Discovering What Mattered: Detecting Unknown Treatment as Breaks in Panel Models

Moritz Schwarz[1,2,3*]

[1]Climate Econometrics, Nuffield College, University of Oxford
[2]Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany
[3]Faculty of Economics and Management, Technische Universität Berlin, Berlin, Germany
[4]Department of Economics, University of Victoria

Job Market Paper

Click here for most up-to-date version.

Co-authored with Felix Pretis*[1,4]

Abstract

Effective policy design requires knowing which interventions have made a measurable difference. Yet researchers often observe that outcomes such as greenhouse gas emissions, productivity, or growth have changed without knowing which policies or shocks were responsible. Rather than evaluating the effects of a known policy, our approach starts from the outcome itself: we detect structural breaks in panel data and attribute them to possible causes. We develop an econometric framework that detects unknown treatment assignment and timing as structural breaks in fixed-effects panel models. We show that the estimation of heterogeneous treatment effects in a dummy-saturated Two-Way Fixed Effects (TWFE) model nests the standard staggered difference-in-differences design as a special case. We demonstrate that machine learning methods — such as indicator saturation or the adaptive LASSO — can recover the underlying treatment without prior knowledge. Crucially, by relying on established properties of indicator saturation, this approach allows for controlling the False Discovery Rate of "discovered" treatments. This provides a theoretical foundation for recent work in climate econometrics (e.g., Stechemesser et al., 2024; Koch et al., 2022, Pretis 2022) where the effectiveness of policy are often uncertain. We demonstrate the method's utility for climate policy evaluation by identifying the effectiveness of the Swedish carbon tax (Andersson, 2019) in a data-driven manner, and replicating Abadie and Gardeazabal (2003) to detect the economic impact of ETA terrorism. The methods are freely available in the open-source R package *getspanel*.

JEL Classification: C21, C23, C52
Keywords: Panel Data, Two-Way Fixed Effects, Treatment, Policy Evaluation, Difference-in-Differences; Break Detection, Indicator Saturation, Adaptive LASSO, Machine Learning;

*Author contact: fpretis@uvic.ca and moritz.schwarz@ouce.ox.ac.uk. We thank Anders Bredahl Kock, Sophia Carodenuto, Nicolas Koch, Andrew B Martinez, and Nolan Ritter for helpful comments and suggestions. MS gratefully acknowledges financial support from the Einstein Foundation. M.S. is employed by the Austrian Ministry of Finance. The views expressed here are those of the authors and not those of the Ministry of Finance or the Austrian government.

1 Introduction

Designing effective policies depends on understanding which interventions have actually driven measurable change. Governments and researchers routinely evaluate whether specific programs, taxes, or regulations achieved their intended effects, traditionally estimating the effects of known treatments, such as: does a carbon tax reduce emissions, or does terrorism affect GDP per capita? These designs place specific, known interventions at the centre of analysis and estimate their effects on outcomes of interest and are known as 'Forward Causal Questions' (Gelman 2011, Gelman & Imbens 2013, Mill 1843). While this approach has been highly successful, it presupposes that the relevant interventions are already known to the researcher or policymaker. As a result, it may miss policies, shocks, or behavioural shifts that have meaningful effects but are not known *a-priori*.

In practice, policy makers often face the problem that they observe that outcomes have changed but do not know why. Emissions may decline, productivity may decrease, or insurance losses may rise – yet the underlying cause remains unclear. The relevant question then becomes not whether a known policy caused a given effect, but what events or interventions produced the observed change. Identifying such unknown but impactful policies requires tools that can uncover the timing and location of interventions directly from the data rather than relying on prior knowledge.

This challenge is particularly acute in climate econometrics, where there are numerous possible policy interventions but there is uncertainty around which are effective. The number of possible interventions is regularly unmanageably large for a forward-causal analysis. Consequently, an econometric analysis conditional on a fixed, known treatment date may suffer from misspecification and fail to identify the true policy impact. A growing literature has thus turned to break detection to agnostically identify effective climate policies by searching for significant shifts in emission trajectories directly. For instance, Koch et al. (2022) use break detection to identify reductions in road-transport emissions, Stechemesser et al. (2024) apply these techniques to detect successful climate policies globally, and Pretis (2022) examines British Columbia's carbon tax. Further Tebecis & Crespo Cuaresma (2025) compiled a global dataset of structural breaks in greenhouse gas emissions, and Tebecis (2023) evaluated Austrian climate policies using a similar approach.

