COMBINING PROCESS TRACING AND SYNTHETIC CONTROL METHOD: BRIDGING TWO WAYS FOR MAKING CAUSAL INFERENCE IN EVALUATION RESEARCH

Federico Podestà

IRVAPP (Research Institute for the Evaluation of Public Policies)
Bruno Kessler Foundation
Via S. Croce, 77
38122 Trento, Italy
+39 0461 312291
podesta@irvapp.it

Abstract

This paper discusses potential ways of triangulation between two methods of evaluation in single-case studies: the synthetic control (SC) method and the process tracing (PT) method. Both are designed to examine certain events/programs that take place in given cases but view these events/programs from different causal perspectives. Seeing an event/program as a cause, SC estimates its impact on one or more outcomes. Conversely, starting from a certain outcome, PT discloses the causes which generated it. Hence, one can start from the causal explanation reached via one of the two methods and then proceed to examine that explanation through the other method. Once the causes of an outcome are traced via a PT analysis, that account can be validated by estimating the effects of those causes via SC. Equally, once the impact of a certain event is estimated through SC, causal mechanisms traceable via PT can be exploited to refine that impact evaluation.

Key words

Synthetic control method; Process tracing; Causal inference; Case study method.

Introduction

Evaluators often operate within a single-case framework. In these cases, the focus is on particular events (natural disasters, social and political conflicts, etc.) or large-scale interventions (labour market policies, welfare state programmes, etc.) that take place at an aggregate level and affect single geographic or administrative areas (countries, regions, cities, etc.).

Among others, two methods can be adopted in this framework: the synthetic control (SC) method and the process tracing (PT) method. SC is a quantitative method. It fully adheres to the impact evaluation perspective based on the counterfactual model of causality. It was initially introduced by Abadie and Gardeazabal (2003) and subsequently developed by Abadie et al. (2010; 2015) and by numerous econometricians (for an overview, see e.g., Abadie 2019). Conversely, PT is a qualitative method designed to craft suitable explanations of specific outcomes. Referring to the historian's methodology of explanation, George (1979) first proposed using PT in the social sciences. PT was then extensively developed by several scholars (e.g., George and Bennett 2005; Hall 2006; Collier 2011; Mahoney 2012) and, more recently, introduced into evaluation scholarship (e.g., Befani and Mayne 2014; Schmitt and Beach 2015; Befani and Stedman-Bryce 2017; Busetti and Dente 2017).

This paper discusses potential ways of triangulation between SC and PT. Using SC and PT in conjunction can be an important development for single-case evaluation. The two methods follow alternative causal perspectives, allowing researchers to make and combine different claims about a certain relationship between a cause X and an effect Y. This is precisely what several evaluation scholars are currently pursuing. More precisely, although it is common to adopt the counterfactual model to estimate the impact of a cause X on a certain outcome Y, there are other causal approaches to be considered when one is engaged in evaluation (Befani 2012; Mayne 2012).

As a counterfactual method, SC is inherently designed to answer the question: "What is the impact of the policy/event X that takes place in case Z?". For instance, SC's inventors employed it to answer the following research questions: (1) What is the economic impact of the terrorist conflict in the

Basque Country (Abadie and Gardeazabal 2003)? (2) What was the effect of Proposition 99, a large-scale tobacco control program implemented by California in 1988 (Abadie et al 2010)? (3) What was the economic impact of German reunification in 1990 (Abadie et al 2015)? Considering this last example, the impact of German reunification is estimated by comparing the post-1990 values of per capita GDP in the presence of the reunification (factual case) with the post-1990 values of per capita GDP in the absence of the reunification (counterfactual case). Since the values of per capita GDP for the counterfactual case are de facto missing, SC is precisely adopted to estimate them.

On the other hand, PT is grounded in generative causality built around a scientific-realist understanding of explanation. It is used to trace causal processes and discern intervening mechanisms underlying the occurrence of a specific event Y which took place in a single case. By starting with a particular outcome, PT moves backward allowing one to discern the causal chain that generated it (Mahoney 2012). Accordingly, PT serves to answer the question: "Which causes determined the occurrence of the event/policy Y in case Z?". Therefore, it was employed to answer the following research questions: (1) Which conditions triggered social revolutions in France, China and Russia (Skocpol 1979)? (2) What led to the US invasion of Iraq in 2003 (Harvey 2012)? (3) Which factors explain the status quo of UK and US welfare states in the 1980s (Pierson 1994)?

These few sentences reveal that a typical divide between the quantitative and qualitative approaches emerges when contrasting SC and PT: the distinction between causes of effects and effects of causes (Mahoney and Goertz 2006). Estimating the impact of certain events/policies, SC is an effects-of-causes-oriented method. Conversely, tracing the causal process of a certain event, PT is a causes-of-effects oriented method.

These different causal perspectives may raise some doubts about how fruitful a dialogue between SC and PT could be. If PT is a causes-of-effects-oriented method, how can it contribute to answering questions such as "What is the impact of that event/policy X?". Analogously, if SC is an effects-of-causes-oriented method, how can it help us answer "Why did this event/policy Y take place?". This paper intends to dispel these doubts and show that a fertile triangulation between SC and PT can

precisely be achieved referring to their different causal perspectives. One can start from the causal explanation provided by one of the two methods and then assess how and under which circumstances the other method can, from its standpoint, contribute to that causal explanation.

If PT starts from the outcome Y to conclude that it is generated by the cause X through the mediation of the mechanism M, SC can be then adopted to counterfactually test the effect exercised by X and/or M on Y. More precisely, if SC is applicable to estimate the impact of X and/or M, the occurrence of Y can be evaluated in relation to the presence/absence of X and/or M.

On the other hand, if SC starts from the event/policy X to estimate its impact on the outcome Y, PT can be employed to unfold the mechanism M through which Y arises. In other words, PT can complement SC by opening up the black box of the causal relation. This might appear to be not so dissimilar to the traditional role of PT in supplementing the correlational approach of large-N comparison (Bennett and Checkel 2015: 9). However, in this case, the contribution concerns causal relations pertaining to individual cases precisely examined *via* SC.

