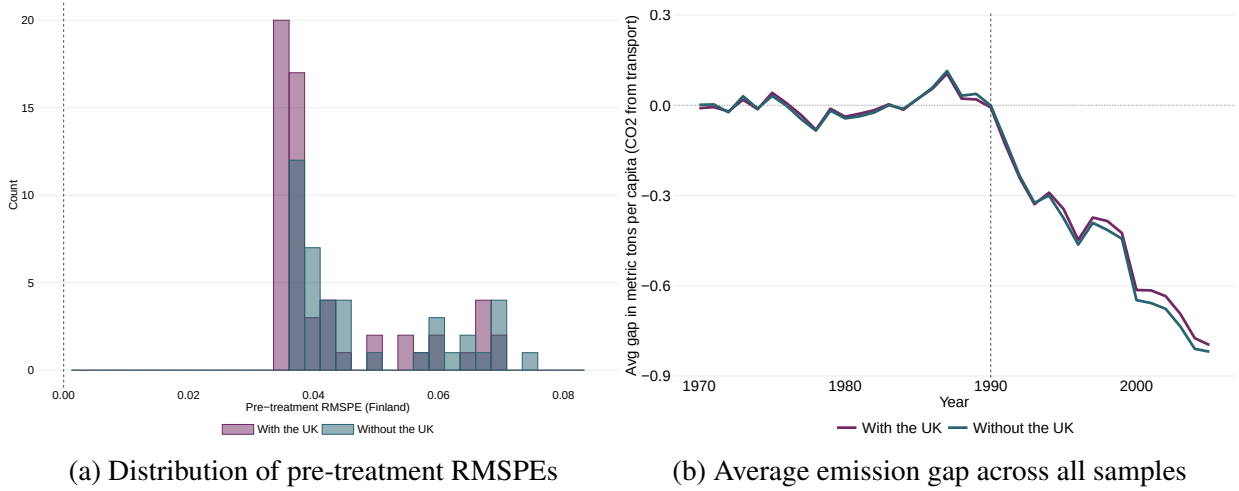
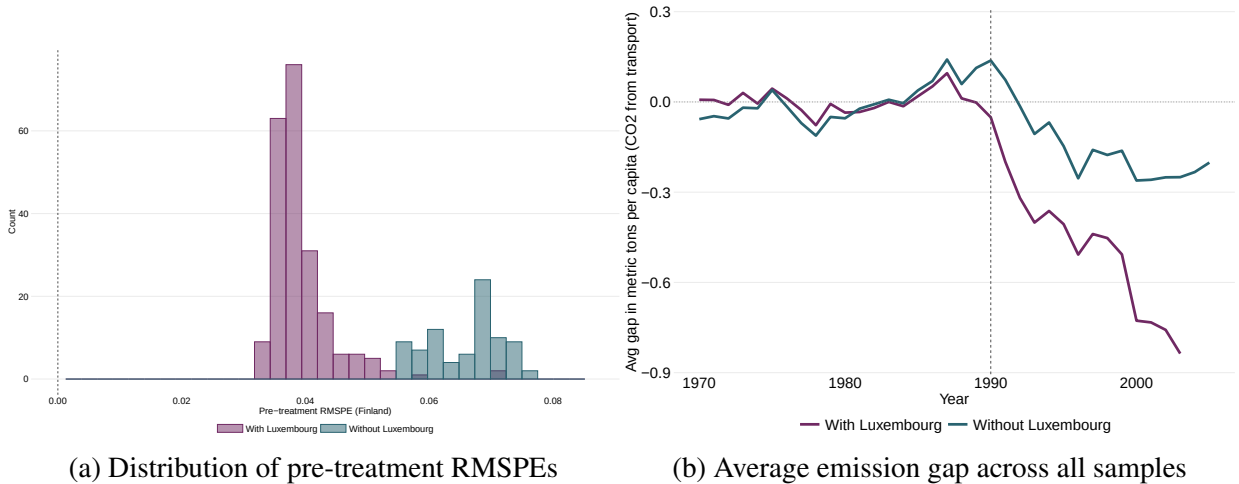


Figure 4: Samples with and without Luxembourg

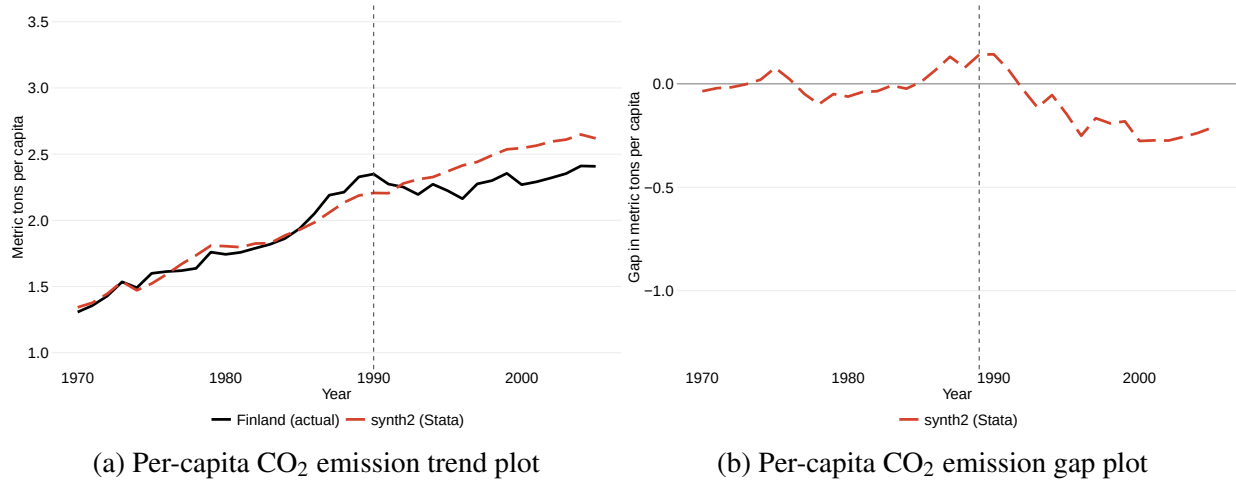


4.3 Replication excluding the influential donor

Having identified Luxembourg as the influential donor, I re-estimate the synthetic control using the full donor pool excluding Luxembourg. Figure 5 illustrates the results of this exercise. Two changes are immediately clear: the pre-treatment fit deteriorates markedly compared to the

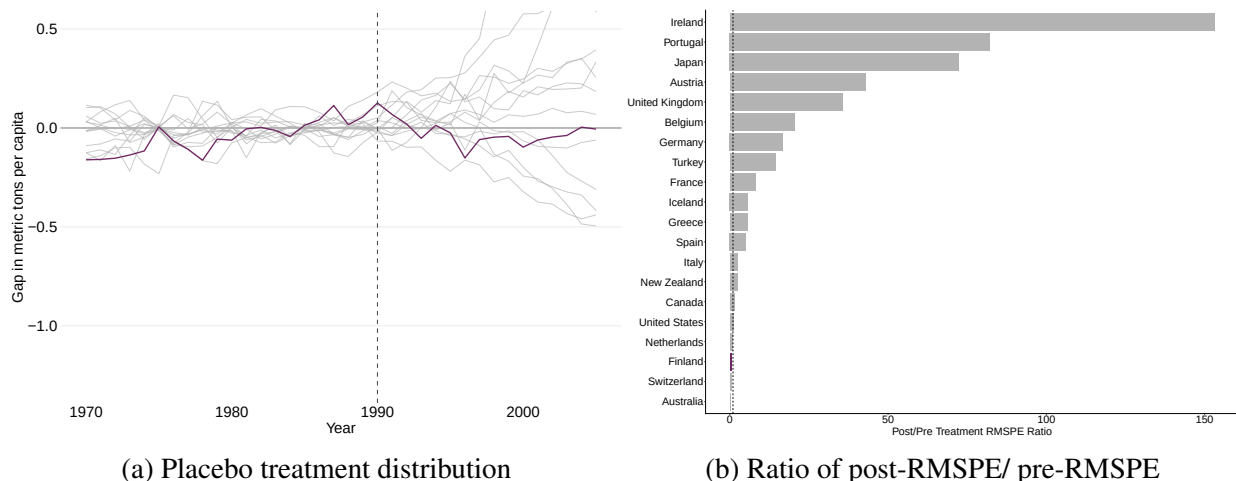
original results (see Figure 1), and the post-treatment divergence, representing the estimated policy effect, is substantially attenuated. The estimated average effect drops from 19% to 6% over the post-treatment period, roughly one-third of the average estimate reported in Mideksa (2024).¹⁶

Figure 5: Replication results excluding Luxembourg



I next carry out the permutation inference procedure using the donor pool excluding Luxembourg. Figure 6 (a) shows that once Luxembourg is excluded, Finland's estimated treatment effect is no longer distinguishable from the placebo distribution. Figure 6 (b) further shows that Finland's post-intervention to pre-intervention RMSPE ratio is no longer the largest in the sample; it falls substantially and ranks 18 out of 20. This indicates clearly that the result reported in the original study does not survive the removal of Luxembourg from the donor pool.¹⁷

Figure 6: Placebo distributions without Luxembourg



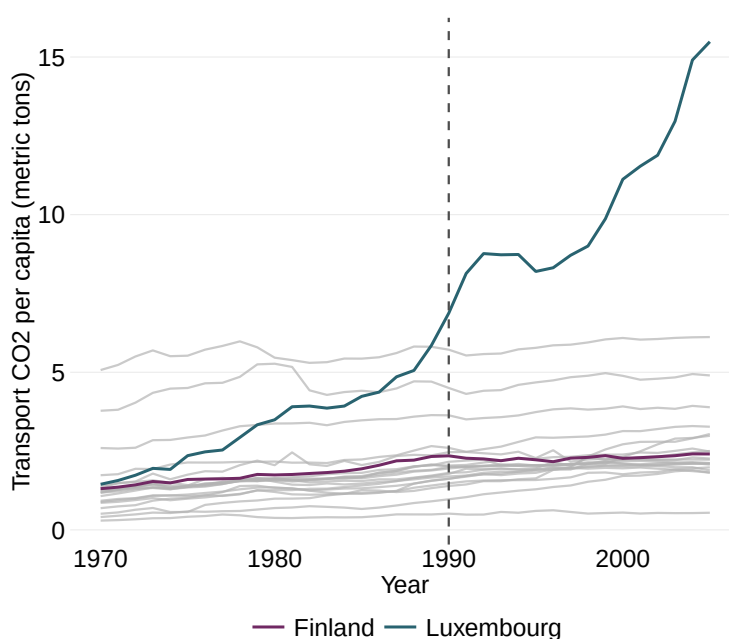
Should Luxembourg have been included in the donor pool? Figure 7 plots the raw outcome trends of Luxembourg against those of Finland and the rest of the donor pool. Luxembourg is a clear

¹⁶Estimated effect is approximately one-third of the original: $-6\% / -19\% \approx 0.32$

¹⁷The corresponding results for `synth` in Stata and `Synth` in R are presented in Figures D1 and D2 in Appendix D, and confirm this finding across all software implementations.

outlier: its emissions rise steadily and continue to increase after 1990, a trajectory not shared by the remaining donors. Excluding Luxembourg is consistent with Andersson (2019), who omits it from the donor pool on the grounds of unusually high non-domestic fuel consumption. Due to its very low fuel taxes, Luxembourg’s per capita fuel consumption is roughly three to four times that of most other European countries (Sterner, 2006). These low prices attract fuel tourism, as drivers from neighboring countries cross the border to refuel, which inflates Luxembourg’s recorded emissions. This matters for the synthetic control because the counterfactual inherits the post-treatment trajectories of its donors: as Luxembourg’s emissions rise after 1990, so do those of synthetic Finland, widening the gap to actual Finland regardless of any effect of the tax. Therefore, it can be reasonably argued that Luxembourg fails the donor selection guidance discussed in Section 3: its recorded emissions are driven by fuel tourism, which is unrelated to the determinants of Finnish emissions. In other words, its outcome is generated by a different structural process than Finland’s.

Figure 7: Transport CO₂ emissions metric tons per capita: Finland and donor pool countries, 1970–2005



5 Replication of Bretschger and Grieg (2024, Energy Policy)

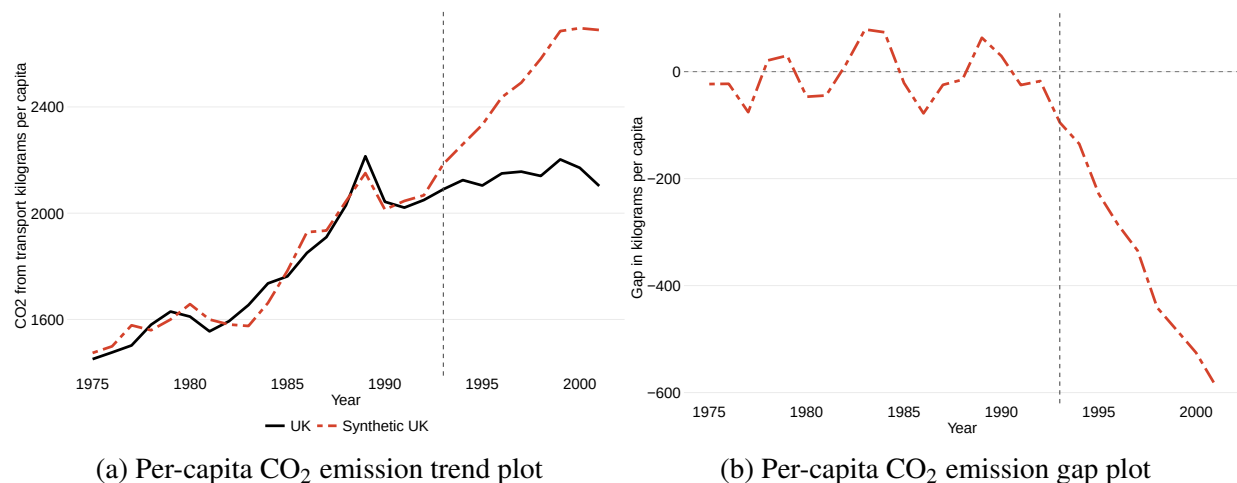
Bretschger and Grieg (2024) evaluate the impact of the UK’s fuel duty escalator, introduced in 1993, which acted as an implicit CO₂ tax on transport sector emissions. As a result of this tax, UK gasoline prices went from among the lowest in Europe to among the highest. The authors evaluate the policy using the SCM. To construct synthetic UK, the study uses a donor pool of 13 OECD countries: Australia, Belgium, Canada, Denmark, France, Greece, Ireland, Italy, Japan, New Zealand, Portugal, Spain, and the United States. OECD membership is used to ensure similarity to the UK. From this set, the authors exclude countries that implemented a CO₂ or comparable fuel tax during the study period (Finland, the Netherlands, Norway, and Sweden), countries whose emissions are distorted for other reasons (Austria and Luxembourg), and countries with insufficient

data. This yields the final pool of 13 donors. The data, which span the period 1975 to 2001, are divided into an 18-year pre-treatment period (1975–1992) and a 9-year post-treatment window (1993–2001). The donor weights are determined by minimizing the pre-treatment discrepancy between the UK and a weighted average of the donor units across a set of key predictors. The predictors in the authors’ main specification are GDP per capita, gasoline consumption per capita, the percentage of urban population, human capital, and pre-treatment outcome values in 1975, 1980, 1985, and 1990. This specification closely follows that of Andersson (2019), except that the authors omit motor vehicles per 1 000 people and add human capital as a predictor.¹⁸ Bretschger and Grieg (2024) estimate that, by 2001, UK per capita emissions are 0.500 metric tons (about 21%) lower than those of synthetic UK, with an average reduction of 0.352 metric tons per capita (about 13%) over the post-treatment period.