While these studies demonstrate the value of detecting unknown treatment episodes directly from data, the econometric foundations linking these applications to the broader program evaluation literature have not been formally established. We propose a formal econometric framework to bridge this gap. We conceptualize the detection of unknown treatment episodes as a variable selection problem in high-dimensional panel data. Our primary theoretical contribution is to establish the equivalence between structural break detection and the estimation of heterogeneous treatment effects. We demonstrate that a dummy-saturated TWFE model – augmented with appropriate penalization or selection criteria – nests the standard staggered difference-in-differences design (e.g., Wooldridge 2025, Goodman-Bacon 2021) as a special case where treatment timing is known.

This equivalence implies that the tools of high-dimensional statistics can be repurposed for causal discovery. By starting from a saturated model that includes potential step-shifts or impulses for every unit and time period, we utilize Indicator Saturation methods (Hendry et al. 2008) and variants of the LASSO (Zou 2006) to detect and estimate treatment. Unlike ad-hoc exploratory analysis, this approach is grounded in established statistical properties of the chosen selection algorithms, allowing us to control the false discovery rate. This transforms policy evaluation from a rigid testing of known hypotheses into a data-driven generation of new ones.

This approach operationalises what Gelman & Imbens (2013) call 'Reverse Causal Questions' or finding

the causes of effects. As they put it: "*Reverse causal reasoning is different; it involves asking questions and searching for new variables that might not yet even be in our model*". Our proposed approach extends the familiar two-way fixed effects (TWFE) estimator, commonly interpreted as a difference-in-differences design in a standard treatment setting, to situations where neither treatment assignment nor timing are known. We show that conventional difference-in-differences models are a special case of a more general setup in which treatment episodes appear as structural breaks in the fixed effects. By allowing the data to reveal which units experienced such shifts and when, the model can detect heterogeneous treatment effects as well as staggered treatment without requiring prior specification of the intervention.

We frame the detection of treatment interventions through structural breaks as a problem of variable selection. Starting from a saturated TWFE model that includes a full set of potential step-change functions representing possible treatments for each unit and time, we apply machine-learning and model-selection methods such as the adaptive LASSO or general-to-specific (GETS) algorithms to identify the relevant subset of breaks (similarly recent work extends this discovery process to Bayesian frameworks: Konrad et al. 2026 develop Bayesian algorithms designed for indicator saturation in high-dimensional settings). These detected shifts can then be interpreted as previously-unknown treatment events that produced the observed changes in the modelled outcome. Once identified, they generate hypotheses to be attributed to potential causes, for example policies or shocks, in a post-estimation analysis. This transforms policy evaluation from a process of testing known hypotheses into one of generating and testing new ones, providing a novel way to investigate the causes of detected effects.

Our approach provides an econometric foundation for a new class of hypothesis-generating tools that link structural-break detection to the modern literature on staggered and heterogeneous treatment effects (see e.g. Roth et al. 2023, Wooldridge 2025). In doing so, we bridge the gap between exploratory data-driven causal discovery and formal causal inference. By embedding break detection in a TWFE framework, we demonstrate that the same principles that underpin difference-in-differences estimation can be used to identify unknown interventions when the timing and assignment of treatment are unknown *a-priori*. This general framework also allows for a detection of heterogeneous treatments, including the detection of evolving treatment sizes through the use of broken linear trends.

Once structural breaks are detected, the next step is to attribute them to potential causes. This post-detection attribution stage is inherently hypothesis-generating: researchers can investigate which policies, shocks, or events most plausibly explain the timing and pattern of the identified breaks. Attribution can proceed through structured database searches that match break dates and locations to documented policy actions, or through systematic text-based or archival searches. Advances in artificial intelligence may further enhance this process. Large language models (LLMs), for instance, could support automated attribution by linking detected breaks to contemporaneous events across multiple sources while maintaining transparency through verifiable reference data. Integrating such attribution strategies extends the empirical reach of break detection from identifying where effects occurred to explaining why they did.

We illustrate our method with two distinct applications. First, to demonstrate the method's specific utility for climate policy, we replicate and expand the analysis of the Swedish carbon tax by Andersson (2019). We show that our method detects the impact of the 1991 carbon tax on transport emissions as a structural break without prior knowledge of the tax's introduction, recovering treatment estimates consistent with Andersson's synthetic control results. Second, we use regional data from Spain to show that the economic impact of ETA terrorism can be detected directly from the data without prior knowledge of the occurrence of terrorism, replicating the results of Abadie & Gardeazabal (2003).