These ways of triangulation are illustrated through the paper without the pretense of developing a systematic guide for combining the two methods. Instead, some leading applications of SC and PT are simply considered to highlight how these methods can (or cannot) be pursued. Moreover, since PT has been only recently introduced in evaluation scholarship, it was necessary to select a set of applications belonging to other scholarships. However, a consideration of these applications remains illustrative for evaluation research.

The rest of the paper is structured as follows. In the next two sections, the basic elements of the two methods under examination are outlined in order to familiarize the reader with them. The fourth and fifth sections discuss how PT and SC can complement each other respectively. The final section concludes.

The SC method in a nutshell

SC is used in a between-case framework. In order to answer causal questions about the effects of a certain event/policy, it prescribes comparing the case exposed to the event/policy of interest, usually called the "treated unit", with an appropriately constructed comparison unit, called the "synthetic control unit". The construction of that unit is done in accordance with Mill's method of difference and by improving on traditional procedures used in comparative research for selecting control units. As is well known, Mill's method of difference implies selecting a control unit with the same characteristics as the case of interest with the exception of the phenomenon under scrutiny. When these conditions are satisfied, the comparison is "controlled". Nevertheless, selecting one real case which meets these conditions is often a heroic task. Consequently, students who follow traditional comparative strategies are frequently criticised for the arbitrariness of their selection. Hence, SC addresses these shortcomings by constructing the synthetic control unit through a data-driven control-group selection procedure (Abadie et al. 2015: 496-7).

That procedure is *de facto* based on the assumption that the characteristics of the treated unit can generally be much more accurately approximated by a combination of comparison units than by any single control unit. Accordingly, the synthetic control unit corresponds to a weighted average of some control units belonging to the so-called donor pool, i.e., a reservoir of potential comparison units. Specifically, the weights assigned to each control unit are chosen so that (1) the distance between the values for the treated unit and the synthetic one prior to the event/program under scrutiny are minimized with respect to a set of (observed) predictors and (2) the outcome trajectories for the actual and the synthetic unit best resemble each other in the pre-treatment period (i.e. prior to the occurrence of the intervention of interest). Therefore, the resulting synthetic control unit can best reproduce the characteristics of the actual case in the pre-treatment period.

In this way, SC not only facilitates comparative case studies based on Mill's method of difference, but also permits the reproduction of suitable counterfactuals: the evolution of the outcome variable for the synthetic control unit in the post-treatment period coincides with what would have occurred in the absence of the intervention under examination[1].

In view of this, the estimation of the effect of the intervention under examination is computed as the difference between the values of the outcome variable of the treated unit and the values of the outcome variable of the synthetic control unit in the treatment period, i.e. since the occurrence of the event/policy under scrutiny (Abadie et al. 2010; 2015)[2].

The logic behind the SC method can be better clarified by reference to its original application by Abadie and Gardeazabal (2003) in estimating the economic impact of the terrorist conflict in the Basque Country. In that study, the treated unit was the Basque Country since 1968, when ETA claimed its first victim.

The outcome was measured *via* real per capita GDP. Accordingly, Y1 refers to the values of per capita GDP in the presence the terrorist conflict, and Y0 denotes the values of per capita GDP in the absence of that treatment. Hence, the yearly gap between Y1 and Y0 corresponds to the impact of the terrorism on per capita GDP. Nevertheless, since the yearly trajectory of Y0 is not obviously observable, it had to be reproduced.

To this end, an annual regional-level panel data for the period 1955-1997 was employed so that the pre-treatment period lasted from 1955 to 1967, while the treatment period lasted from 1968 to 1997. The dataset included the Basque country plus the other Spanish regions that constituted the set of potential comparison units. For the pre-terrorism characteristics, a standard set of economic growth predictors was employed. It included sectoral share, human capital, investment ratio, and population density.

The data-driven procedure mentioned above resulted in positive weights for two regions, Catalonia and Madrid, with the values of 0.85 and 0.15 respectively, and the value of zero for the other potential controls. Accordingly, a synthetic Basque Country was constructed as a weighted average of

Catalonia and Madrid, which best reproduced values of economic-growth predictors and per-capita GDP of the Basque Country (Y0). In particular, the per-capita GDP trajectory for the synthetic unit almost overlaps with that of the actual unit in the 1955-1967 period (see Figure 1). Consequently, the synthetic Basque Country provided a reasonable approximation of the per-capita GDP that would have been experienced by that region over the treatment period in the absence of terrorism. Hence, the Y1 and Y0 trajectories were compared over the 1968-1997 period to estimate the effect of terrorism. In doing this, Abadie and Gardeazabal (2003) found a pronounced negative effect. As Figure 1 shows, per capita GDP of the two units trended similarly until 1975. However, from that year, when ETA's terrorist activity became a large-scale phenomenon, the two outcome lines diverged progressively and the synthetic one increasingly descended below the treated one ((Y1<Y0) denoting precisely the negative impact of terrorism.

Figure 1 – Per capita GDP for the Basque country and the synthetic Basque country (Abadie and Gardeazabal 2003).

Figure 1 about here

As anticipated, the SC method proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010; 2015) was progressively developed by numerous econometricians in order to overcome its drawbacks. In line with the aim of this article, one of these drawbacks and the corresponding remedy must be mentioned. The original version of SC suffers from a classical shortcoming often attributed to the impact evaluation perspective: it examines a mere link between the treatment and the outcome variable. In other words, the standard version of the SC method provides an estimation of the total effect exercised by the event/policy of interest, but does not make it possible to perform any causal mediation analysis. In the impact evaluation perspective, mediation analysis is a statistical framework which precisely permits us to take into account the fact that the impact exercised by the treatment variable on the outcome variable is mediated by an intervening variable, called the mediator (M) (e.g. Keele et al. 2015).

For this reason, Mellace and Pasquini (2019) proposed the mediation-analysis-synthetic-control (MASC) method. It is *de facto* a generalization of the standard SC method and makes it possible to break down the total effect of a certain event/policy into its direct and indirect components referring to a certain mediator. Specifically, through the MASC method the total effect can be broken down into the indirect effect, which goes through a mediator, and the residual effect. To compute these effects, three distinct outcome variables must be considered: Y00 representing the values of the outcome in the absence of both the treatment and the mediator, Y11 denoting the values of the outcome in the presence of both the treatment and the mediator, and Y01 representing the values of the outcome in the absence of the mediator but in the presence of the treatment. Hence, the total effect corresponds to the difference between Y11 and Y00, the residual effect to the difference between Y11 and Y01, and the indirect effect coincides with the difference between the total and the residual effect (for AN empirical application of MASC, see the fourth section).