5.1 Replication of main findings

Figure 8¹⁹ presents the replication of the main results of Bretschger and Grieg (2024), using the original predictor specification and the data provided by the authors, in order to replicate the original study as closely as possible. I estimate the model using the `synth2` package in Stata.²⁰ The replication reproduces the original results. Figure 8 (a) shows the per capita transport emission trajectory of the UK alongside its synthetic counterpart. Following the introduction of the fuel duty escalator in 1993, a clear divergence emerges between actual UK emissions and the estimated counterfactual, i.e., the emissions path the UK would have followed in the absence of the policy. Figure 8 (b) plots the gap between actual and synthetic UK emissions, which represents the estimated treatment effect over time.

Figure 8: Replication results



¹⁸The original study provides no definition of the human capital variable beyond describing it as an index from the Penn World Table (PWT), and does not report which PWT version is used. I assume it corresponds to the PWT human capital index, which is constructed from average years of schooling and an assumed rate of return to education.

¹⁹The outcome axes in Bretschger and Grieg (2024) are labeled in tons per capita, but the magnitudes (up to 3,000) indicate that emissions are reported in kilograms per capita.

²⁰I do not repeat the cross-software check here, as the results in the Mideksa replication were consistent across software implementations.

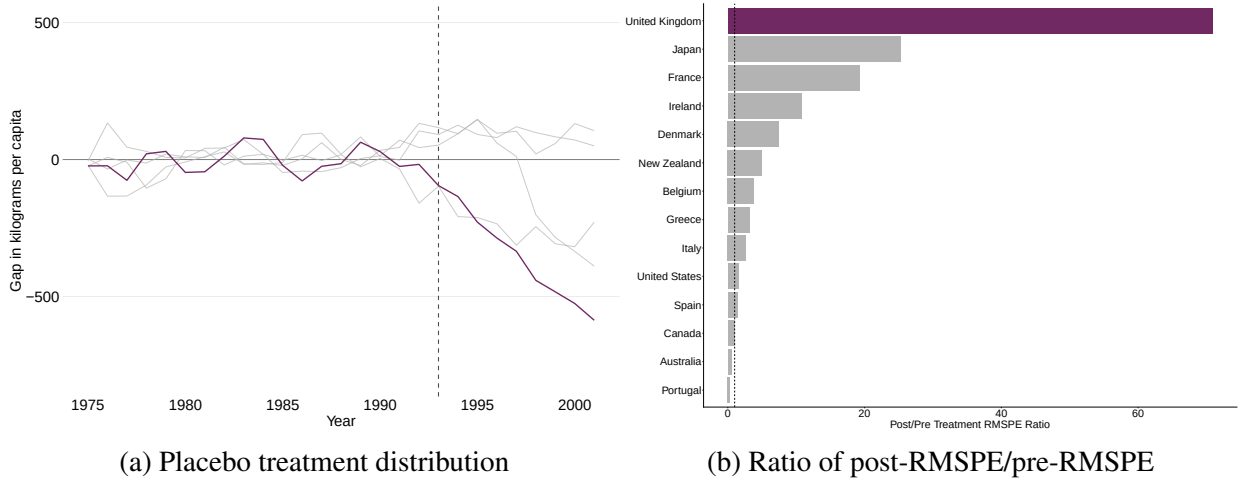
Table 4 compares the donor weights used to construct the synthetic control in the replication with those reported in the original study (shaded in gray). The replicated weights resemble the original ones, with the same countries receiving weights of similar magnitude.

Table 4: Donor countries and synthetic control weights

Donor Country	(Bretschger & Grieg, 2024)	synth2 (Stata)
USA	0	0
Canada	0	0
Ireland	0.197	0.198
Belgium	0	0
France	0	0
Spain	0	0
Portugal	0	0
Italy	0	0
Greece	0.253	0.261
Denmark	0	0
Japan	0.176	0.165
Australia	0	0
New Zealand	0.375	0.375
$\Sigma \approx$	1	1

Figure 9 presents the results of the in-space placebo (permutation) test used to assess the statistical significance of the estimated effect, as detailed in Section 2.2. Figure 9 (a) shows the gap trajectories for all placebo units with a pre-treatment fit comparable to that of the UK, defined as those whose pre-treatment RMSPE does not exceed the UK's by more than 25%. Figure 9 (b) plots the ratio of post-treatment to pre-treatment RMSPE for the UK and all placebo units. The UK's post-treatment gap is the most negative among the placebo gaps in panel (a), and its RMSPE ratio is the largest in the sample in panel (b), implying a permutation p -value of $1/14 \approx 0.07$. This is the lowest attainable p -value with 14 units. Both results are consistent with Bretschger and Grieg (2024).

Figure 9: Placebo distributions



5.2 Sensitivity to donor pool composition

Next, I assess the sensitivity of the results of Bretschger and Grieg (2024) to the composition of the donor pool. I estimate the synthetic control on all 286 subsets of 10 donors from the pool of 13 countries.²¹ Unlike in the Finnish case, the smaller donor pool makes full enumeration feasible, so the results do not depend on random sampling. With subsets of 10, each donor is included in 220 subsets²² and excluded from 66, and each exclusion occurs alongside a different combination of the remaining donors.²³ For each subset, I append the UK and re-estimate the synthetic control. This allows me to assess how sensitive the estimated treatment effect is to the inclusion of each donor unit that received a positive weight in the original specification.

I conduct this resampling exercise for each of the four donors that receive a positive weight in Bretschger and Grieg (2024): Ireland, Greece, Japan, and New Zealand. I additionally include Belgium, Spain, and Portugal, which receive positive weights when the synthetic control is estimated with `Synth` in R, the package used for the resampling exercise. For each donor, I compare the subsets that include it with those that exclude it, in terms of both pre-treatment RMSPE and the average post-treatment gap. As with Luxembourg in the Finnish case detailed in Section 4, New Zealand is the exception (Figure 10). The pre-treatment RMSPE distribution of subsets containing New Zealand lies to the left of, and closer to zero than, the distribution of subsets excluding it, with virtually no overlap between the two. The same bifurcation is evident in the estimated treatment effects: subsets including New Zealand yield a substantially larger post-treatment divergence than those excluding it. For all other donors, inclusion makes little difference. These results are detailed in Appendix H. Appendix I and Appendix J illustrate four

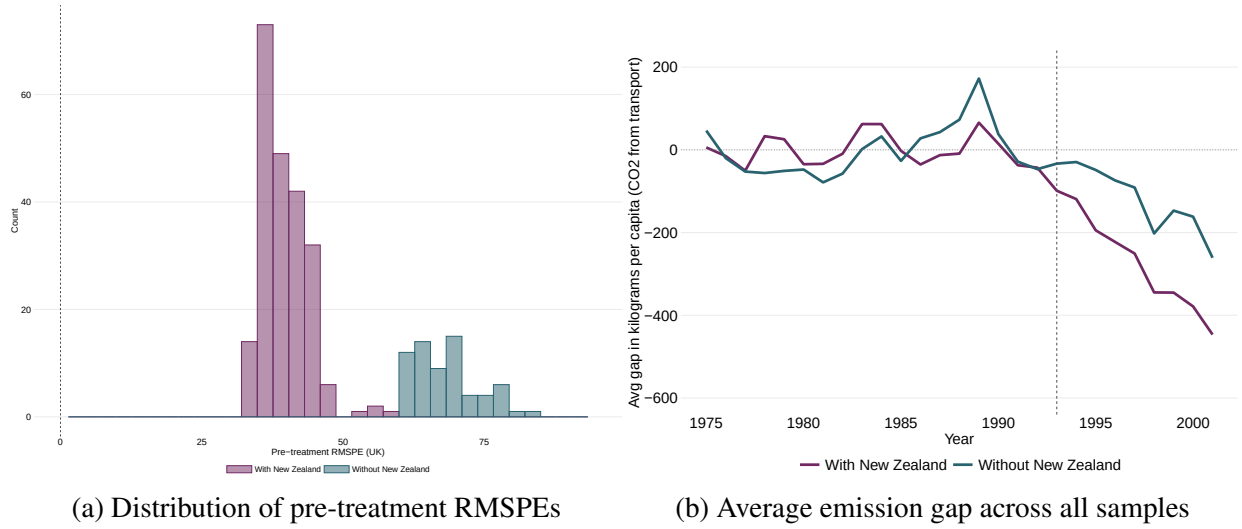
²¹ $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, so $\binom{13}{10} = \frac{13!}{10!3!} = 286$.

²² Of the $\binom{13}{10} = 286$ possible donor pools, each donor appears in a share $10/13$ of them, i.e. $\frac{10}{13} \times 286 = 220$ pools.

²³ Each subset of 10 omits three of the 13 donors, so each of the 66 subsets without a given donor also omits a different pair of the remaining 12. The effect of excluding a donor is then observed across many compositions of the rest of the pool. With subsets of 11, each donor would be omitted alongside only one other, giving 12 such compositions, and with subsets of 12 the exercise reduces to the standard leave-one-out test.

draws from subsets with and without New Zealand, respectively, making the role of New Zealand’s inclusion and exclusion clearly visible in individual draws.

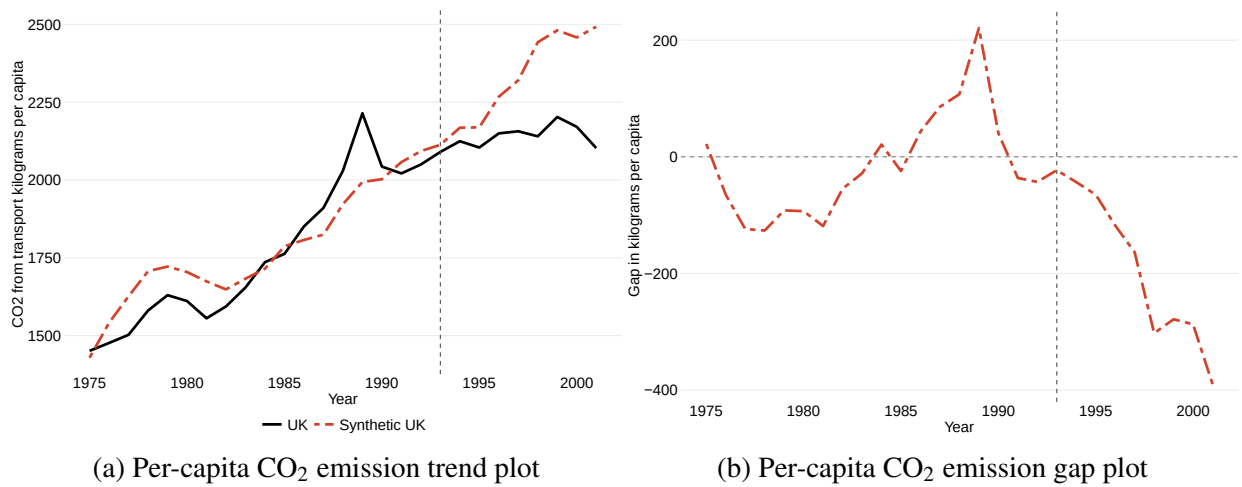
Figure 10: Samples with and without New Zealand



5.3 Replication excluding the influential donor

Next, I re-estimate the synthetic control using the full donor pool excluding New Zealand, the influential donor identified in Section 5.2. Figure 11 illustrates the results of this exercise. Two changes are immediately clear: the pre-treatment fit deteriorates markedly compared to the original results (see Figure 8), and the post-treatment divergence, representing the estimated policy effect, is substantially attenuated. The estimated effect averages approximately -7% over the post-treatment period, about half the average estimate of -13% reported in Bretschger and Grieg (2024).

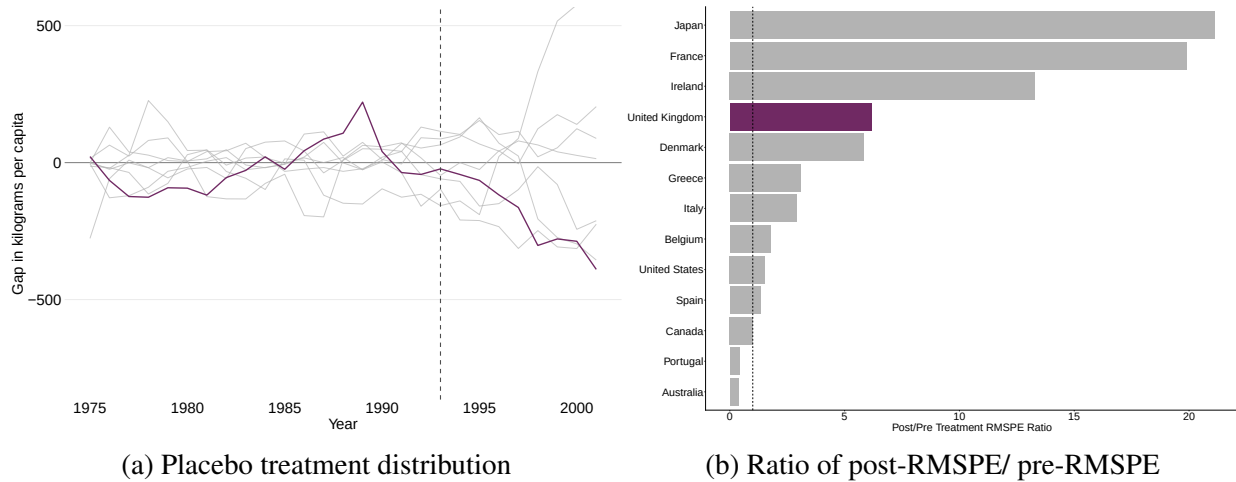
Figure 11: Replication results excluding New Zealand



I next carry out the in-space placebo test using the donor pool excluding New Zealand. Figure 12 (a) shows that, once New Zealand is excluded, the UK’s estimated treatment effect is no longer the largest negative gap in the placebo distribution. Figure 12 (b) further shows that the UK’s ratio

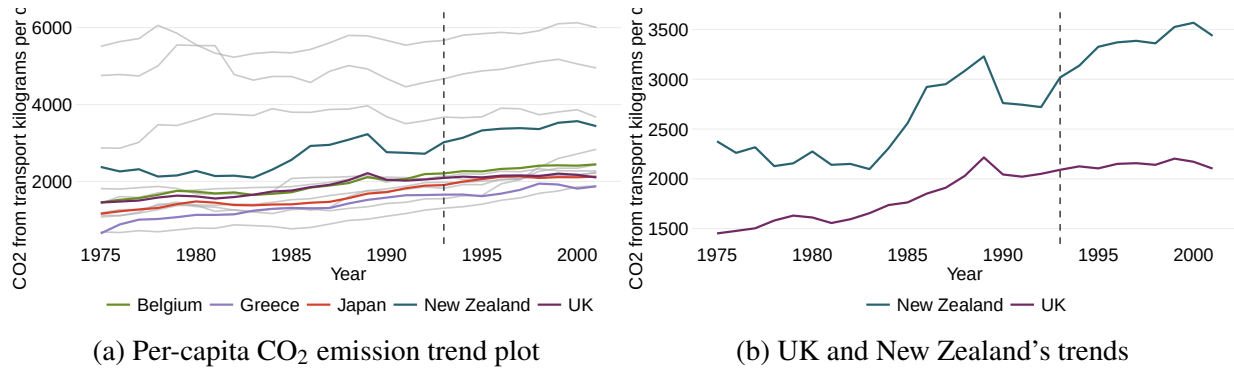
of post-treatment to pre-treatment RMSPE falls well within the range of placebo ratios, ranking 4th of 13 units. Without New Zealand, the estimated effect is therefore indistinguishable from the placebo effects. The results reported in the original study do not survive the removal of New Zealand from the donor pool.