By connecting break detection to heterogeneous-treatment models, this paper develops the econometric underpinnings of an emerging class of data-driven policy-evaluation tools. We make the methods available in the freely-available open-source R package *getspanel*. The framework offers a general, theory-based approach for identifying unknown interventions and provides the foundation for a growing literature that applies these methods to study economic, climate, and other policy domains. Our proposed panel break detection approach can be readily implemented using our accompanying R-packages *gets* (Pretis et al. 2018) and *getspanel* (Schwarz & Pretis 2026).

1.1 Related Literature & Contribution

We connect two distinct strands of literature: the econometric analysis of structural breaks in time series and panels, and the program evaluation literature on difference-in-differences (DiD) with heterogeneous effects.

Break detection to assess the impact of policy has been commonly used in time series analysis. A non-exhaustive list of examples in the time series literature ranges from Perron (1989) detecting breaks in GNP time series attributed to the Great Depression and an oil price shock, Hendry (2020) identifying policy interventions in UK CO_2 per capita emissions, Estrada et al. (2013) quantifying the impact of the Montreal Protocol on CFC emissions and subsequently temperatures, to Apergis & Lau (2015) identifying whether breaks in Australian electricity markets align with policy interventions. Piehl et al. (2003) also use the detection of breaks in time series to assess treatment effectiveness of a youth homicide prevention programme in Boston. However, most time series applications do not have control groups, making a clear causal interpretation of any break difficult.[1]

In the panel data literature, structural break detection has largely focused on testing for breaks in slope coefficients of random variables or common factors, rather than identifying treatment effects via fixed effects. Chan et al. (2008) extend the Andrews (1993) test to panels, while Baltagi et al. (2016) and Bai (1997) study heterogeneous panels with breaks. More recently, Qian & Su (2016) and Li et al. (2016) have applied LASSO-type estimators to detect common breaks or breaks in interactive fixed effects. However, these contributions typically view breaks as nuisance parameters or structural instabilities in the slope coefficients of controls, rather than as the parameter of interest representing a treatment effect.

Our primary contribution is to explicitly link structural break detection in panel fixed effects to the estimation of treatment effects. We show that detecting a step-shift in a unit's fixed effect is formally equivalent to estimating a treatment effect in a difference-in-differences design where the treatment timing and assignment are unknown. This aligns our work with the rapidly expanding literature on DiD with staggered adoption and heterogeneous treatment effects (Roth et al., 2023; Callaway & Sant'Anna, 2020; Goodman-Bacon, 2021). Specifically, Wooldridge (2025) demonstrates that heterogeneous treatment effects can be consistently estimated in a TWFE framework using interactions of treatment times and unit dummies. We show that our break detection procedure nests Wooldridge's specification: the "unknown" treatment effects are simply the interaction terms selected by model selection procedures from a saturated model.

Our paper thereby provides a theoretical foundation for recent applied papers detecting treatment as structural breaks, particularly in the context of climate and environmental policy evaluation. Koch et

[1] A causal interpretation in time series nevertheless is possible where breaks occur in some conditioning variables under super exogeneity (Bazinas & Nielsen, 2015). This has been shown first in Engle et al. (1983) as causal relations invariant to shocks (referred to as super exogeneity). Under such super exogeneity causal identification is possible, see e.g. Martinez (2020), Mukanjari & Sterner (2018), or Pretis (2021) for relevant examples of this. Where super exogeneity does not hold, has not been tested, or is difficult to establish, however, a causal interpretation of structural breaks is more difficult.

al. (2022) used break detection to identify major reductions in road-transport emissions and attribute them to policy mixes. Stechemesser et al. (2024) applied similar techniques in *Science* to detect climate policies that achieved large emission reductions worldwide. Pretis (2022) examined Canadian CO_2 emissions and British Columbia's carbon tax, and Tebecis & Crespo Cuaresma (2025) compiled a global dataset of structural breaks in greenhouse gas emissions. Tebecis (2023) further evaluated Austrian climate policies using the same approach. These studies demonstrate the value of detecting unknown treatment episodes directly from data, yet until now the econometric foundations underlying these applications have not been formally established. We provide this theoretical basis and clarify how such applications relate to the broader literature on program evaluation with heterogeneous and staggered treatment. Note that operationalising reverse-causal modelling does not imply reverse causality between variables but rather the identification of when and where interventions occurred. It can serve as the first step in a broader empirical strategy, helping researchers detect potential treatment episodes and subsequently evaluate their causal impact. In this way, it complements rather than replaces standard program-evaluation techniques.