The PT method in a nutshell

In tracing the causal chain of a certain outcome, PT is not only aimed at discerning the initial cause, but also the causal mechanism[3]. A causal mechanism is defined in this perspective as the causal process and intervening steps between an initial cause and a final outcome. Specifically, since it is assumed that an outcome is generated by a sequential chain which includes an intricate blend of causes, any causal explanation is reached by identifying the mechanisms through which the outcome of interest is generated (Bennett 2008a; George and Bennett 2005; Hall 2006). In other words, given that the notion of causal chain and underlying process imply that there is not too much proximity between the initial cause and the outcome of interest, one may rely on the intermediating mechanisms to make causal inference (Mayntz 2004).

In order to trace processes, researchers usually rely on particular information, called causal process observations (CPOs). A CPO can be defined as ‘‘[a]n insight or piece of data that provides information about context, process or mechanism, and that contributes distinctive leverage in causal inference” (Collier, Brady, and Seawright 2010). CPOs can thus be viewed as a piece of evidence resulting from some combination or accumulation of empirical observations and other contextual information about certain processes occurring within a single case. In other words, CPOs reflect in-depth knowledge of sequential processes rather than data collected as part of a systematized array of variables. CPOs may be either qualitative or quantitative in nature. However, each piece of evidence is qualitatively different so that they remain reciprocally non-comparable (Gerring 2007).

Furthermore, causal inference is made *via* within-case analysis. In other words, PT is a within-case method irrespective of the number of cases considered. If a single case shows a strong positive result with respect to the outcome of interest (e.g. the occurrence of a natural disaster) this is selected and then individually examined. This is precisely because, with PT, causation is not established through cross-case comparison, but through uncovering traces which connect events over time within a context of a single case. Hence, even if more cases are contrasted to assess the relevance of certain factors in determining similar or different outcomes, PT implies that their processes are traced singularly and separately to explain their specific case-outcomes (Bennett and Elman 2006; Tarrow 2010).

A widely-discussed example of this empirical strategy is Skocpol’s book State and Social Revolutions (1979). As Mahoney (1999: 1164) observed, in explaining the causes of revolution in China, Russia and France, Skocpol not only adopted cross-case comparison, but also analysed those national outcomes as the product of unique, temporally ordered, and sequentially unfolding events that occur within cases. Therefore, she de facto adopted the PT approach. By reference to pertinent CPOs, complex causal chains for each revolutionary transformation in the three countries examined were elucidated in the book.

In the case of France, Skocpol detected three main causal factors underlying the breakdown of state authority in the eighteenth century: agrarian backwardness, international pressure, and state autonomy. These factors, in turn, were disaggregated into thirty-seven discrete steps that connect structural causes to the outcome of interest. It is superfluous to present a full summary of the overall causal process connecting each event. However, the combination of property rights relations that prevented agricultural innovations (event 1), a tax system that discouraged agricultural innovation (event 2), and sustained growth that discouraged agricultural innovation (event 3) conjointly caused agricultural backwardness (event 4). The latter was the cause of the weak domestic market for industrial goods (event 5) which in turn produced the failure to achieve industrial breakthrough (event 8), and so on (for more details, see Mahoney (1999)).

Like Skocpol (1979), numerous scholars aim to present an explanation of the entire causal chain, addressing and shedding light on each individual step. This is especially the case when PT is simply used to explain the occurrence of a certain outcome in a single case and not to test or build a hypothesis. On the other hand, when it is applied to a single case to test a more general hypothesis the focus is not on an articulated causal chain, but on only a few elements. More precisely, when the existing literature indicates an association between phenomenon X (e.g., democracy) and phenomenon Y (e.g., peace), but there is uncertainty about the existence of a causal mechanism (e.g., political transparency) linking the two phenomena, PT can be adopted to evaluate whether that mechanism actually links X and Y in a particular case (George and Bennett 2005; Blatter and Haverland 2012; Bennett and Checkel 2015; Beach and Pedersen 2013).

This approach is also adopted in evaluation scholarship. As Schmitt and Beach (2015) observe, when one adopts PT for case evaluation study, the focus should be exclusively on the link between a single cause X and a particular outcome Y. To keep the analysis manageable, it is best to trace each of the distinct mechanisms separately from each other, unless their effects are significantly intertwined. Indeed, in their PT evaluation, these authors focus their attention more narrowly on two elements: the

governance effectiveness of budget support inputs, and direct outputs at the level of induced outputs, neglecting other interrelated causal links.

In this vein, Mahoney (2012) proposed a more structured procedure to use PT to test a causal hypothesis. He argued that the analyst should first combine pre-existing generalizations with specific observations (i.e. CPOs) from a single case and then evaluate the validity of the hypothesis *via* two tests commonly used in PT scholarship: the hoop test and the smoking gun test. A hoop test prescribes that a certain CPO must be present for a hypothesis to be valid. Failing a hoop test rejects a hypothesis but passing a hoop test does not corroborate a hypothesis. Conversely, smoking gun tests prescribes that if a certain CPO is present, then the hypothesis must be valid. Passing a smoking gun test lends decisive support in favour of a hypothesis, though failing a smoking gun test does not eliminate a hypothesis (Van Evera 1997; Bennett 2008a; Bennett 2010). Since a mechanism M is normally more closely situated to X and/or Y, Mahoney (2012) recommended using these two tests referring to CPOs about the relation between X and M or M and Y, rather than pieces of evidence about the direct link between X and Y.[4]

Complementing SC analysis *via* the PT method

This section aims to elucidate how PT can be used to complement an impact evaluation obtained *via* the SC method. To this end, two applications of the latter method are contrasted. They are the study on the economic impact of German re-unification (Abadie et al. 2015) and the already discussed analysis of the economic impact of the terrorist conflict in the Basque Country (see section 2). They are chosen because they make clear a remarkable difference regarding the supplementary use of causes-of-effects explanation and, thus, they precisely demonstrate how the adoption of that explanation can complement a SC study.