Figure 12: Placebo distributions without New Zealand



Should New Zealand have been included in the donor pool? Figure 13 (a) plots the raw outcome trends of New Zealand against those of the UK and the rest of the donor pool, and panel (b) focuses on the UK and New Zealand. Unlike Luxembourg in the case of Mideksa’s study, New Zealand is not an outlier in levels; what sets it apart is the shape of its trajectory. Its emissions fall until about 1983, rise sharply in the second half of the 1980s, fall back around 1990, and rise steadily after 1993. These phases reflect country-specific developments. The early decline coincides with domestic energy-security measures following the second oil shock, including the “carless days” of 1979–1980 (Johnston et al., 1983). The later rise reflects a surge in imported used vehicles, one of the oldest light vehicle fleets among developed countries, and high car dependence. Between 1990 and 2016, New Zealand’s transport emissions grew by 71.3%, compared with 1.7% in the UK (Hasan et al., 2019). New Zealand therefore fails the donor selection guidance discussed in Section 3: its emissions are shaped by country-specific developments unrelated to the UK, and they continue to rise after 1993, inflating the counterfactual and widening the estimated gap. Its emissions are generated by a different structural process than the UK’s, and it should not have been included in the donor pool.

Figure 13: CO₂ from transport kilograms per capita: UK and donor pool countries, 1975–2001



6 Replication of Runst and Thonipara (2020, Energy Economics)

Sweden introduced a CO₂ tax in 1991 at €26 per tonne of CO₂, raising it steadily thereafter. The sharpest adjustment occurred between 2001 and 2004, when the rate rose from around €40 to €100 per tonne. Runst and Thonipara (2020) exploit this increase to evaluate its effect on residential CO₂ emissions. Using annual data on residential CO₂ emissions per capita for 18 European countries between 1990 and 2016, they treat 2001 as the intervention date and estimate the effect using, among other methods, the SCM. Their SCM design is notable for the number of estimates it produces: three alternative donor pools, or samples, are crossed with five predictor specifications.²⁴ The samples are defined by the prevailing CO₂ tax rate in potential control countries.

6.1 Replication of main findings

I replicate the results for Sample II, which Runst and Thonipara (2020) describe as yielding their “best estimate” of the effect of the CO₂ tax. Sample II restricts the donor pool to countries with no CO₂ tax or a tax below €20 per tonne, and further excludes Italy, Greece, and the Netherlands, which imposed relatively high general energy taxes. This reduces the donor pool from 17 countries in the full sample to 8. This replication is deliberately narrower in scope than that carried out for Mideksa (2024) and Bretschger and Grieg (2024). The objective is to demonstrate how interpolation bias arises in a case where the mechanism is unusually transparent: the donor pool is small, the weights are concentrated in a structurally dissimilar unit, and the consequences of removing that unit are visible in the raw trajectories.

Table 5 reports the donor weights for synthetic Sweden across the five specifications estimated on Sample II. The column labeled *Spec 3 Replication* gives the weights I obtain; they reproduce

²⁴The five specifications differ in the combination of lagged outcomes and covariates used as predictors. Specification 1 uses the lags for 1990, 1994, and 2000; specification 2 uses the three lags immediately preceding treatment (1998, 1999, and 2000); and specification 3 uses all pre-treatment lags. Specification 4 retains the lag structure of specification 1 and adds heating degree days, GDP per capita, the oil price, and the electricity price as covariates. Specification 5 includes a single lagged outcome, for 1995, alongside the same covariates.

the published values exactly. One feature stands out: Spain is the largest contributor in every specification, carrying between 0.488 and 0.865 of the total weight. Figure K1 in [Appendix K](#) shows the replication results for specification 3 in Sample II: the trajectories of Sweden and synthetic Sweden, together with the corresponding gap plot.

Table 5: Donor countries and synthetic control weights

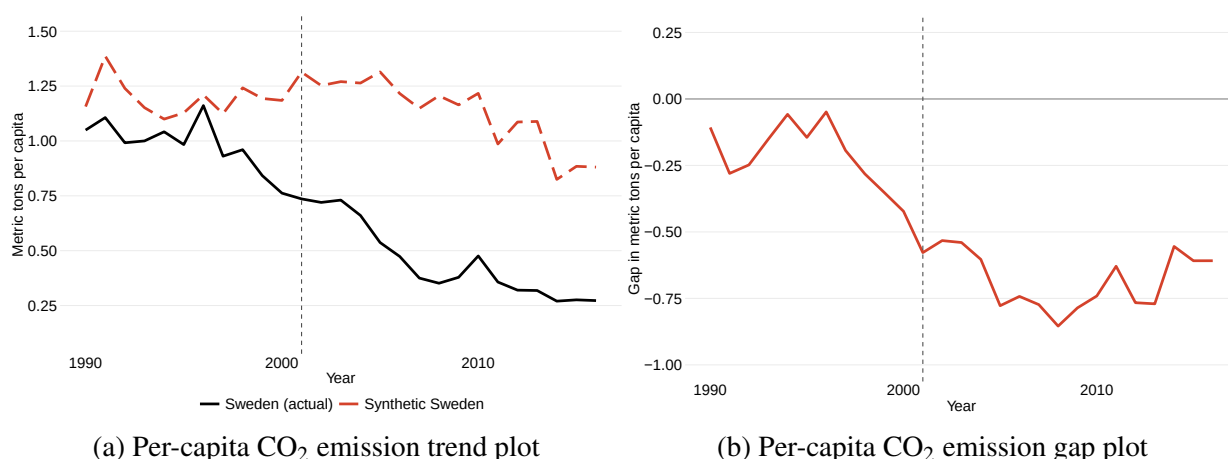
Donor Country	Spec 1	Spec 2	Spec 3	Spec 3 Replication	Spec 4	Spec 5
Austria	0	0	0	0	0	0
Belgium	0	0	0	0	0	0
Czech Republic	0	0	0	0	0	0
France	0	0.289	0.385	0.385	0.492	0.413
Germany	0	0	0	0	0	0
Poland	0.135	0	0.03	0.03	0	0.099
Slovakia	0	0	0	0	0	0
Spain	0.865	0.711	0.585	0.585	0.505	0.488
$\Sigma \approx$	1	1	1	1	1	1

6.2 Sensitivity to donor pool composition

Next, I examine how sensitive the estimated effect is to the inclusion of Spain. The authors do not report the results of a leave-one-out test. Since Spain carries the most weight across all specifications, I remove it from the donor pool and re-estimate. The results are shown in Figure 14. Synthetic Sweden now lies above actual Sweden in every year, and the gap reaches about -0.6 metric tons per capita already by 2000, before the treatment. France receives the full weight, and this is consistent across all specifications. The synthetic control is therefore simply France's emissions trajectory.²⁵

²⁵I do not apply the resampling exercise to Runst and Thonipara (2020). Once Spain is excluded, Sweden lies outside the convex hull of the remaining donors, and the optimization assigns the full weight to France in every specification. The resulting estimate is no longer a synthetic control but a comparison with a single country, so resampling would add nothing: every subset without Spain collapses in the same way.

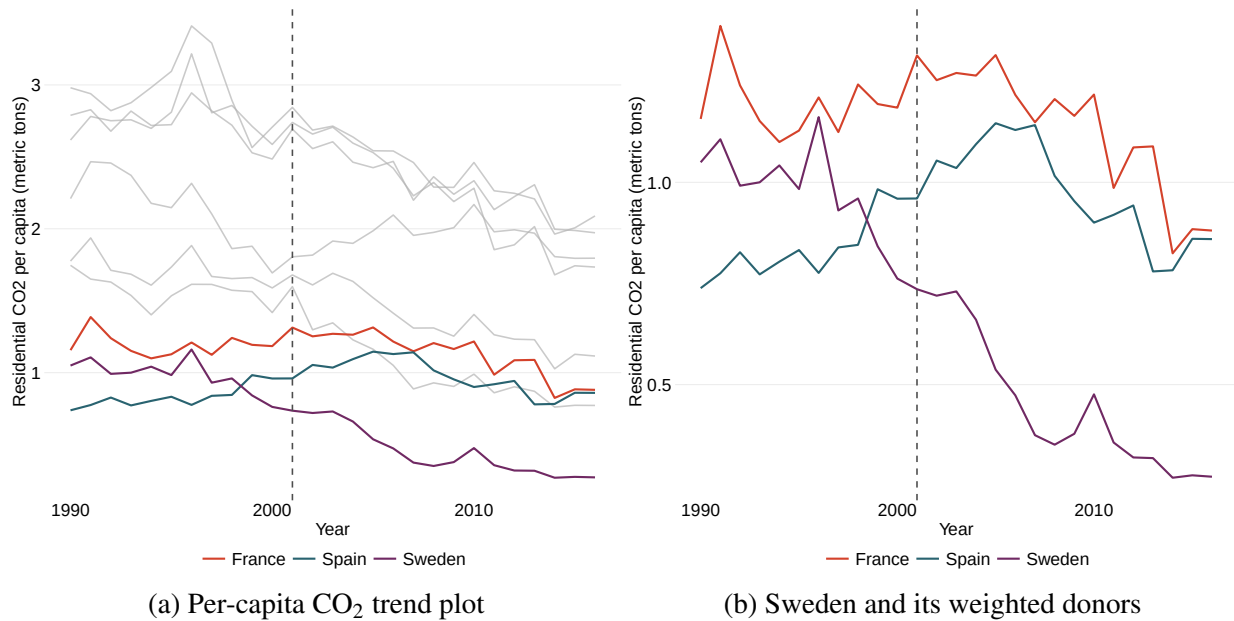
Figure 14: Replication results excluding Spain



This result warrants closer examination. Figure 15 plots the raw emissions series: panel (a) shows all countries in the sample, with Sweden, Spain and France highlighted, and panel (b) restricts attention to those three. The figure makes clear why Spain receives most of the weight in every specification. For most of the pre-treatment period, it is the only country in the donor pool whose emissions lie below Sweden's. Because the synthetic control is a convex combination, its level is bounded below by the lowest donor, so with every other donor above Sweden, no combination of them can reach Sweden's level, and Spain is required to pull the average down. This is also why the entire weight falls to France once Spain is dropped.

Spain's trajectory, however, does not resemble Sweden's. It rises through the pre-treatment period, crosses Sweden before the tax is introduced, and continues to rise thereafter until around 2008. The pre-treatment match is a match in levels between two units moving in opposite directions. Spain's continued growth after 2001 pulls the counterfactual upward, and the estimated effect reflects that divergence.

Figure 15: Per capita residential CO₂ emissions: Sweden and donor pool countries, 1990–2016



Should Spain have been included in the donor pool? According to Blázquez et al. (2013), the Spanish residential final energy demand grew by 73% between 1990 and 2008, at an annual rate roughly four times that observed in other European countries, and the residential electricity demand rose by nearly a third between 2000 and 2008 alone. The drivers were structural: rapid population growth accompanied by an even faster expansion of the housing stock, and falling household size (Blázquez et al., 2013). Electricity use does not itself enter direct residential CO₂ emissions, but the same drivers raise household demand for heating fuels, which does. None of these developments has a Swedish counterpart. Spain’s inclusion means that the counterfactual for Sweden is built primarily on a country undergoing rapid growth in residential energy demand over precisely the period used to construct it. The post-2000 divergence between Sweden and synthetic Sweden seems to reflect that trajectory.

7 Trimming the donor pool

Using simulated data, Abadie and Vives-i-Bastida (2025) show that trimming the donor pool based on the Euclidean distance from the treated unit eliminates structurally different units from the pool, thereby reducing interpolation bias. This section examines whether this remedy carries over to small-donor-pool settings such as the three empirical examples discussed above.

The simulations in Abadie and Vives-i-Bastida (2025) is based on the grouped factor model of Ferman (2021), in which units come in pairs that share a common factor structure. The treated unit therefore shares its factor structure with exactly one of the 199 units in the donor pool, and no treatment effect is imposed. The donors are ranked by distance to the treated unit in predictor space (the predictors are the pre-intervention outcomes) and the closest q percent are retained. As q falls, the weight given to donors sharing the treated unit’s factor structure increases, causing

the spurious post-treatment effect to disappear. Most of this improvement is realized by $q = 0.10$, which leaves approximately 20 donors in the pool. These 20 donors are the closest survivors of an initial pool of 199 units.

I apply the design of Abadie and Vives-i-Bastida (2025) to a donor pool of size $J = 19$ to demonstrate its limited applicability to the small-donor-pool settings typical of country-level policy evaluations. At this scale the synthetic control assigns most of the weight (about 84% on average) to the correct donor without any trimming, and the spurious post-treatment effect is absent from the outset. The bias the quantile rule corrects is thus a feature of large donor pools in this design, where unrelated units are numerous enough to be combined into an apparent pre-treatment fit. [Appendix L](#) details the results of this exercise.

A quantile rule is also purely ordinal. Euclidean distances are always defined and always admit a ranking, whatever their magnitudes, so retaining the closest q percent discards a fixed share of the pool regardless of whether any donor is a good match. The rule will exclude units even when every donor lies close to the treated unit, and will retain the nearest unit even when it lies far from it. The ranking therefore cannot answer the question the guidance actually poses: whether a given donor is similar enough to the treated unit to belong in the pool at all. Table 6 illustrates this point. It ranks each donor by the Euclidean distance between its pre-treatment outcomes and those of the treated unit, for each of the three empirical examples. In none of the three cases is the influential donor the most distant. Luxembourg ranks 18th of 20 and New Zealand 9th of 13, while in Runst and Thonipara (2020), Spain is the closest donor of all. A trim that retains the closest 85% of donors would remove Luxembourg, but nothing in the ranking itself indicates that the cut-off should lie there, and the same rule would keep New Zealand and Spain

Table 6: Euclidean distance between each donor and the treated unit

Rank	Mideksa (2024)		Bretschger and Grieg (2024)		Runst and Thonipara (2020)	
	Donor	Distance	Donor	Distance	Donor	Distance
1	Switzerland	0.64	Belgium	342	Spain	1.56
2	United Kingdom	0.83	France	537	France	1.91
3	Belgium	0.88	Denmark	843	Slovakia	2.58
4	Germany	1.01	Italy	1,200	Austria	3.50
5	France	1.09	Japan	1,320	Poland	5.28
6	Iceland	1.23	Spain	1,546	Germany	7.65
7	Netherlands	1.35	Ireland	2,108	Belgium	7.89
8	Austria	1.55	Greece	2,258	Czech Republic	8.49
9	New Zealand	1.85	New Zealand	3,289		
10	Japan	2.36	Portugal	3,824		
11	Italy	2.41	Australia	7,809		
12	Ireland	2.73	Canada	13,381		
13	Spain	2.96	United States	16,243		
14	Greece	3.70				
15	Portugal	5.00				
16	Turkey	6.10				
17	Australia	6.46				
18	Luxembourg	8.11				
19	Canada	12.50				
20	United States	17.00				

Notes: Euclidean distance between each donor and the treated unit: Finland for Mideksa (2024), the UK for Bretschger and Grieg (2024), and Sweden for Runst and Thonipara (2020). Distances are computed on pre-treatment outcomes over the pre-treatment periods of each study (metric tons per capita for Mideksa, and Runst and Thonipara; kg per capita for Bretschger and Grieg). Donor countries are ranked from closest to most distant. Shaded cells indicate the influential donors identified for each of the studies.