There are a range of nuances to our proposed approach. First, if treatment assignment and timing is known (and happens to have a large effect), then imposing interacted treatment dummies allowing for heterogeneous treatment effects for the known intervention in a TWFE estimator is effectively equivalent to agnostically detecting a break in this fixed effect (if that is the only break retained) and estimating the model post-break detection. The estimated model with an imposed break or a single retained break is *identical* (subject to accounting for any selection bias). In other words, if the intervention was known, we could simply run a TWFE estimator, which will be equivalent to having found the one particular intervention and then assessing the estimated model.

Second, our idea is modular with respect to known treatment. If there is a known intervention, we can impose it into the model without selection and estimate its impact, while at the same time searching for additional breaks. This allows us to assess the impact of a known policy while also detecting potentially unknown interventions, effectively implementing the theory-embedding approach described in Hendry & Johansen (2015). It is worth noting that our approach concentrates on causes of effects by first identifying effects. If there are no effects, naturally we cannot find any corresponding cause. Thus, for unknown treatment we cannot distinguish between no treatment or a zero treatment effect. For known treatment, however, this is not a concern as it can easily be embedded as a forced a-priori treatment variable that is introduced into the model independent of the selection. We can also restrict the search for breaks and treatment to a subset of units if we suspect that some units may be treated and are certain that others are not.

Third, our conceptual approach is also modular in terms of the choice of detection method. We can use different machine learning methods of our choice to detect breaks (i.e. treatment), depending on the preferred properties of the selection algorithm. For example, if our main concern is the false-positive detection rate, then we can choose to use methods that control the false discovery rate (such as general-to-specific selection methods, henceforth 'gets'). If instead we care about computational speed, we could use regularised estimators, such as the (adaptive) LASSO. More recently, this high-dimensional selection problem has been extended to Bayesian frameworks. Notably, Konrad et al. (2026) develop the Bayesian Indicator Saturated Model (BISAM), which builds on indicator saturation principles to provide a coherent Bayesian method for identifying structural shifts and outliers.

There are of course some constraints to our methods. First, when detecting breaks in individual fixed effects, each treated unit will be identified with a separate treatment dummy. While this allows for straightforward heterogeneous treatment effects, it means that we do not gain power if there are multiple

units that received the same treatment. Thus, our TWFE break detection approach mirrors the use of interactions to identify known heterogeneous treatment effects (Wooldridge, 2025) and lends itself to panels with longer time series and smaller cross-sectional dimensions with heterogeneous treatment (similar to settings encountered when using synthetic controls).

Second, all break detection methods evaluate the presence of breaks relative to a specified underlying model. If the model is not well-specified, then breaks that we detect may simply reflect model mis-specification. This is of course also a problem in conventional TWFE difference-in-differences settings, however, can be amplified in our setting if we attempt to attribute a 'spurious' break to an event. This effect can be mediated by selecting at tight significance levels to control the false-positive rate (when using gets, see section 3.1.1) or by making use of robust estimators less sensitive to observations falling outside the specified model (such as embedding break detection in a wider outlier-robust estimation framework, e.g. Impulse Indicator Saturation, IIS – see Hendry et al. 2008; Jiao & Pretis 2020, Jiao et al. 2021).

Third, once a treatment effect is detected in the form of a break, it has to be attributed to a potential cause. In other words, the detection of a break generates potential hypotheses around the potential cause of the break. Attributing breaks to potential causes can be done through structured searches in the existing literature and records, and LLMs may be useful tools in this setting. While attribution can be a challenge and requires subject-specific knowledge, it offers an opportunity to learn from the data. A search for a potential cause of an effect is comparable to arguing that a known intervention was exogenous (or as-if randomly assigned) in a conventional programme evaluation application.