The study of the economic impact of German reunification was included in a methodological paper essentially aimed at discussing the use of the SC method in comparative research. Consequently, the authors introduced that SC analysis as a mere application of the more general SC method. The estimation of impact exercised by German reunification on the selected outcome, i.e. per capita GDP for West Germany was thus illustrated simply to describe the steps one must take to implement the SC method for a comparative case study. No information about the mechanisms through which the policy intervention of interest exercised its impact was provided. Only some previous studies about the economic costs of the reunification for West Germany were mentioned. As a result, the black box of the causal relation of interest was left closed.

On the other hand, the study of the economic impact of terrorism in Basque Country was included in a substantive article. Although SC was applied for the first time, Abadie and Gardeazabal (2003) remained mainly engaged with their research topic. In performing this case evaluation study, these authors, quite surprisingly, took several steps markedly consistent with the recommendations regarding using PT as a method for testing hypotheses (see section 3). First, they argued that the terrorism in the Basque Country was used as a case study to investigate the more general relationship between political conflict and economic growth. Hence, Abadie and Gardeazabal combined pre-existing generalizations with specific observations of the case of interest. Concerning pre-existing generalizations, they referred to the literature about the association between political stability/conflict and economic performances. Second, and more importantly, in addition to the SC estimation of the effects of the terrorism activity, they provided a qualitative causes-of-effects account, i.e., a brief history of the Basque Country over the treatment period of 1968-1997 (see Section 2). In providing that account, these authors took a further important step in line with the PT scholarship: making causal inference providing evidence about intervening mechanisms regarding the relationship between an initial cause and a subsequent outcome. Considering that, in this case study, terrorism represents the initial cause X and economic growth the subsequent outcome Y, Abadie and Gardeazabal presented some relevant pieces of evidence about the intermediating mechanism M. Specifically, they noted

that, in order to finance its operations, ETA used kidnappings-for-ransom, extortion, and, less frequently, robberies. The main targets of such money-raising activities were Basque entrepreneurs, who consequently started to abandon the Basque Country. In addition, these authors cited some literature as evidence that the terrorist conflict deterred domestic and foreign direct investment in the Basque Country. Therefore, the economic downturn caused by terrorist conflict would be mediated by a progressive reduction of investment, in turn due to pressures (directly or indirectly) suffered by local, national or foreign entrepreneurs.

Comparing these two studies, what can one conclude about the use of PT to supplement a SC analysis? If one refers essentially to the reliability of SC results, a complementary causes-of-effect explanation will appear superfluous. The total effect exercised by German reunification is *per se* reliable. In this sense, the SC method can be regarded as self-sufficient. Therefore, one might wonder what contribution the causes-of-effect account about the Basque Country may offer to the corresponding SC analysis. The causal inference reached *via* that account may appear as the standard contribution of PT to a more general association between two events X and Y (see section 3). The causal-mechanism unfolded for the Basque case serves to demonstrate how political conflict and economic development are linked in a single case. Nonetheless, the SC analysis performed on this region plays quite a similar role with respect to the general association between political conflict and economic development. It provided evidence of the negative economic impact of the terrorist conflict. In other words, as single-case studies, both SC and PT – and not only PT – shed light on a more wide-ranging association. PT contributes with an effects-of-causes account, while SC contributes a causes-of-effects account. However, one should at this point wonder how the two methods can speak to each other to further develop the single-case study. The causal mechanism unfolded by Abadie and Gardeazabal (2003) *de facto* opens the door to further SC analysis. In particular, since terrorist tensions would have progressively reduced (local and foreign) investment (see above), a MASC analysis (see section 2) can be performed to estimate to what extent the impact exercised by terrorism on per capita GDP was mediated by that intervening variable.

Figure 2 precisely shows that the total and negative impact that arose since the mid-1970s (see section 2) was essentially due to the more pronounced negative indirect effect which went through investment. Nevertheless, the figure also reveals quite a surprising positive residual effect which in part compensates the detrimental impact attributable to the indirect effect. Clearly, this residual effect could work through other, possibly unobserved, causal pathways. Their identification goes beyond the aim of this paper but indicates that a further causes-of-effects analysis might be performed to precisely discern those causal pathways.

Figure 2 – Total, indirect and residual effect of terrorism in the Basque Country.
Figure 2 about here

This triangulation between PT and MASC highlights another relevant issue. Although MASC allows an assessment of the extent to which the impact of a treatment on an outcome is mediated through another variable, it remains an effects-of-causes-oriented method. In other words, even if mediation analysis is often viewed as a framework for examining causal mechanisms, that examination does not per se permit to discern 'how' the effect is generated. PT remains indispensable for this purpose.
Furthermore, if PT analysis can supplement the SC method in identifying causal mechanisms, that identification can be strengthened by providing causes-of-effect accounts for the comparison units of the case of interest too. In other words, one may perform what Tarrow (2010) calls 'dual-process tracing' in order to trace and then contrast causal mechanisms of treated and comparison units. This would, moreover, make PT more compatible with SC logic. In fact, although processes would be traced for these units in accordance with within-case logic, a between-case analysis would be also carried out for them. Clearly, the selection of comparison units is, in this case, ensured by the transparency of the SC method – i.e. the explicit contribution of each potential comparison unit in constructing the synthetic unit. As Abadie et al (2015: 508) themselves note, by explicitly specifying the set of units that are used for comparison, the SC method does not preclude but facilitates detailed qualitative analysis and comparison between the case of interest and the set of comparison units selected by the SC method.

Accordingly, since the synthetic Basque Country was constructed as a weighted average of Catalonia and Madrid (see section 2), the socio-economic processes that occurred in these two regions during 1968-1997 period (i.e. the treatment period for Basque Country) may be traced and then contrasted with the corresponding processes for Basque Country. Specifically, given that these two regions did not experience terrorism in that period, this could make it possible to evaluate whether they denoted opposite causal mechanisms from those observed in Basque Country. More precisely, one could evaluate whether the relationship between political stability and economic growth observed in those regions was mediated by a significant permanence of local entrepreneurs as well as relevant domestic and foreign investments – i.e., the contrary of what happened in the Basque Country. Moreover, an in-depth examination of these two regions might also help to discern the causal pathways underlying the above-mentioned residual effect estimated for the Basque Country.