8 Discussion

The three replications share a common pattern. In each case, the synthetic control achieves a good pre-treatment fit, but it does so by relying on a donor that is a poor comparison for the treated unit. Luxembourg's recorded transport emissions are inflated by fuel tourism, New Zealand's trajectory reflects country-specific shocks, and Spain's emissions match Sweden's in levels while moving in the opposite direction. None of these donors is well suited to represent the treated unit's counterfactual, yet each is rewarded by the optimization precisely because it improves the pre-treatment fit. Excluding it attenuates the estimated effect or leaves no credible counterfactual at all.

This is the mechanism of interpolation bias. The synthetic control reproduces the treated unit as a weighted average of donors, and a close pre-treatment match is taken as evidence that the average also approximates the treated unit's untreated outcome after the intervention. That inference holds only if the donors are structurally similar to the treated unit. When the match is achieved by combining a dissimilar donor with other donors, the pre-treatment fit reflects the flexibility of the weighting rather than a shared underlying process, and after treatment the counterfactual inherits the donor's own trajectory. The Runst and Thonipara (2020) example is the clearest case: Sweden can only be matched by averaging donors on either side of it, and the unit that makes this possible is moving in the opposite direction. In the other two cases, the mechanism differs in its details, but the consequence is the same: pre-treatment fit, the prerequisite of SCM credibility, is improved by exactly the donor that undermines the counterfactual.

8.1 Why did the leave-one-out tests not detect the influential donors?

The leave-one-out test (Section 2.3), which assesses whether the estimated effect depends on any single donor, is the standard diagnostic for donor sensitivity and, by extension, one of the few available checks for interpolation bias. In principle, it should reveal the donor dependence documented in the three empirical examples. Runst and Thonipara (2020) do not report the test. Mideksa (2024) and Bretschger and Grieg (2024) both do, and conclude that their results are robust to the exclusion of any single donor. The remainder of this section examines why.

Mideksa (2024) acknowledges the sensitivity of his results to the exclusion of Luxembourg, though only partially. He first concludes that the result is qualitatively robust to removing any country with a positive weight, provided the pre-treatment fit remains acceptable. He then notes that dropping Luxembourg produces the poorest pre-treatment fit among all leave-one-out perturbations, and characterizes the resulting estimates as a lower bound, although a poorer fit makes an estimate less reliable rather than biased in a known direction. He then argues, on probabilistic grounds, that if each leave-one-out possibility is equally likely, the exclusion of Luxembourg is only one of six scenarios. This reasoning misinterprets the purpose of the test. As discussed in Abadie (2021), the leave-one-out test is not a probabilistic exercise but a diagnostic of whether any single donor drives the estimated treatment effect. The relevant question is not how often Luxembourg would be absent by chance, but whether its exclusion changes the result. It does: the estimated effect shrinks to about a third of its original size, and the pre-treatment fit deteriorates sharply. Because the SCM makes the composition of the counterfactual transparent, this is exactly the outcome that should prompt scrutiny of Luxembourg's inclusion, rather than reassurance that it represents one scenario in six.

As an additional robustness test, Mideksa conducts what he describes as an extreme test, stating that "Abadie (2000) does not suggest this extreme test and it is unseen in the synthetic control

literature.”²⁶ The test entails excluding all countries with weights above 15%. In the first round, countries with a total weight of 77.5% are removed; Luxembourg survives, since its weight of 10.2% falls below the threshold.²⁷ In the second round, countries with a total weight of 86.9% are removed, and Luxembourg survives again.²⁸ The test therefore cannot detect the problem by design: it targets donors with large weights, whereas Luxembourg’s influence far exceeds its weight. This again highlights the distinction between high-weight and high-influence donors.

Bretschger and Grieg (2024) also report a leave-one-out test, re-estimating the synthetic control while excluding in turn each of the four donors with a positive weight. They note that sensitivity to a single exclusion would indicate that the results are driven by developments in that donor country. However, rather than examining each exclusion separately, they report the mean effect across the four re-estimates, a reduction of 0.325 tonnes per capita per year, close to the main estimate. Averaging defeats the purpose of the test. The question it asks is whether any single exclusion changes the result, and averaging dilutes the one exclusion that matters among those that do not.

In both cases, the leave-one-out test produced the relevant signal; however, a single re-estimate per donor can be dismissed as merely one scenario among several or as one estimate within an average. The resampling exercise differs from the leave-one-out test in three key respects. First, instead of generating a single re-estimate per excluded donor, it produces a distribution of estimates with and without each donor. Second, because the remaining donors vary across subsets, the method demonstrates whether the consequences of excluding a donor persist across different compositions of the donor pool. Third, it compares the pre-treatment fit with the estimated effect, clarifying whether the donor influencing the treatment effect is also the one on which the fit depends. The influence of a donor is thus indicated by a systematic separation between two distributions, which is considerably more difficult to dismiss than a single re-estimate.

8.2 Influential donors and the transparency of synthetic controls

In each of the empirical examples above, I investigated the suitability of the influential donor only after it had been identified. One might question this sequence, but it in fact reflects one of the strengths of the method. Investigating a donor after it has been identified as influential is not a departure from standard practice but the use Abadie (2021) envisages for the method. He states:

²⁶The reference appears to be an error. No such Abadie (2000) paper on synthetic controls exists, nor does it appear in the paper’s bibliography. Additionally, the method itself was introduced three years later, in Abadie and Gardeazabal (2003).

²⁷The countries excluded are the UK (43.2%), New Zealand (15.9%), and Turkey (18.4%). Their weights add up to 77.5% of the total weight.

²⁸Mideksa (2024) does not specify which countries are removed, but based on the reported weights, I infer that they are Switzerland (9.4%), the UK (43.2%), New Zealand (15.9%), and Turkey (18.4%). The weights of these countries sum to 86.9%, and no other combination of donors sums to this total.

Synthetic controls make explicit the contribution of each comparison unit to the counterfactual of interest. Moreover, because the synthetic control coefficients are proper weights and are sparse [...], they allow a simple and precise interpretation of the nature of the estimate of the counterfactual of interest. [...] Simplicity and transparency of the counterfactual allows the use of the expert knowledge to evaluate the validity of a synthetic control and the directions of potential biases. (Abadie, 2021, p. 406)

The weights, however, reveal each donor's contribution to the counterfactual, not its influence on the estimated effect, and the two can diverge. In Mideksa (2024), Luxembourg receives 10.2% of the weight, well below the UK's 43.2%, yet it is the only donor whose exclusion overturns the result. Mideksa's own extreme test likewise targets donors by weight rather than influence, which is why Luxembourg survives both rounds. The resampling exercise addresses this problem. By comparing distributions of pre-treatment fit and estimated effects with and without each donor, it identifies high-influence donors irrespective of their weight. It also constrains interpretation. A single leave-one-out estimate, or a test that targets donors by weight, leaves the researcher free to judge whether the result is robust; a systematic separation between two distributions leaves far less room for that judgment. In this sense, the resampling exercise extends the transparency of the synthetic control from weight to influence. An influential donor is not necessarily an unsuitable one. A donor may drive the estimated effect precisely because it is the closest structural match to the treated unit. Influence is therefore a reason for scrutiny, not for exclusion. Whether an influential donor belongs in the pool remains a matter of expert judgment, and it is this judgment that Sections 4.3, 5.3, and 6.2 supply.

8.3 Policy and replication

All three cases highlight the importance of replicating and re-analyzing published estimates, yet replications of this kind remain rare in economics (Ankel-Peters et al., 2025). Each of the original studies reports a good pre-treatment fit and, in two cases, a leave-one-out test, and each would pass a check of computational reproducibility. The sensitivity documented here becomes visible only when the analytical choices themselves are revisited, in this case the composition of the donor pool. Stevenson and Fischman (2026) argue that econometric practice treats the researcher as if they mechanically implemented the method, and that ignoring the researcher's role leads to inflated and overconfident estimates. This makes replication that revisits the researcher's choices all the more important. Where published estimates carry direct policy weight, as they do in the evaluation of CO₂ taxes, the cost of leaving such choices unexamined is correspondingly higher.

The consequences of leaving primary estimates unchecked extend beyond the individual study. Döbbling-Hildebrandt et al. (2024) conduct a systematic review and meta-analysis of the ex-post carbon pricing literature, drawing on 483 effect sizes from 80 causal evaluations across 21 CO₂ pricing schemes. They find immediate and substantial emission reductions for at least 17 of

these policies, with statistically significant reductions ranging from 5% to 21% across schemes, or 4% to 15% after correcting for publication bias. Runst and Thonipara (2020) and Mideksa (2024) are among the primary studies included.²⁹ As shown above, neither estimate survives a defensible change to the donor pool. This adds to concerns already raised about the strength of the evidence underlying the meta-analysis (Piseddu et al., 2026), and points to a distinct source of fragility in the primary studies. This is not a criticism of the meta-analysis itself, but a point about what evidence synthesis can and cannot do. Effect sizes are extracted as reported, and their sensitivity is not evident in the published output. When primary estimates are not scrutinized individually, aggregation carries them forward, and the aggregate reaches policy audiences. Döbbeling-Hildebrandt et al. (2024) is cited in the World Bank’s *State and Trends of Carbon Pricing 2025* (World Bank, 2025). Both Mideksa (2024) and Runst and Thonipara (2020) are also included in the OECD’s Climate Policy Effectiveness Network, a data platform that compiles ex-post evaluations of climate policies so that policymakers and researchers can compare their effectiveness across countries, policy types, and sectors (OECD, 2025).³⁰ The platform is part of the OECD’s Inclusive Forum on Carbon Mitigation Approaches. Estimates that are not robust to the composition of the donor pool thus enter directly into the evidence base intended to guide policy design.

9 Conclusion

This paper began from a simple observation: the methodological literature is clear that donor units should be similar to the treated unit, but offers no objective criterion for deciding which units qualify. Using the empirical CO₂ pricing literature, I showed that this ambiguity translates into markedly different choices in practice, even in studies applying near-identical designs to neighboring, and arguably very similar, countries.

I then examined what follows from those choices. Replications of Mideksa (2024), Bretschger and Grieg (2024), and Runst and Thonipara (2020) show that in each case the estimated effect rests on a single structurally unsuitable donor: Luxembourg for Finland, New Zealand for the UK, and Spain for Sweden. Resampling the donor pool identifies these donors systematically, and excluding them either renders the estimated effect indistinguishable from placebo effects or leaves no credible counterfactual at all. The published conclusions therefore rest on the composition of the donor pool rather than on the policy itself. The recommended robustness tests do not reliably reveal this. Where leave-one-out tests were reported, the relevant signal was present but was explained away, and the influential donors are not those furthest from the treated unit in Euclidean distance, nor necessarily those with the largest weights.

²⁹Döbbeling-Hildebrandt et al. (2024) cite the CESifo working paper version of Mideksa (2024).

³⁰See <https://oecd-main.shinyapps.io/CPEN/>. Bretschger and Grieg (2024) is not included in either this platform or Döbbeling-Hildebrandt et al. (2024), likely because it was published after their literature searches were completed.

Resampling subsets of the donor pool offers researchers a straightforward way to identify influential donors. By re-estimating the synthetic control across many donor combinations, it reveals which donors drive the estimated effect, regardless of their weight or their distance from the treated unit. An influential donor is not necessarily an unsuitable one, but its influence shows researchers where scrutiny is needed: whether the donor is a suitable comparison for the treated unit must then be assessed on substantive grounds, as in the three cases examined here.

The underlying vulnerability is general. Although the SCM selects weights optimally given the donor pool it is handed, the composition of that pool is far from innocuous. A unit that is structurally unsuitable as a comparison may nonetheless receive positive weight and, in combination with other donors, reproduce the pre-treatment trajectory of the treated unit. The result is a good fit that does not reflect a shared underlying process, and that may conceal interpolation bias. Because a good pre-treatment fit is treated as a prerequisite for using the SCM, the very condition researchers seek to satisfy can reward the donor that undermines the counterfactual.

Two directions for future research follow. First, objective criteria for selecting the donor pool in small-sample settings are still needed; donor selection rules suited to country-level evaluation remain to be developed. While resampling subsets of the donor pool provides a diagnostic for detecting influential donors, it does not establish *ex ante* what constitutes sufficient similarity to the treated unit. Second, the findings make the case for more replication. Applying the resampling exercise to the wider SCM literature on carbon pricing would show how common donor dependence is, and whether evidence syntheses such as Döbbeling-Hildebrandt et al. (2024) change when donor-sensitive estimates are accounted for.

References

- Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2), 391–425.
- 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.
- Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2), 495–510.
- Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, 93(1), 113–132.
- Abadie, A., & Vives-i-Bastida, J. (2025). Synthetic controls in action. *Advances in Economics and Econometrics: Twelfth World Congress*.
- Andersson, J. J. (2019). Carbon taxes and CO2 emissions: Sweden as a case study. *American Economic Journal: Economic Policy*, 11(4), 1–30.
- Ankel-Peters, J., Fiala, N., & Neubauer, F. (2025). Is economics self-correcting? Replications in the American Economic Review. *Economic Inquiry*, 63(2), 463–485.
- Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), 3–32.
- Aus dem Moore, N., Brehm, J., & Gruhl, H. (2025). Driving innovation? Carbon tax effects in the Swedish transport sector. *Journal of Public Economics*, 248, 105444.
- Baranzini, A., Goldemberg, J., & Speck, S. (2000). A future for carbon taxes. *Ecological Economics*, 32(3), 395–412.
- Basaglia, P., Behr, S. M., & Drupp, M. A. (2025). Fuel taxation and environmental externalities: Evidence from the world’s largest environmental tax reform. *CESifo Working Paper No. 11949*.
- Blázquez, L., Boogen, N., & Filippini, M. (2013). Residential electricity demand in Spain: New empirical evidence using aggregate data. *Energy Economics*, 36, 648–657.
- Bretschger, L., & Grieg, E. (2024). Carbon taxes, CO2 emissions, and the economy: The effects of fuel taxation in the UK. *Energy Policy*, 195, 114359.
- Chernozhukov, V., Wüthrich, K., & Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536), 1849–1864.
- Commonwealth of Australia. (2014). Clean energy legislation (carbon tax repeal) act 2014 [Act No. 83, 2014]. <https://www.legislation.gov.au/C2014A00083>
- Döbbeling-Hildebrandt, N., Miersch, K., Khanna, T. M., Bachelet, M., Bruns, S. B., Callaghan, M., Edenhofer, O., Flachsland, C., Forster, P. M., Kalkuhl, M., et al. (2024). Systematic review and meta-analysis of ex-post evaluations on the effectiveness of carbon pricing. *Nature Communications*, 15(1), 4147.