Nonetheless, one must note that the study on the economic impact of the terrorist conflict in the Basque Country was extremely lucky in performing this kind of triangulation between SC and PT. Only two regions were involved in constructing the synthetic Basque and one of these, Catalonia, was given a very high weight (0.85). This means that this region alone quite adequately approximates the most relevant characteristics of Basque Country, excluding terrorism.

However, exploiting potential control units to perform further PT analyses is conditioned by their number and, more importantly, by the values of their weights. In fact, in several cases more control units are used to construct the synthetic unit and their weights are quite low, denoting limited similarity with the treated unit. For instance, in the analysis of the economic impact of German reunification, the synthetic West Germany was given by a weighted average of Austria (0.42), US (0.22), Japan (0.16), Switzerland (0.11) and the Netherlands (0.09) – with the weights in parentheses.

Complementing PT analysis with the SC method

This section shows how and under what conditions SC can be adopted to enhance PT analysis. Although a causal chain reconstructed through PT can be regarded as valid per se, some criticism may arise. One reason derives from the fact that PT makes causal inference *via* a mere tracing of actual process. Furthermore, even when a causal relation is assessed via the hoop test and the smoking gun test (see section 3) scepticism may persist. This is because these tests are far more subjective than standard statistical tests. They lack the precision and decisiveness of formal tests (Skarbek 2020).

Accordingly, several PT developers suggest counterfactual analyses can validate the causal inference made *via* PT. In fact, in counterfactual framework, the outcome is specified in case of both X and non-X (Kay and Baker 2015; Blatter and Haverland 2012; Collier 2011).

Using the SC method is clearly one way of accomplishing this. It makes it possible to construct synthetic units which precisely represent what would have occurred in the absence of the cause of interest (see section 2). Furthermore, since SC is inherently designed to estimate the impact of an event/policy that took place after a particular point of time, it appears suitable to the sequential logic underlying PT. Accordingly, SC can be employed to evaluate the occurrence of a certain outcome in relation to the presence/absence of a putative cause identified *via* PT. Specifically, if SC demonstrates that that cause is irrelevant for the occurrence of the outcome, PT analysis will have to be reviewed.

However, SC cannot be used to counterfactually test all causes-of-effects accounts. The first reason concerns limited data availability. Indeed, one may start a case evaluation study *via* PT, but the time series necessary to assemble the panel dataset required for performing the complementary SC analysis might simply be unavailable. A second constraint regards the complexity of the causal chain reconstructed for the case under examination. Specifically, a single SC analysis cannot enable us to assess an entire and articulated causal chain like that of Figure 2 (see section 3). Even if one adopts MASC instead of the standard SC, one can only assess causal relations between a few elements.

However, this is not a relevant limitation for evaluation research. As observed in Section 3, focus should be exclusively on the link between a single cause X and a particular outcome Y.

Moreover, outcome operationalization and some problems of comparability between treated and control units can involve other problems of SC applicability. To show this, a leading study of political science can be considered, i.e., Harvey's (2012) study of the 2003 Iraq War. Through PT analysis, this author provided a detailed tracing of all key decisions, choices and tactics adopted by George W. Bush during that period. The reconstructed sequence of decisions that led to war, i.e. the outcome of interest, is definitely complex. Therefore, in line with the argument developed before, SC applicability is hereafter assessed in testing the simpler and direct impact of George W. Bush's presidency (X) on the invasion of Iraq (Y). The corresponding counterfactual is: 'If George W. Bush had not been elected president, the United States would not have invaded Iraq'.

Since the SC method inherently works with continuous variables for measuring the outcome, it is impracticable when referring to a dichotomous variable denoting the occurrence vs. the non-occurrence of US invasion of Iraq. Nonetheless, proxying that event *via* quantitative indicators involves serious problems as well. In fact, an obvious solution would be using the number of US troops who arrived in Iraq. This would imply that the corresponding US time series would denote zero values for the pre-treatment period (e.g. from 1992 to the George W. Bush 2000 election) as well as for 2001-02 period and a value of about 130,000 for 2003. Using the SC method to reproduce that trajectory would be nonsensical. This is because, if one considers a set of OECD countries as potential control units, all the corresponding time series for the outcome variable would be entirely composed of zero values. Consequently, any weighted average of the available control units to best reproduce the outcome trajectories for the US would end up mathematically irrelevant precisely because it would produce the same result. However, beside any SC calculation, Graph 1 of Figure 3 reproduces the logic of this hypothetical counterfactual exercise. The counterfactual and US trajectory for the outcome variable would be overlapped until 2003. From that year, the two lines would diverge, but

the effect they denote would be trivial. In the absence of George W. Bush, none of the US troops would have arrived in Iraq, so the Bush impact would correspond to 130,000 troops (130,000 - 0). Serious problems would also arise if military expenditure as a percentage of GDP were employed to measure the outcome of interest. In this case, the US trajectory for the outcome variable would not be reproducible because, during the pre-treatment period, the US level of military expenditure was, on average, significantly higher than that of any potential control unit. Indeed, this problem could be addressed through the synthetic-difference-in-differences estimator recently developed by Arkhangelsky et al. (2021). Nonetheless, the Bush impact on the Iraq war specifically would be indiscernible. This is because, after a prolonged period of decrease, US military expenditure had already begun to increase again since 2001 as a consequence of the Afghanistan war (see Graph 2 of Figure 3).

Using SC to test the above-mentioned counterfactual would also imply a problem of unit comparability. The US, the treated unit, experienced the Bush presidency from 2000 onwards. Consequently, in order to know what would have occurred in its absence, the synthetic US should be constructed using a reasonable set of potential control units. Assuming that the negation of the actual antecedent, i.e., George W. Bush's election, is not equivalent to the election of his rival, Al Gore, a certain set of affluent democracies could be used as potential controls. If one assumes that no other affluent democracy experienced the Bush presidency, all these democracies can be included in the donor pool. On the other hand, if one assumes that the Bush presidency is comparable to right-wing governments of other developed countries, the countries that were exposed to those governments prior to 2000 must be excluded from the donor pool.