- Douenne, T., & Fabre, A. (2022). Yellow vests, pessimistic beliefs, and carbon tax aversion. *American Economic Journal: Economic Policy*, 14(1), 81–110.
- Ferman, B. (2021). On the properties of the synthetic control estimator with many periods and many controls. *Journal of the American Statistical Association*, 116(536), 1764–1772.
- Ferman, B., Pinto, C., & Possebom, V. (2020). Cherry picking with synthetic controls. *Journal of Policy Analysis and Management*, 39(2), 510–532.
- Fernando, S. (2019). The environmental effectiveness of carbon taxes: A comparative case study of the Nordic experience [14–15 February 2019]. *Proceedings of the First International Research Conference on Carbon Pricing*, 349–369.
- Firpo, S., & Possebom, V. (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, 6(2), 20160026.
- Green, J. F. (2021). Does carbon pricing reduce emissions? A review of ex-post analyses. *Environmental Research Letters*, 16(4), 043004. <https://doi.org/10.1088/1748-9326/abdae9>
- Hasan, M. A., Frame, D. J., Chapman, R., & Archie, K. M. (2019). Emissions from the road transport sector of New Zealand: Key drivers and challenges. *Environmental Science and Pollution Research*, 26(23), 23937–23957.
- Johnston, W., Elliott, J., Fletcher, I., & Lauder, G. (1983). New Zealand's carless day scheme: People's responses to restriction of car use. *Transport Reviews*, 3(4), 399–410.
- Kaul, A., Klößner, S., Pfeifer, G., & Schieler, M. (2022). Standard synthetic control methods: The case of using all preintervention outcomes together with covariates. *Journal of Business & Economic Statistics*, 40(3), 1362–1376.
- Leroutier, M. (2022). Carbon pricing and power sector decarbonization: Evidence from the UK. *Journal of Environmental Economics and Management*, 111, 102580.
- Mideksa, T. K. (2024). Pricing for a cooler planet: An empirical analysis of the effect of taxing carbon. *Journal of Environmental Economics and Management*, 103034.
- Nordhaus, W. (2019). Climate change: The ultimate challenge for economics. *American Economic Review*, 109(6), 1991–2014.
- OECD. (2025, May). *The effects of climate policies on emissions: Evidence from a comprehensive and systematic review of the ex-post empirical literature* (Inclusive Forum on Carbon Mitigation Approaches Papers No. 7). OECD Publishing. Paris. <https://doi.org/10.1787/ec1f7f0d-en>
- Piseddu, E., Brodeur, A., Rose, J., Sievert, M., & Ankel-Peters, J. (2026, April). *The power of carbon pricing: A comment on Döbbeling-Hildebrandt et al. (2024) and its press release* (IZA Discussion Paper No. 18579). Institute of Labor Economics (IZA). Bonn. <https://www.iza.org/publications/dp18579>
- Pizer, W. A. (2002). Combining price and quantity controls to mitigate global climate change. *Journal of Public Economics*, 85(3), 409–434.
- Pretis, F. (2022). Does a carbon tax reduce CO₂ emissions? evidence from British Columbia. *Environmental and Resource Economics*, 83(1), 115–144.

- Runst, P., & Höhle, D. (2022). The German eco tax and its impact on CO2 emissions. *Energy Policy*, 160, 112655.
- Runst, P., & Thonipara, A. (2020). Dosis facit effectum why the size of the carbon tax matters: Evidence from the Swedish residential sector. *Energy Economics*, 91, 104898.
- Stern, N. (2008). The economics of climate change. *American Economic Review*, 98(2), 1–37.
- Sterner, T. (2006). *Survey of transport fuel demand elasticities* (tech. rep.). Naturvårdsverket. Stockholm.
- Stevenson, M. T., & Fischman, J. B. (2026). Humans in the loop: The next frontier in the credibility revolution. *Journal of Economic Literature*, 64(3), 801–828.
- Vives-i-Bastida, J. (2022). Predictor selection for synthetic controls. *arXiv preprint arXiv:2203.11576*.
- World Bank. (2025, June). *State and trends of carbon pricing 2025* (Report). World Bank. Washington, DC. Retrieved September 3, 2026, from <http://hdl.handle.net/10986/43277>
- Xiang, D., & Lawley, C. (2019). The impact of British Columbia’s carbon tax on residential natural gas consumption. *Energy Economics*, 80, 206–218.

Appendices

Appendix A Comparison of synthetic control estimates across software packages

Figure A1: Replication results across different software packages

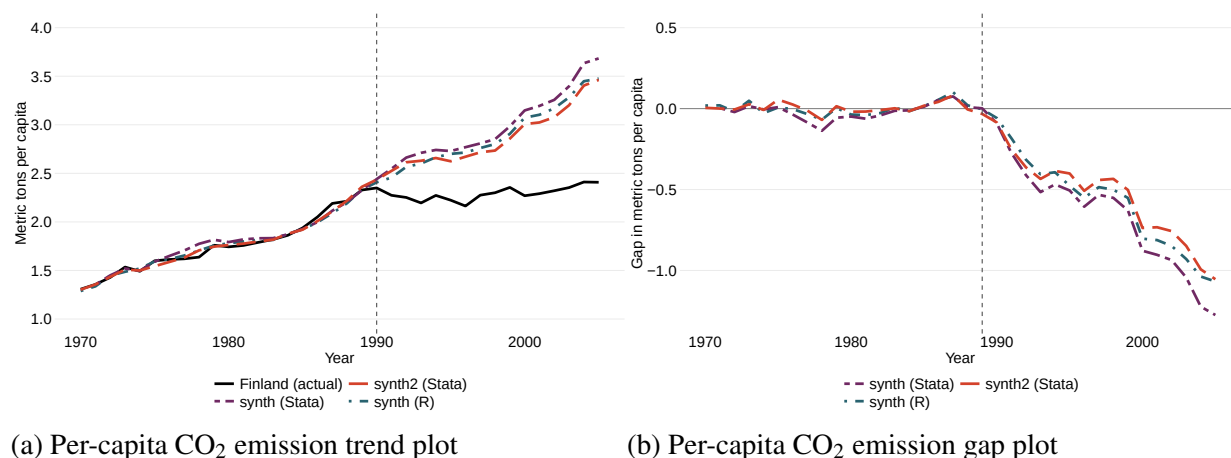


Table A1 compares the unit weights assigned in the construction of the synthetic control across software implementations, with the first column (shaded in gray) reporting the weights from the original study. The weights produced by `synth2` most closely resemble those of the original study, with the same countries receiving weights of similar magnitude.³¹ Notwithstanding these numerical differences, all three software implementations assign weights to a similar set of donor countries, and the qualitative conclusions evident in Figure A1 are consistent across implementations.

³¹The original study uses `synth` in Stata with the `nested` and `allop` optimization options. However, these options have been reported to be unreliable in more recent versions of Stata (such as version 18), frequently causing convergence failures (see <https://www.statalist.org/forums/forum/general-stata-discussion/general/1375084-nested-option-in-synth-ado> for a discussion). Additionally, the `nested` and `allop` optimization options are sensitive to the order in which predictors are specified, which can yield different weights (Fernando, 2019). Therefore, the weights reported under `synth` are estimated without the `nested` and `allop` optimization options. This issue does not occur in the newer `synth2` package; consequently, the synthetic weights generated via `synth2` most closely resemble those reported in the original study.

Table A1: Donor countries and synthetic control weights across software packages

Donor Country	(Mideksa, 2024)	synth (Stata)	synth2 (Stata)	Synth (R)
Australia	0	0	0	0
Austria	0	0	0	0
Belgium	0	0.118	0	0
Canada	0	0	0	0
Switzerland	0.094	0	0.095	0.099
Germany	0	0	0	0
Spain	0	0	0	0
France	0	0	0	0
UK	0.432	0	0.44	0
Greece	0	0	0	0
Ireland	0	0	0	0
Iceland	0	0	0	0
Italy	0	0.22	0	0.263
Japan	0	0	0	0
Luxembourg	0.102	0.112	0.099	0.076
Netherlands	0	0.007	0	0
New Zealand	0.159	0.122	0.164	0.432
Portugal	0	0	0	0
Turkey	0.184	0.323	0.175	0.122
USA	0.029	0.098	0.026	0.006
$\Sigma \approx$	1	1	1	1

Appendix B Replication of Mideksa (2024)’s predictor balance table

Table B1 reports the average predictor values for Finland, synthetic Finland, and the OECD donor average over 1971–1989, the window over which Mideksa averages the predictors, with the columns shaded in gray corresponding to the values reported in the original study. The comparison confirms that synthetic Finland more closely approximates the predictor values of actual Finland than the unweighted OECD average. Since the weights produced by `synth2` do not exactly replicate those of the original study, some discrepancy between the two synthetic Finland columns is to be expected. More noteworthy is the discrepancy in the OECD average columns, which, given that both analyses draw on the same underlying dataset, should not arise. Upon inspection, this appears to stem from Mideksa computing the OECD average over the full sample period (1970–2005) rather than the pre-treatment period alone. This inflates the gap between the OECD average and actual Finland, making synthetic Finland appear a considerably better fit than it does when the average is restricted to the pre-treatment period, as done here.

Table B1: Predictor balance

	Finland	Synthetic Finland		OECD Average	
		(Mideksa, 2024)	<code>synth2</code>	(Mideksa, 2024)	pre-treatment average
Urban population (%)	71.35	71.30	71.67	73.10	71.06
GDP per capita (PPP)	16,722.56	16,730.78	16,890	21,793.98	17,204.41
Gasoline consumption per capita	285.80	374.64	379.02	418.94	389.61
Motor vehicles (per 1000 people)	291.67	314.59	320.4	421.76	341.51

Notes: Columns shaded in gray report values from Mideksa (2024). Finland’s predictor averages are identical across the original study and the replication, as both draw on the same underlying dataset. The replicated synthetic Finland and OECD average columns (`synth2`) are computed over the predictor averaging window 1971–1989, where the OECD average is the unweighted mean across all donor countries. The synthetic Finland values are the weighted averages of donor country predictors using the unit weights derived from the `synth2` implementation. As noted in the text, Mideksa (2024) reports OECD averages computed over the full sample period (1970–2005), which accounts for the discrepancy between the two OECD average columns.

Appendix C Placebo inference: main results

Figure C1: Ratio of post-RMSPE/pre-RMSPE using `synth` in Stata

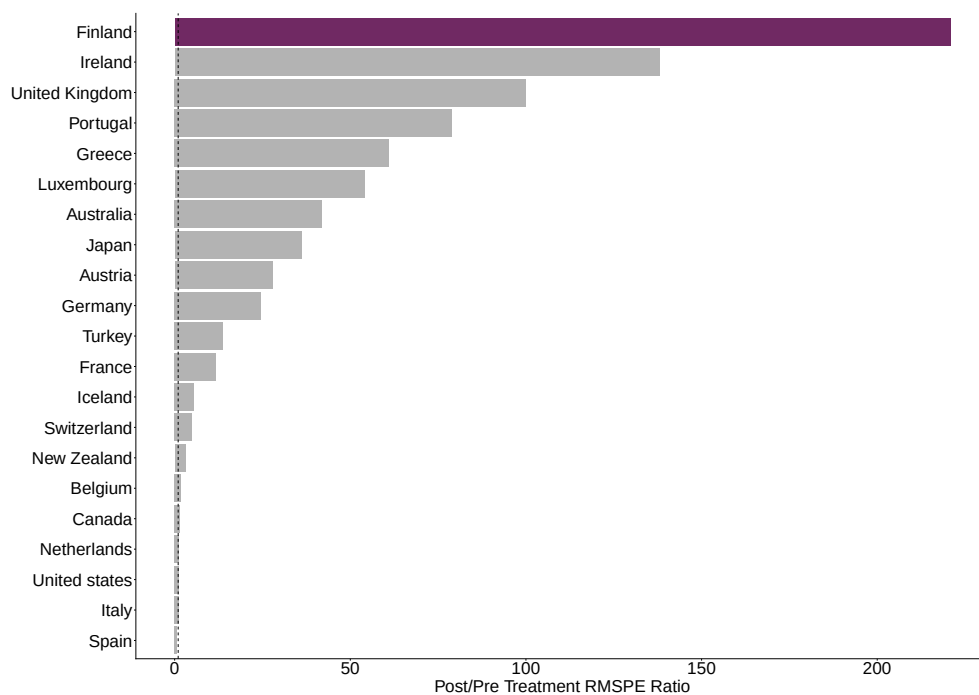
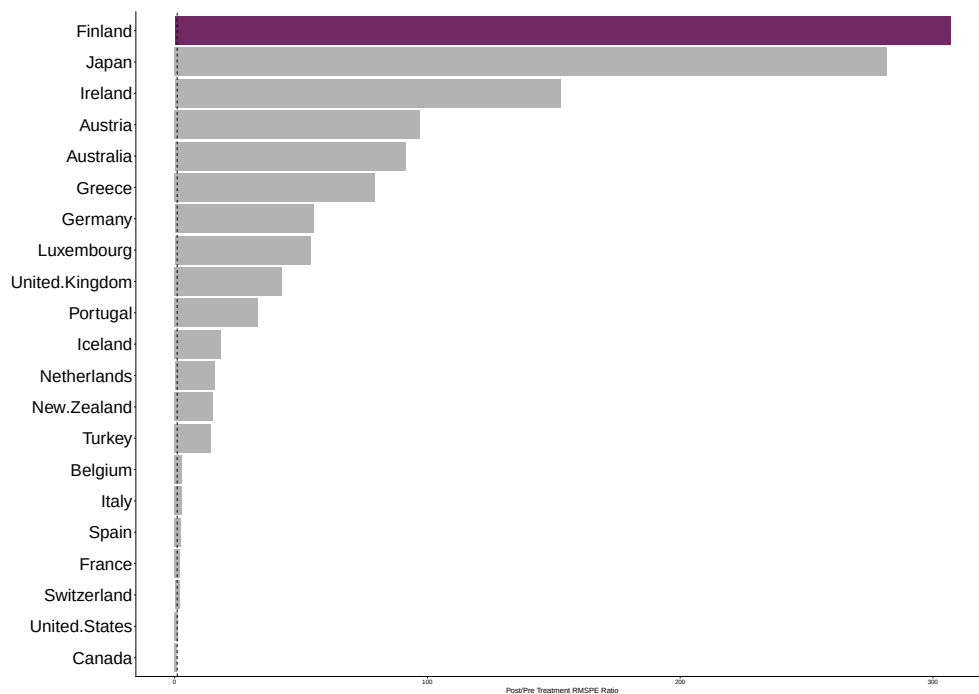


Figure C2: Ratio of post-RMSPE/pre-RMSPE using `synth` in R



Appendix D Placebo inference: results without Luxembourg

Figure D1: Ratio of post-RMSPE/pre-RMSPE using `synth` in Stata

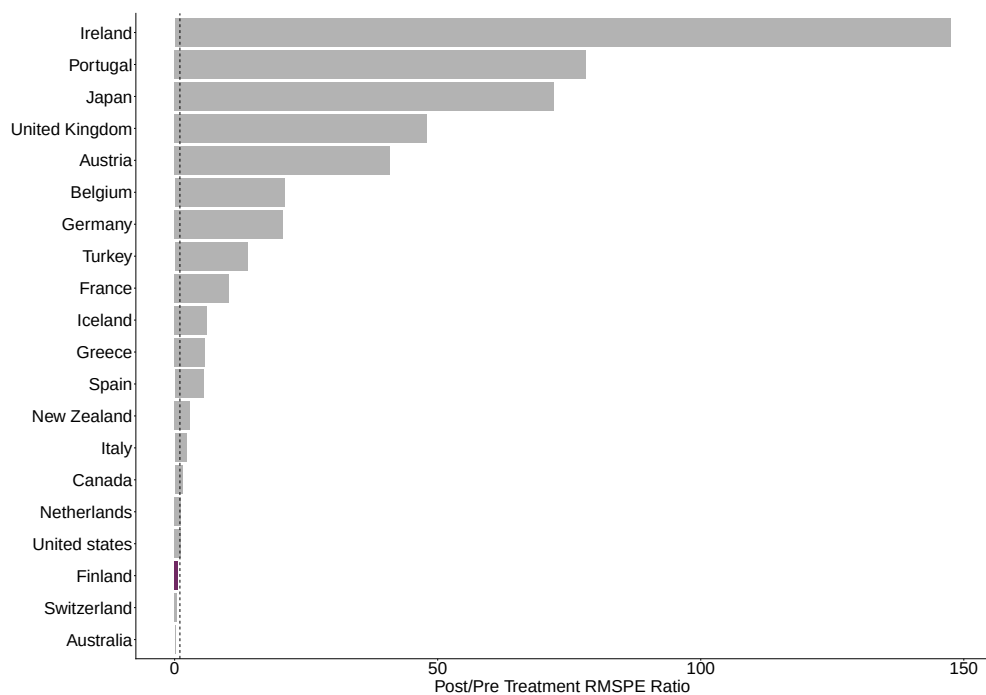
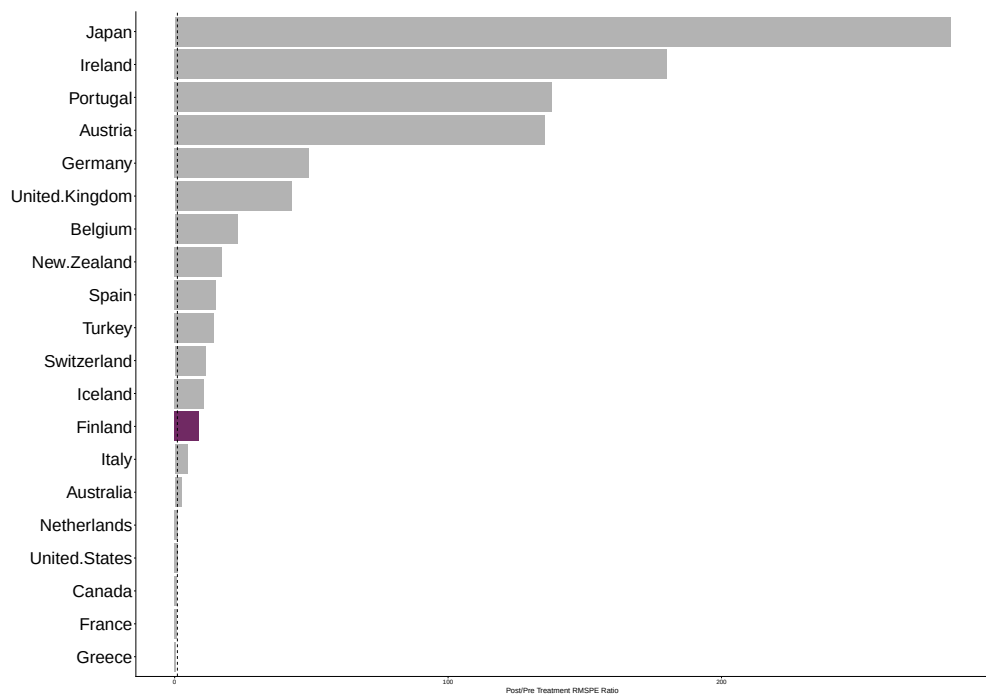
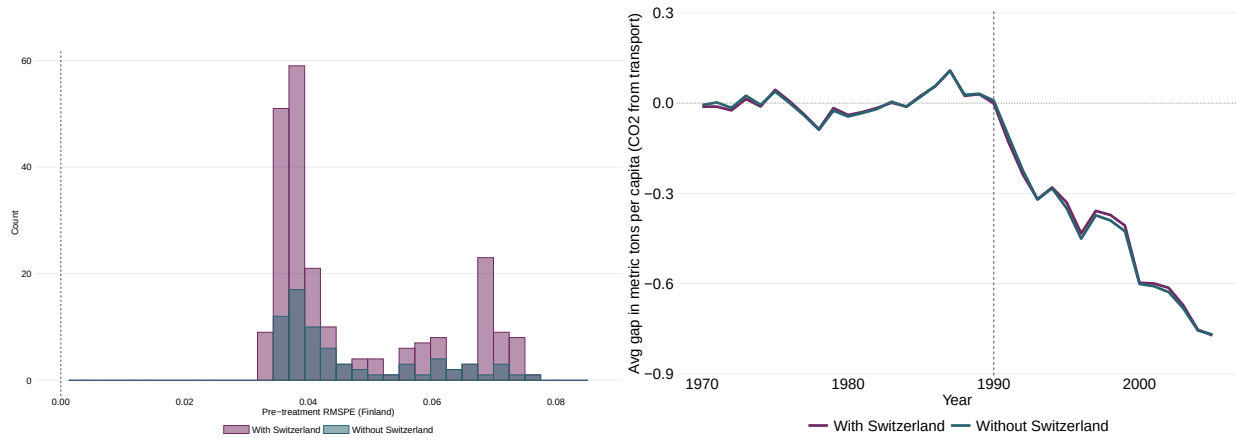


Figure D2: Ratio of post-RMSPE/pre-RMSPE using `synth` in R



Appendix E Subsample analysis: results excluding other donors that received positive weights in Mideksa (2024)

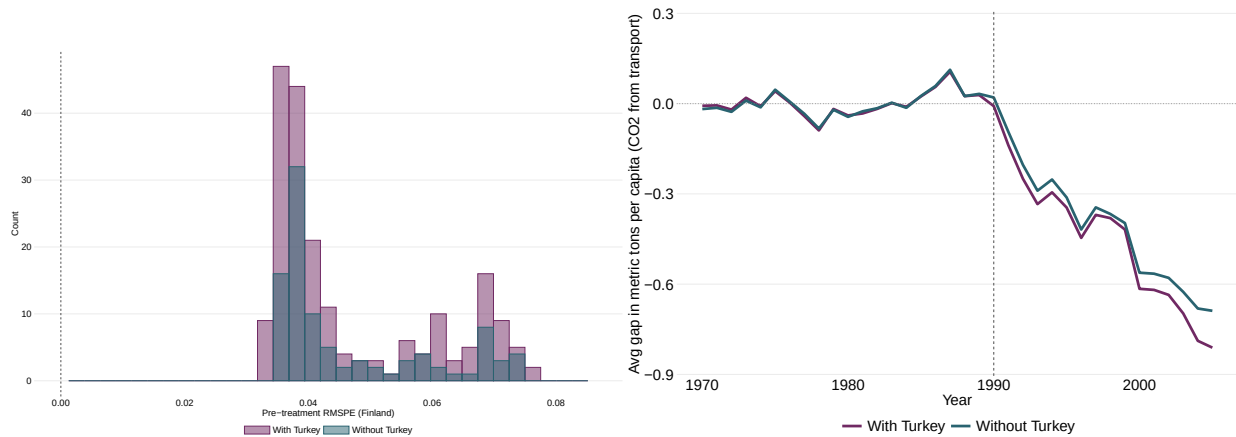
Figure E1: Samples with and without Switzerland



(a) Distribution of pre-treatment RMSPEs

(b) Average emission gap across all samples

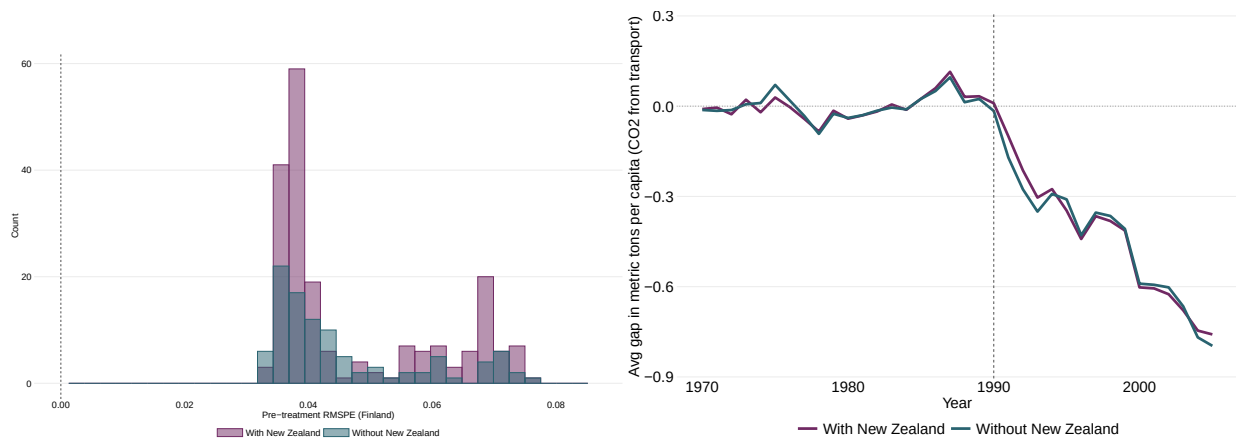
Figure E2: Samples with and without Turkey



(a) Distribution of pre-treatment RMSPEs

(b) Average emission gap across all samples

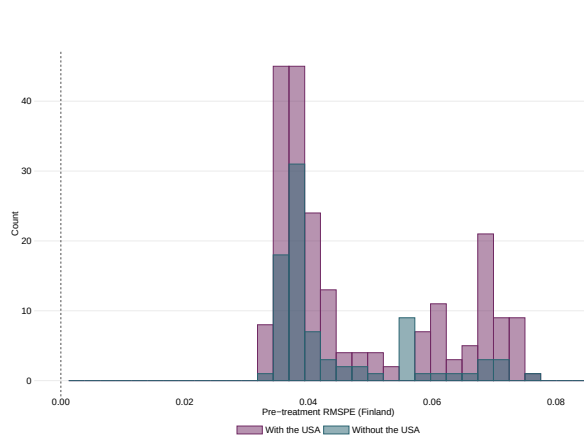
Figure E3: Samples with and without New Zealand



(a) Distribution of pre-treatment RMSPEs

(b) Average emission gap across all samples

Figure E4: Samples with and without the USA

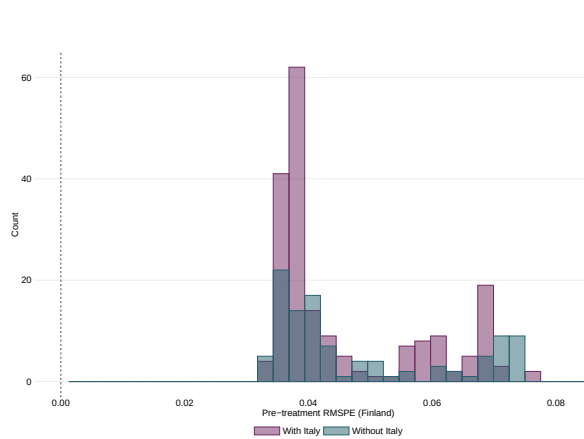


(a) Distribution of pre-treatment RMSPEs

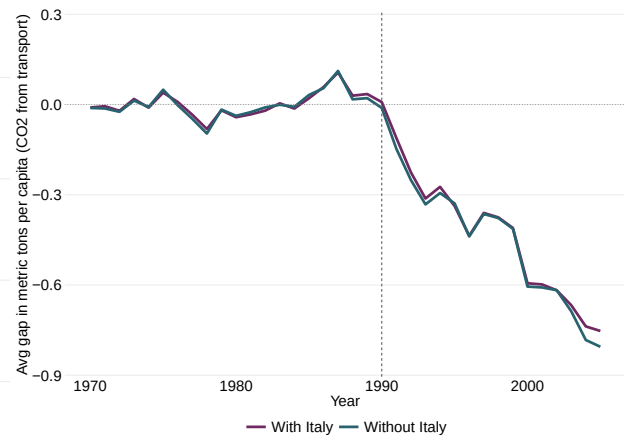


(b) Average emission gap across all samples

Figure E5: Samples with and without Italy



(a) Distribution of pre-treatment RMSPEs



(b) Average emission gap across all samples

Appendix F Subsample analysis: results with Luxembourg

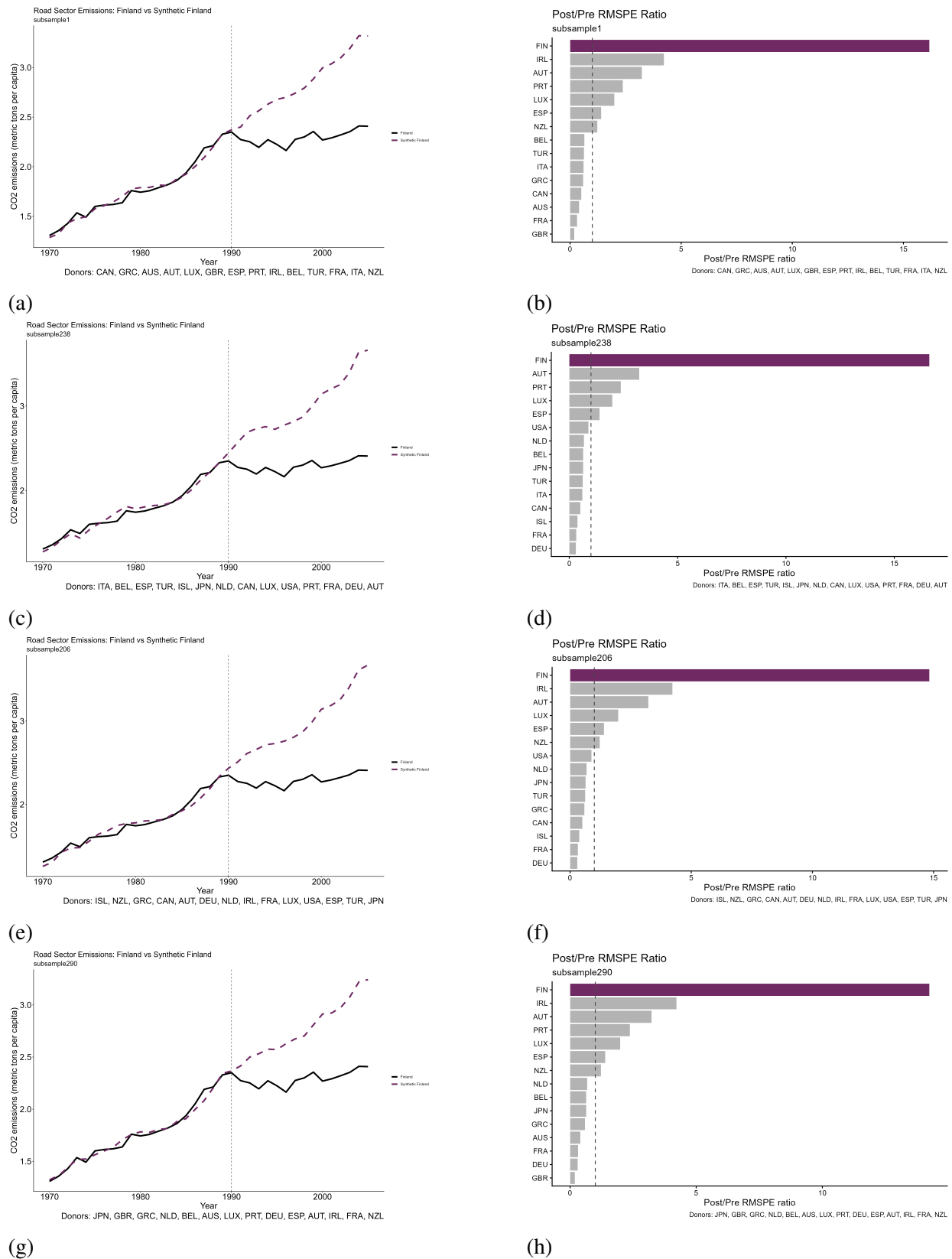


Figure F1: Four illustrative synthetic control estimates drawn from subsamples that include Luxembourg. Figures (a), (c), (e), and (g) show the path plots of Finland versus synthetic Finland. Figures (b), (d), (f), and (h) show the corresponding post/pre-RMSPE ratios; the vertical line marks a ratio of 1, where post-treatment and pre-treatment RMSPE are equal.