Figure 3 – SC analysis on US invasion of Iraq: US troops and military expenditure.
Figure 3 about here

That said, let us consider a causes-of-effects study counterfactually test *via* SC method. It is Pierson's (1994) analysis of Reagan and Thatcher's welfare state retrenchments, counterfactually replicated by

Podestà (2020). Pierson observed that at the end of the so-called conservative resurgence of the 1980s, when Thatcher and Reagan left power, the UK and US welfare states ended up substantially unaltered. That outcome was attributed to dense networks of interest groups developed around numerous welfare policies. Tracing and contrasting the decision-making of several social programs, Pierson showed that retrenchments were only successful for those programs with less organized recipients. In this sense, Reagan and Thatcher's plans to dismantle the welfare state can be regarded as the initial cause X, while the mobilization of organized recipients constitutes the intermediating mechanism of the process M. However, Pierson's analysis did not provide any evidence about the extent to which the retrenchments would have been implemented in the absence of influential groups. Analogously, we do not know whether the retrenchments would have been substantially the same or even smaller in the absence of conservative governments. Stressing the difficulty in testing the first counterfactual, Podestà (2020) performed a SC analysis to test the second one. Specifically, he argued that finding developed countries without organized supporters of the welfare state to be used as potential control units is *de facto* unfeasible. On the other hand – he claimed – one may reasonably assume that, during the 1980s, the Reagan and Thatcher administrations were the only examples of the conservative resurgence[5]. Accordingly, the US and the UK were regarded as the sole treated units, while a sample of developed countries was used as the set of potential controls.

Furthermore, since the SC method inherently works with continuous variables to measure the outcome, six quantitative indicators (i.e. social security transfers, welfare state generosity, unemployment-benefit generosity, pension generosity, total government revenue, and union density) were adopted to operationalise some of the policy areas examined by Pierson. Hence, the trajectories of those indicators, observed in the presence of a conservative administration, were contrasted with corresponding trajectories that were reconstructed in the absence of such a political circumstance.

The results of this replication broadly confirmed Pierson's substantive conclusion: the conservative revolution of the 1980s did not substantially alter the UK or US welfare state. More precisely, in many policy arenas, the two conservative governments did not provoke noticeable retrenchment.

Nevertheless, the estimated impacts of specific components of the two welfare states are, in some cases, different from the conclusions stressed in Pierson's book. For instance, while Pierson concluded that unemployment insurance benefits have undergone extensive retrenchment in both countries, Podestà's SC analysis revealed that if the UK and US had not undergone a conservative revolution, their unemployment generosity would have not been significantly different. These differing findings probably depend on the fact that (1) the causal chain was not entirely counterfactually tested and (2) the quantitative indicators used in the SC analysis differ from the qualitative information adopted by Pierson. In any case, these differences question some of Pierson's conclusions and require some reviews of his causal explanation. As argued above, if an SC analysis highlights a causal effect inconsistent with the PT explanation, the latter should be amended accordingly.

Conclusion

In order to develop the methodology for evaluation scholarship, this paper has discussed potential ways of triangulation between two methods currently central to single-case research: SC and PT. To put these two methods in dialogue with each other, the starting point has been a substantial divergence between them: the causes-of-effects versus effects-of-causes perspective. SC starts from an event/program that took place in a certain case and – assuming it to be a cause – estimates its impact on one or more outcomes. Conversely, PT starts from an outcome and discloses the causal chain which generated it.

Hence, a basic rule in triangulating the two methods is to start from the causal explanation reached *via* one of the two methods and then examine that explanation using the other method, i.e. looking at the causal relation from the opposite standpoint. Accordingly, once the causes of an outcome are

identified *via* a PT analysis, the latter can be complemented and validated via a SC analysis by estimating the effects of those putative causes. On the other hand, once the impact of a certain event on a given outcome is estimated through a SC analysis, supplementary information about causal mechanisms can be provided *via* PT and hence, the effects-of-causes analysis can be refined. These ways of triangulation mean that the analyst must not necessarily start with one method and then proceed with the other. One may start with an effects-of-causes via SC and then proceed with a causes-of-effects perspective via PT or vice versa. Therefore, if the combination of the two methods is successfully performed, the causal relation linking the phenomena of interest can be viewed from opposite standpoints. Clearly, in triangulating the two methods, some inconsistencies may arise in terms of findings. For example, as observed in the previous section, SC disconfirmed some results of the causes-of-effects study of Reagan and Thatcher's welfare state retrenchments. This requests a partial revision of that study. This means that an iterative process of triangulation can be necessary to smooth out certain substantive incongruences resulting from a preliminary dialogue between the two methods.

Nonetheless, some studies considered in the previous pages reveal that the combination of SC and PT cannot be systematically applied. This essentially depends on some constraints in implementing the SC method. Data unavailability, problems of comparability between treated and control units and difficulty of outcome operationalization may limit the adoption of SC to complement PT analysis. Therefore, the possible means of triangulation sketched in the previous pages depend on the phenomena one wants to examine.

Beyond this, a more extensive exchange between SC and PT depends on the stance of the developers of each of the two methods towards the other approach. SC scholars appear quite open to embracing PT. The causes-of-effects account on the Basque Country provided by Abadie and Gardeazabal (2003) constitutes a good example of this predisposition. Indications of this propensity are also signalled by the fact that researchers using any impact evaluation technique, in particular experimental approaches, are currently looking at causal mechanisms (Gerring 2010). On the other

hand, several PT students invoke counterfactual perspectives to complement their explanations. Nonetheless, they appear to be somewhat hesitant to employ SC. To the best of my knowledge, none of the leading PT developers have yet employed this kind of counterfactual strategy in social research, and more precisely in evaluation study.

Acknowledgments:

For helpful comments on previous versions of this paper, I thank Davide Azzolini, Martina Bazzoli, Sergiu Constantin Burlacu, Daniela Sicurelli, and Alessandro Tondini. Furthermore, I would like to thank Nicola Bazoli, Joan Madia, Giovanni Mellace and Alessandra Pasquini for their very helpful discussions on specific aspects of this article. Further thanks go to Alessandra Pasquini for her precious technical support. Finally, I am particularly grateful to the anonymous reviewers and the editorial team who provided very valuable comments on my article.

Endnotes

[1] As often happens in case studies, this is obtained following a merely hypothetical logic (Schmitt and Beach 2015). On the other hand, the above-mentioned procedure makes SC counterfactuals well-grounded.

[2] SC prescribes evaluating the uncertainty of the estimated effect *via* placebo techniques based on the principle of permutation inference (for more details, see Abadie et al. (2010, 2015)). For the sake of simplicity, this issue is ignored through the paper.