Appendix G Subsample analysis: results without Luxembourg

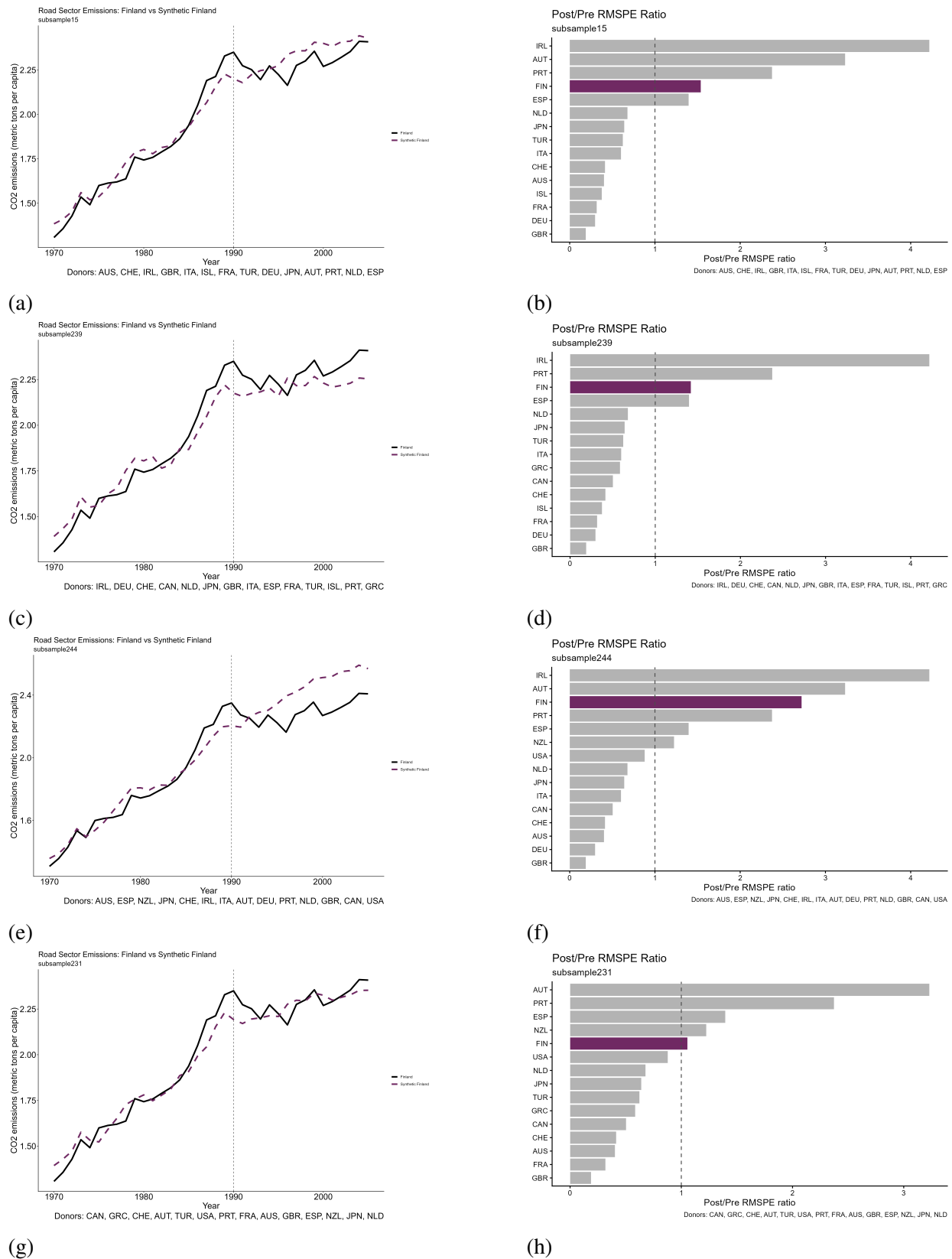


Figure G1: Four illustrative synthetic control estimates drawn from subsamples that exclude Luxembourg. Figures (a), (c), (e), and (g) show the path plots of Finland versus synthetic Finland. Figures (b), (d), (f), and (h) show the corresponding post/pre-RMSPE ratios; the vertical line marks a ratio of 1, where post-treatment and pre-treatment RMSPE are equal.

Appendix H Subsample analysis: results excluding other donors that received positive weights in Bretschger and Grieg (2024)

Figure H1: Samples with and without Greece

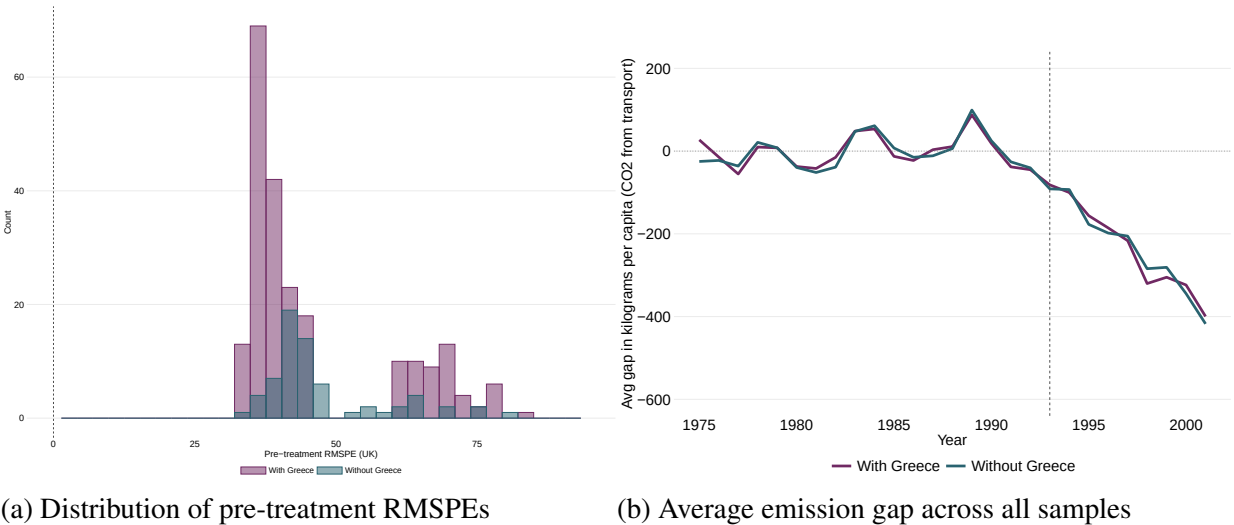


Figure H2: Samples with and without Japan

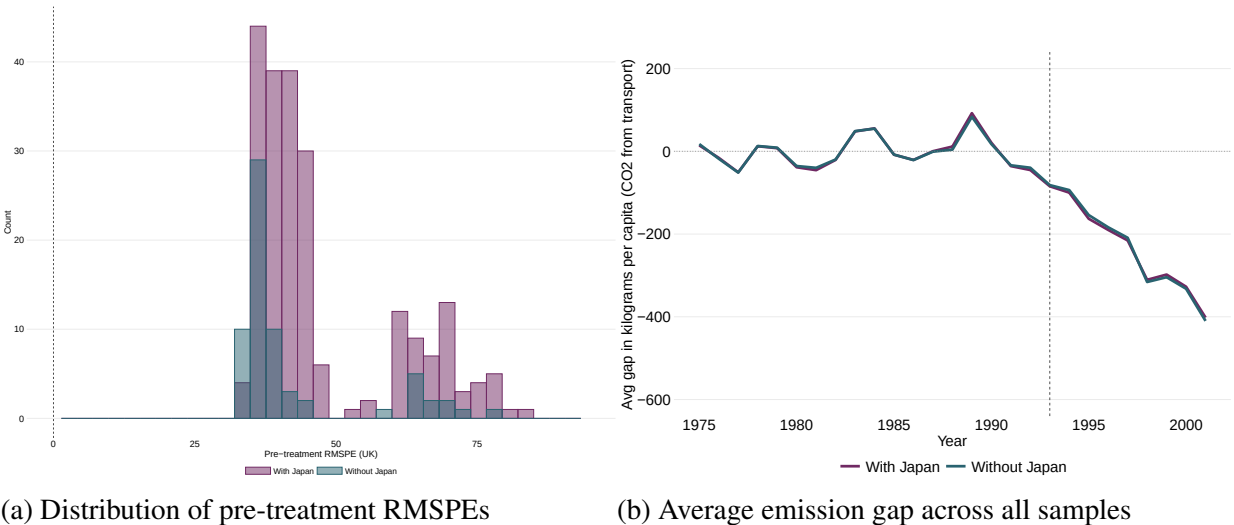
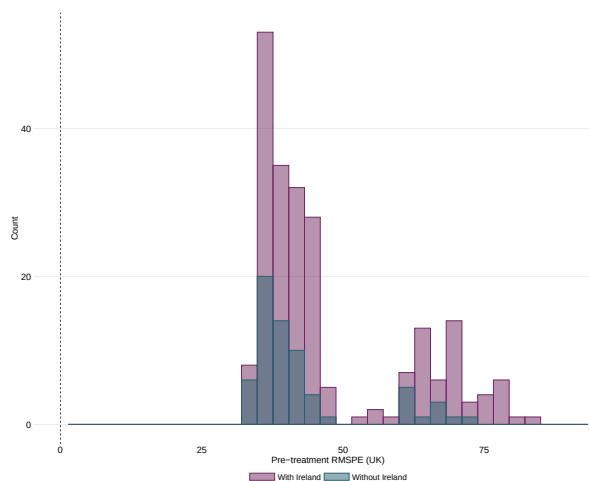
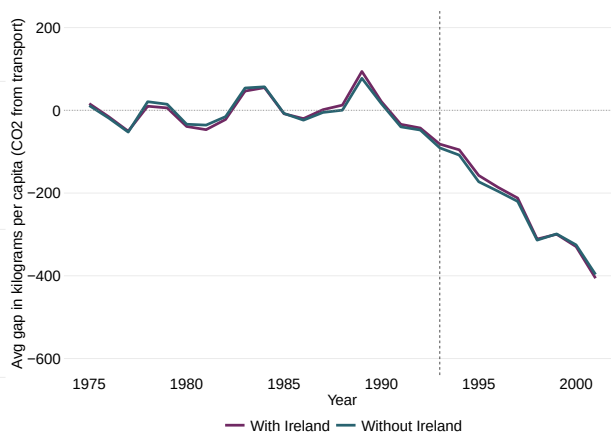