[3] In this sense, PT can be likened to theory-based evaluation (Pawson 2013). However, as Schmitt and Beach (2015) show, PT is more suitable for empirically tracing causal mechanisms that link interventions with outcomes.

[4] Several PT developers have incorporated Bayesian logic in PT. This is in order to provide a more rigorous framework for assessing the type and strength of conclusions we can make using different forms of empirical evidence (e.g., Rohlfing 2013; Befani and Stedman-Bryce 2017; Barrenechea and Mahoney 2019). This issue is ignored in the present paper because it is not essential for the development of its argument on causal inference.

[5] Pierson (1994: 4) himself admitted that the Thatcher and Reagan administrations constitute a crucial test of the welfare state's status because, for the first time since before World War II, two political executives were openly critical of central features of social policy.

References

Abadie A (2021) Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature* 59(2): 391–425.

Abadie A and Gardeazabal J (2003) The economic costs of conflict: A case study of the Basque Country. *American economic review* 93(1): 113–132.

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

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

Alexander G and Lauren PG (1979) Case studies and theory development: the method of structured, focused comparison. *Diplomacy: New approaches in history, theory, and policy*. Free Press.

Arkhangelsky D, Athey S, Hirshberg DA, et al. (2021) Synthetic difference-in-differences. *American Economic Review* 111(12): 4088–4118.

Barrenechea R and Mahoney J (2019) A set-theoretic approach to Bayesian process tracing. *Sociological Methods & Research* 48(3): 451–484.

Beach D and Pedersen R (2019) *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor, MI: University of Michigan Press.

Befani B (2012) Models of Causality and Causal Inference. In: Stern, E, Stame, N, Mayne, J, Forss, K, Davies, R, Befani, B *Broadening the range of designs and methods for impact evaluations,* DFID Working Paper no. 38. UK Department for International Development.

Befani B and Mayne J (2014) Process Tracing and Contribution Analysis: A combined approach to generative causal inference for impact evaluation. *IDS bulletin* 45(6): 17–36.

Befani B and Stedman-Bryce G (2017) Process tracing and Bayesian updating for impact evaluation. *Evaluation* 23(1): 42–60.

Bennett A (2010) Process tracing and causal inference. In: Brady H and Collier D. *Rethinking social inquiry: Diverse Tools, Shared Standards*. New York: Rowman & Littlefield.

Bennett A and Checkel J (2014a) Introduction: Process tracing: From philosophical roots to best practices. In: *Process tracing: From metaphor to analytic tool*. Cambridge: Cambridge University Press.

Bennett A and Checkel J (2014b) (eds) *Process tracing: from metaphor to analytic tool*. Cambridge: Cambridge University Press.

Bennett A and Elman C (2006) Qualitative research: Recent developments in case study methods. *Annual Review of Political Science* 9: 455–476.

Blatter J and Haverland M (2012) *Designing Case Studies: Explanatory Approaches in Small-N Research*. Houndmills: Palgrave Macmillan.

Busetti S and Dente B (2017) Using process tracing to evaluate unique events: The case of EXPO Milano 2015. *Evaluation* 23(3): 256–273.

Collier D (2011) Understanding Process Tracing. *PS: Political Science and Politics* 44(4): 823–830.

Collier D, Brady HE and Seawright J (2010) Sources of leverage in causal inference: Toward an alternative view of methodology. In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. New York: Rowman & Littlefield.

Gerring J (2006) *Case Study Research: Principles and Practices*. Cambridge: Cambridge University Press.

Gerring J (2010) Causal mechanisms: Yes, but…. *Comparative political studies* 43(11): 1499–1526.

Hall PA (2006) Systematic process analysis: when and how to use it. *European Management Review* 3(1): 24–31.

Harvey FP (2011) *Explaining the Iraq War: Counterfactual Theory, Logic and Evidence*. Cambridge: Cambridge University Press.

Kay A and Baker P (2015) What Can Causal Process Tracing Offer to Policy Studies? A Review of the Literature. *Policy Studies Journal* 43(1): 1–21.

Keele L, Tingley D and Yamamoto T (2015) Identifying Mechanisms behind Policy Interventions via Causal Mediation Analysis. *Journal of Policy Analysis and Management* 34(4): 937–963.

Mahoney J (1999) Nominal, Ordinal, and Narrative Appraisal in Macrocausal Analysis. *American Journal of Sociology* 104(4): 1154–96.

Mahoney J (2012) The Logic of Process Tracing Tests in the Social Sciences. *Sociological Methods & Research* 41(4): 570–597.

Mahoney J and Goertz G (2006) A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research. *Political Analysis* 14(3): 227–249.

Mayne J (2012) Making causal claims. *ILAC Brief*.

Mayntz R (2004) Mechanisms in the analysis of social macro-phenomena. *Philosophy of the social sciences* 34(2): 237–259.

Mellace G and Pasquini A (2019) *Identify More, Observe Less: Mediation Analysis Synthetic Control*. arXiv e-prints, arXiv:1909.12073.

Pawson R (2013) *The Science of Evaluation: A Realist Manifesto*. London: SAGE Publications.

Pierson P (1994) *Dismantling the Welfare State? Reagan, Thatcher and the Politics of Retrenchment*. Cambridge Studies in Comparative Politics. Cambridge: Cambridge University Press.

Podestà F (2020) Did the conservative resurgence matter? A counterfactual replication of Pierson's analysis on Reagan and Thatcher's welfare state retrenchments. *Comparative European Politics* 18(5): 819–839.

Rohlfing I (2014) Comparative Hypothesis Testing Via Process Tracing. *Sociological Methods & Research* 43(4): 606–642.

Schmitt J and Beach D (2015) The contribution of process tracing to theory-based evaluations of complex aid instruments. *Evaluation* 21(4): 429–447.

Skarbek D (2020) Qualitative research methods for institutional analysis. *Journal of Institutional Economics* 16(4): 409–422.

Skocpol T (1979) *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge: Cambridge University Press.

Tarrow S (2010) The strategy of paired comparison: toward a theory of practice. *Comparative political studies* 43(2): 230–259.

Van Evera S (1997) *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.

Figure 1 – Per capita GDP for the Basque country and the synthetic Basque country (Abadie and Gardeazabal 2003).