Figure H3: Samples with and without Ireland

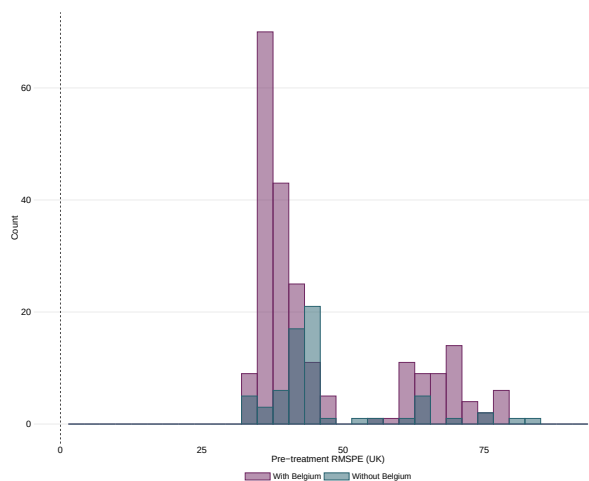


(a) Distribution of pre-treatment RMSPEs

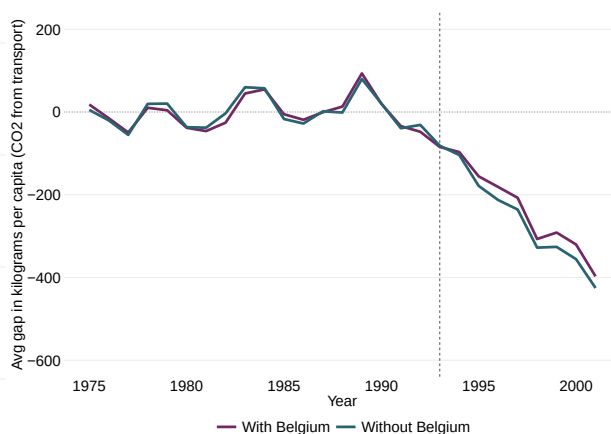


(b) Average emission gap across all samples

Figure H4: Samples with and without Belgium

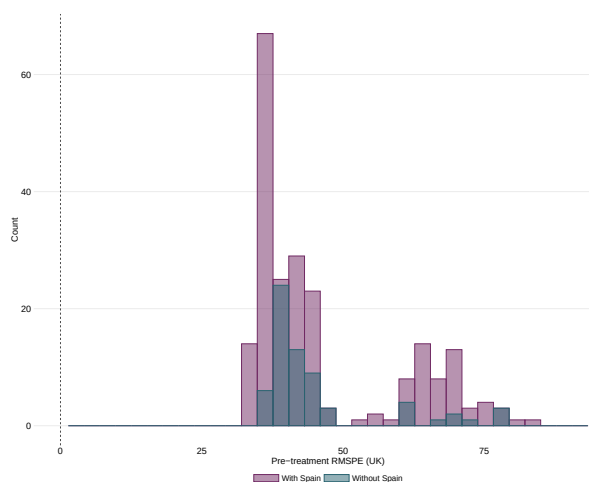


(a) Distribution of pre-treatment RMSPEs

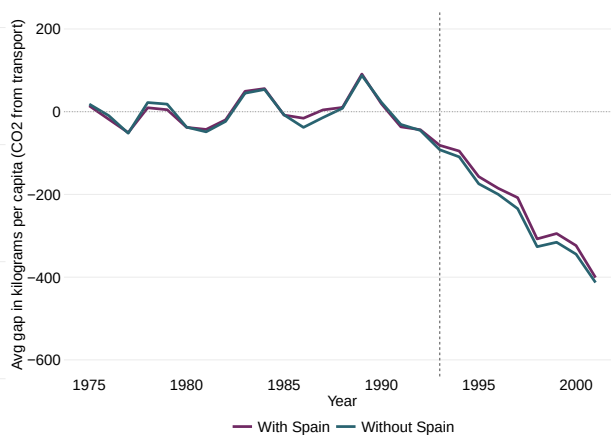


(b) Average emission gap across all samples

Figure H5: Samples with and without Spain

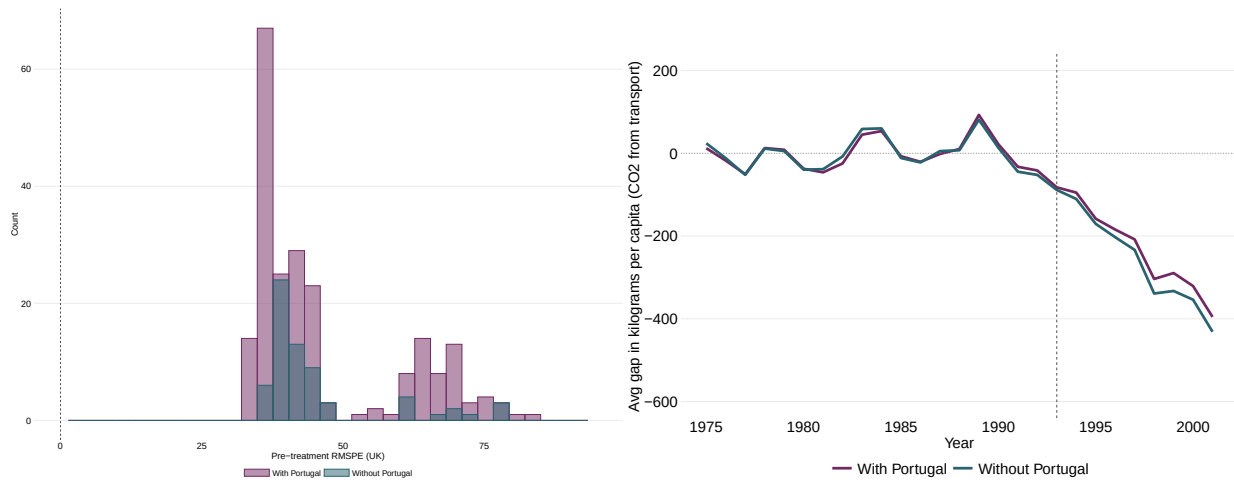


(a) Distribution of pre-treatment RMSPEs



(b) Average emission gap across all samples

Figure H6: Samples with and without Portugal



(a) Distribution of pre-treatment RMSPEs

(b) Average emission gap across all samples

Appendix I Subsample analysis: results with New Zealand

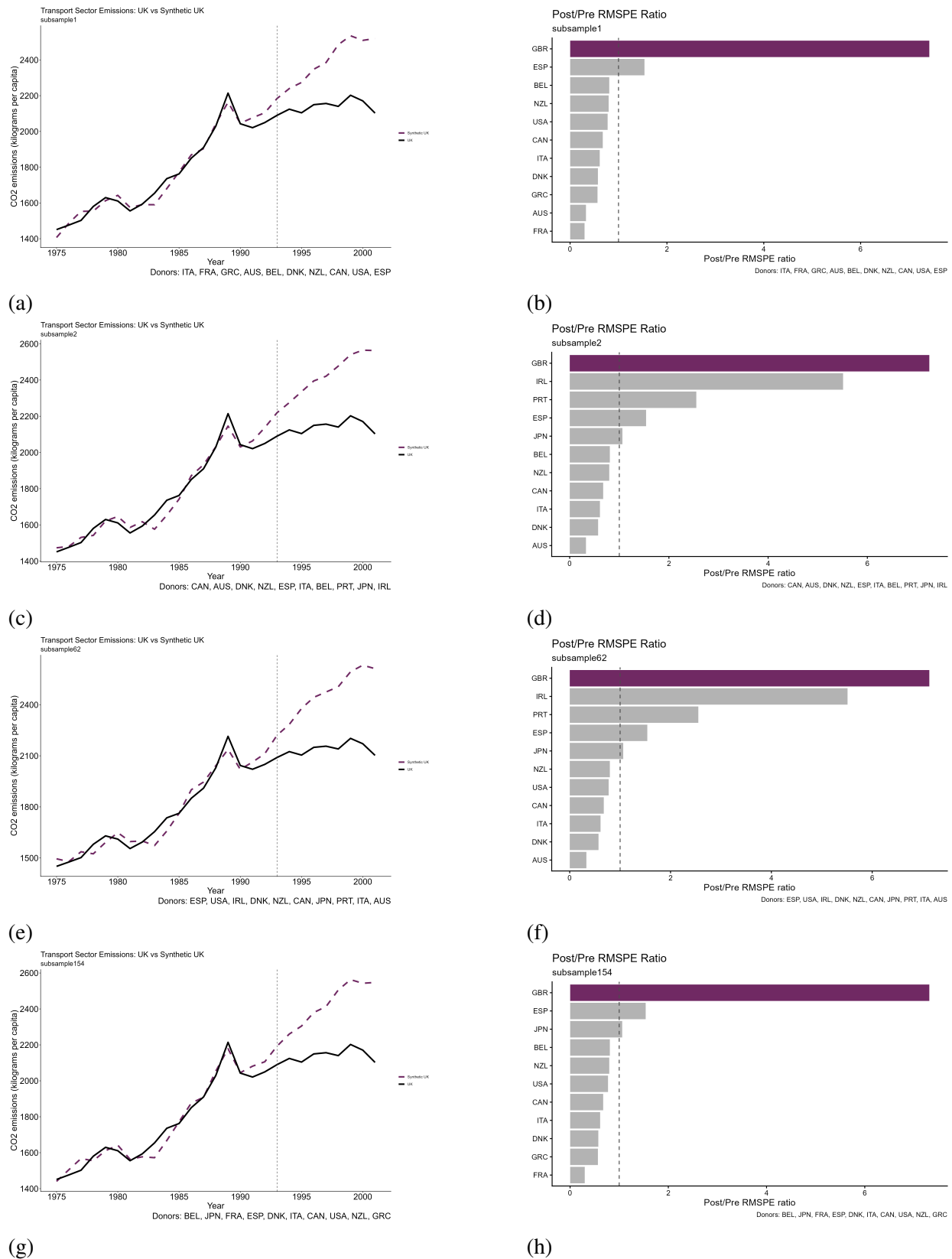


Figure I1: Four illustrative synthetic control estimates drawn from subsamples that include New Zealand. Figures (a), (c), (e), and (g) show the path plots of the UK versus synthetic UK. Figures (b), (d), (f), and (h) show the corresponding post/pre-RMSPE ratios; the vertical line marks a ratio of 1, where post-treatment and pre-treatment RMSPE are equal.

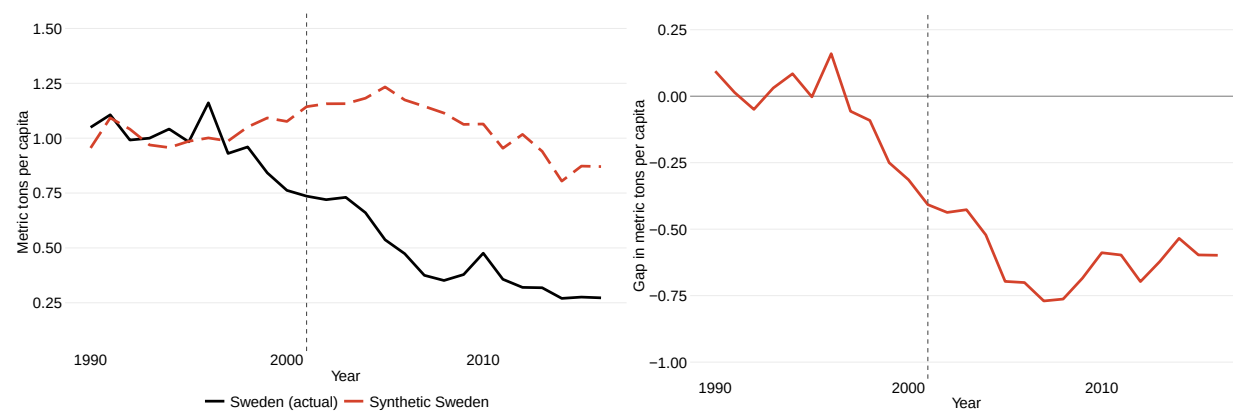
Appendix J Subsample analysis: results without New Zealand



Figure J1: Four illustrative synthetic control estimates drawn from subsamples that exclude New Zealand. Figures (a), (c), (e), and (g) show the path plots of the UK versus synthetic UK. Figures (b), (d), (f), and (h) show the corresponding post/pre-RMSPE ratios; the vertical line marks a ratio of 1, where post-treatment and pre-treatment RMSPE are equal.

Appendix K Replication results of Runst and Thonipara (2020)

Figure K1: Replication results



(a) Per-capita CO₂ emission trend plot

(b) Per-capita CO₂ emission gap plot

Appendix L Trimming in a small donor pool

I adopt the grouped factor model of Abadie and Vives-i-Bastida (2025), parameterized to reflect the small-donor-pool setting typical of country-level policy evaluations:

$$Y_{jt}^N = \delta_t + \lambda_{f(j),t} + \varepsilon_{jt}, \quad (7)$$

where δ_t is a common time effect, $\lambda_{f(j),t}$ a factor specific to group $f(j) \in \{1, \dots, K\}$, and $\varepsilon_{jt} \sim \mathcal{N}(0, \sigma^2)$ an idiosyncratic shock. Each factor follows an AR(1) process,

$$\lambda_{k,t} = \rho \lambda_{k,t-1} + \eta_{kt}, \quad \eta_{kt} \stackrel{iid}{\sim} \mathcal{N}(0, 1). \quad (8)$$

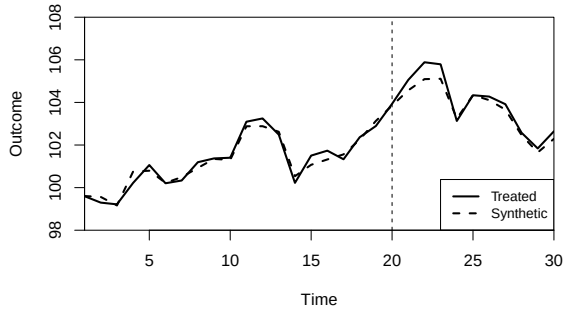
Units are organized into pairs sharing a common factor: $f(1) = f(2)$, $f(3) = f(4)$, and so on. Unit 2 is therefore the only donor sharing the treated unit's factor, and the ideal synthetic control assigns $W_2^* = 1$ and $W_j^* = 0$ for $j \geq 3$. The key departure from Abadie and Vives-i-Bastida (2025) is the size of the donor pool. Their trimming analysis uses $J = 199$; I set $K = 10$ and $J = 19$. The remaining parameters are those of Abadie and Vives-i-Bastida (2025): $T = 30$, $T_0 = 20$, $\rho = 1$, and $\sigma = 0.25$. The treatment effect is zero throughout, so any post-treatment divergence reflects estimation error.

Figure L1 reports a single realization of the DGP under four donor pools: $q = 1$ (panel a), $q = 0.75$ (panel b), $q = 0.50$ (panel c) and $q = 0.25$ (panel d). This is analogous to Figure 10 in Abadie and Vives-i-Bastida (2025), except that $J = 19$. Each panel overlays the treated unit's trajectory (solid) with its synthetic control (dashed). Neither the pre- nor the post-treatment fit improves visibly with trimming, and the post-treatment gap remains close to zero throughout, as expected given that no treatment effect is imposed.

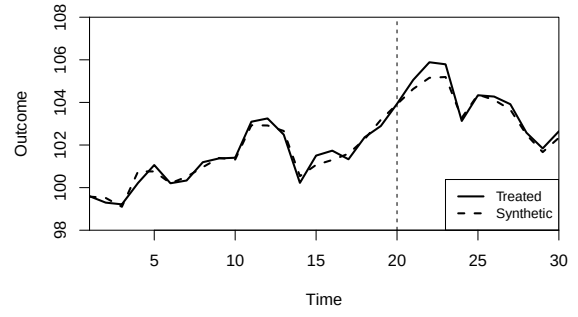
Table L1 shows the donor units, which are ranked by their Euclidean distance to the treated unit. It also shows the weights that they receive at each trim level. Unit 2, which shares the same factor structure as the treated unit and has the smallest distance, absorbs between 86% and 90% of the weight at every trim level. This is why trimming yields no visible gain here: the correct donor was already receiving almost all of the weight.

To verify that the pattern in Figure L1 is systematic rather than an artifact of a single draw, I run 10,000 Monte Carlo replications at each trim level under the same parameterization (Table L2). Unit 2 again attracts approximately 84% of the weight on average, falling to 79% only in the smallest pool ($q = 0.25$). Post-treatment bias is negligible throughout. Trimming therefore brings no improvement in this design.

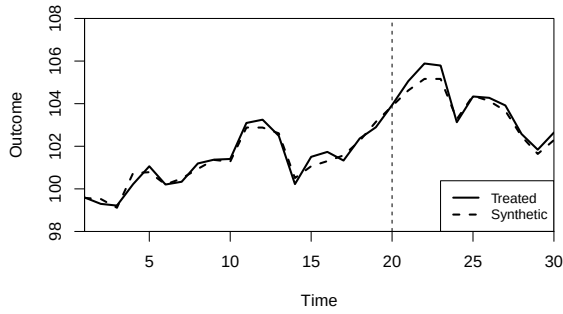
Figure L1: Synthetic control with trimming



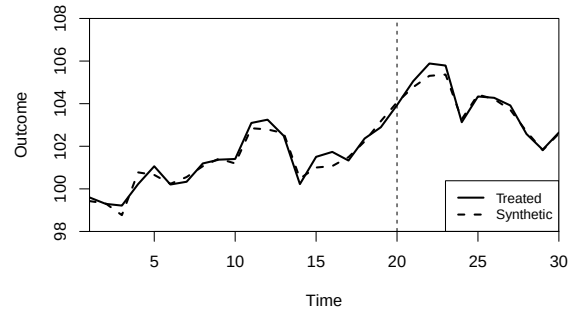
(a) No trimming ($q = 1.00$, 19 donors)



(b) 25% trimming ($q = 0.75$, 14 donors)



(c) 50% trimming ($q = 0.50$, 9 donors)



(d) 75% trimming ($q = 0.25$, 4 donors)

Note: In each graph the solid line is the treated unit and the dashed line is the synthetic control. DGP parameters: $K = 10$, $J = 19$, $T_0 = 20$, $T = 30$, $\rho = 1$, $\sigma = 0.25$. No treatment effect is imposed.

Table L1: Donor distances and synthetic control weights under trimming

Rank	Donor	Distance	w (no trim)	w (75%)	w (50%)	w (25%)
1	Unit 2	1.5527	0.860	0.877	0.873	0.895
2	Unit 3	6.1961	0.000	0.001	0.000	0.005
3	Unit 20	6.2597	0.000	0.001	0.001	0.074
4	Unit 4	6.3999	0.000	0.001	0.000	0.026
5	Unit 19	6.4722	0.000	0.001	0.000	–
6	Unit 17	8.0319	0.001	0.002	0.002	–
7	Unit 18	8.2199	0.083	0.062	0.076	–
8	Unit 8	9.2753	0.000	0.000	0.000	–
9	Unit 5	9.3781	0.035	0.027	0.047	–
10	Unit 7	9.4008	0.000	0.000	–	–
11	Unit 6	9.4919	0.019	0.028	–	–
12	Unit 11	14.8594	0.000	0.001	–	–
13	Unit 12	15.2959	0.000	0.000	–	–
14	Unit 10	15.7711	0.000	0.000	–	–
15	Unit 9	16.0434	0.000	–	–	–
16	Unit 14	16.6509	0.000	–	–	–
17	Unit 16	16.8994	0.000	–	–	–
18	Unit 15	16.9115	0.000	–	–	–
19	Unit 13	17.1037	0.000	–	–	–

Notes: The table reports the result of a single realization of the DGP. The column *Distance* is the Euclidean distance between the donor's and the treated unit's pre-treatment outcome paths, and the rows are sorted by it. The percentage in each w column heading gives the share of donors retained; horizontal rules mark the resulting cut-offs. Dashes indicate donors excluded at that trim level.

Table L2: Synthetic control fit by size of the retained donor pool

Retained (q)	Donors	Mean w_2	Pre-RMSE	Post-RMSE	Post-bias
0.25	4	0.7899	0.4460	0.7715	–0.0010
0.50	9	0.8409	0.3855	0.6091	–0.0077
0.75	14	0.8390	0.3789	0.6081	–0.0017
1.00	19	0.8361	0.3858	0.6283	0.0006

Notes: Averages across 10,000 replications at each trim level. The column *Retained* (q) is the share of donors retained; *Donors* is the resulting pool size; *Mean w_2* is the average weight on unit 2; *Pre-RMSE* and *Post-RMSE* are root mean squared gaps over periods $1-T_0$ and T_0+1-T ; *Post-bias* is the mean post-treatment gap.