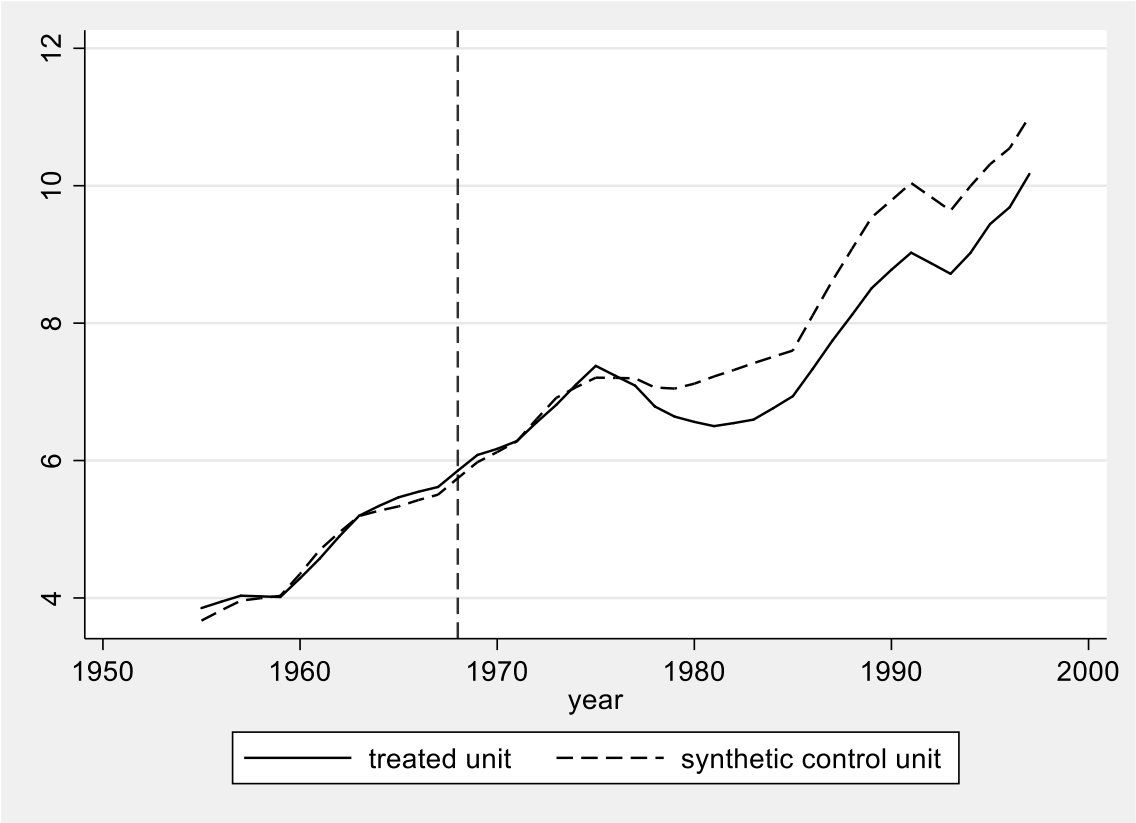

Figure 2 – Total, indirect and residual effect of terrorism in the Basque Country.

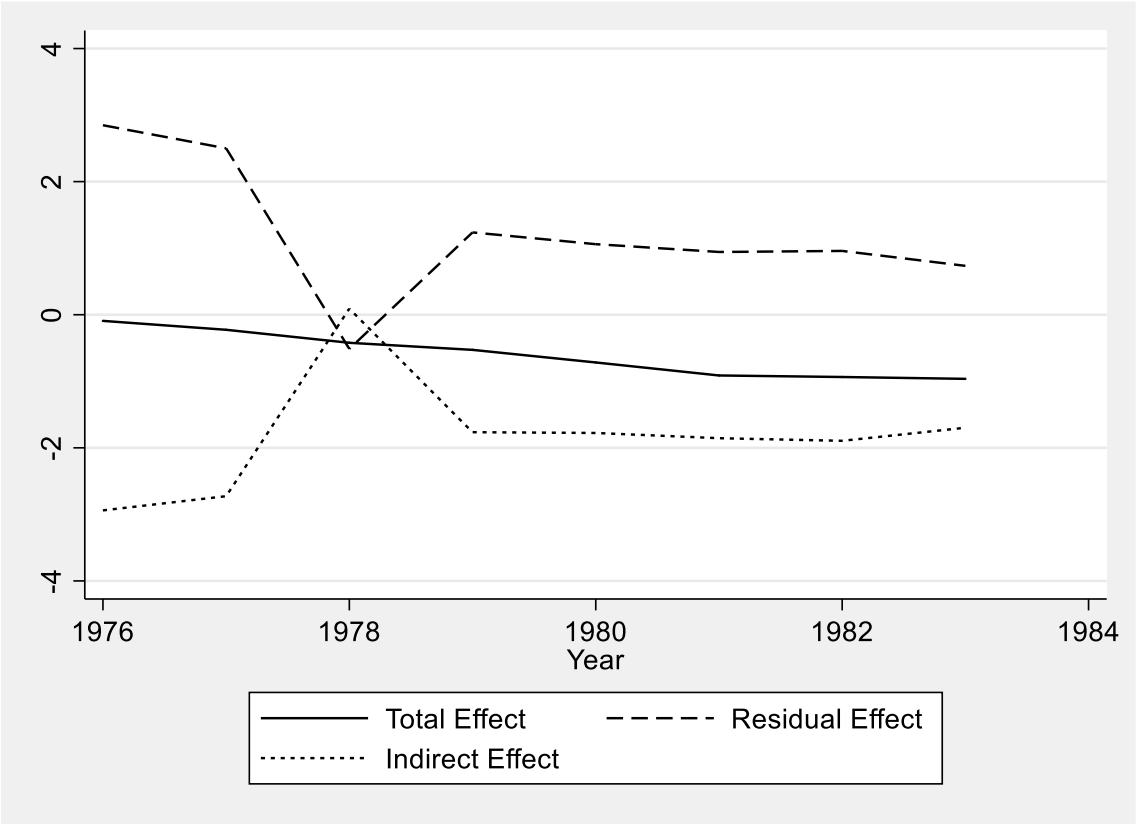

Figure 3 – Iraq war