The roadmap for the remainder of the paper is as follows. In section 6.1 we first provide a simple illustration of how structural breaks are closely linked to treatment evaluation in two-way fixed effects estimators. We consider the standard case of known treatment assignment and illustrate its equivalence to a step-shift break in the treated units' intercept. In section 2.2 we then consider the case where treatment assignment and timing is unknown, and we establish that unknown treatment assignment and timing can be identified using impulse dummies in a saturated regression (for fully time-varying effects) and step-dummies (for piece-wise constant effects). Using recent results from Wooldridge (2025), we show how time-varying and unit-heterogeneous treatment effects can be nested in dummy-saturated models with more variables than observations. We further show that if we detect multiple treatment breaks (in multiple different units), they can be interpreted as time-varying treatment effects in a staggered treatment intervention setting. We discuss our approach in a balanced panel without explicitly discussing control variables, however, the results should generalise to the inclusion of other covariates. In the following section 3 we then briefly discuss two estimation approaches using general-to-specific (gets) selection and the adaptive LASSO (and provide some simulations in the Supplementary material). Finally, we apply our methods to models of Swedish transport CO_2 emissions and Spanish regional GDP per capita in section 4.

2 Conceptual Approach: Break Detection to Detect Unknown Treatment Assignment & Timing

We consider the detection of treatment interventions and subsequent estimation of treatment effects when both treatment assignment and treatment timing are unknown. We show that if treatment assignment and timing are unknown, such treatment can be identified by allowing for potential structural breaks at any point in time for any unit in a model including individual and time fixed effects. Applying machine learning methods allowing for model selection for more variables than observations, we then remove all irrelevant treatment dummies and are left with the resulting model that identifies treatment assignment, timing, and estimates treatment effects conditional on the treatment effects being non-zero.

To aid the reader's understanding, we present a stylised example in Figure 1, which provides guidance for the various aspects of our conceptual approach we touch upon in the following sections. The example in Figure 1 consider two units (e.g. regions), one of which is treated while the other acts as control. If treatment timing and assignment are known (left panels), then we can estimate the effects either as time-varying (first column) or as average treatment effects on the treated. If treatment is unknown (right panels), then treatment effects can be detected by saturating the models with a full set of possible treatment interventions (gray indicators), and retaining only the 'large' ones (red retained indicators). If treatment is sparse affecting only a subset of the units and treatment effects are sufficiently large, then we can detect treatment as a structural break in the treated unit without prior knowledge.

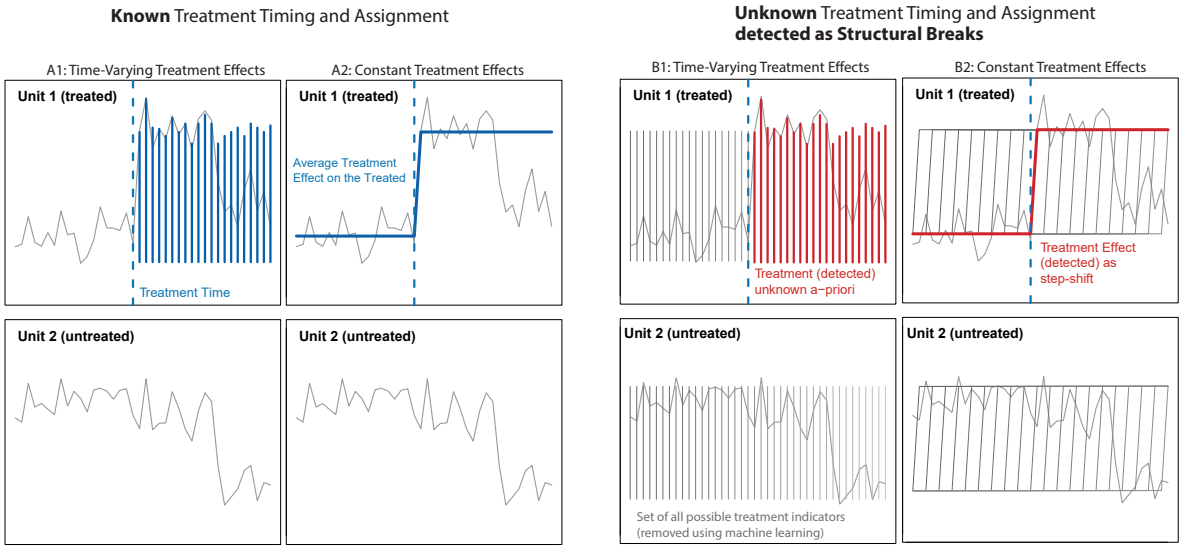

Figure 1: Detecting unknown treatment timing and assignment as structural breaks – a stylised example using artificial data. Left: 'Known' Treatment baseline for time-varying and constant treatment effects. Right: Detecting treatment as breaks using impulses for time-varying and step-shifts. All possible impulse and step-indicators shown in grey, a subset of which (red) identify the true underlying treatment (blue in left panels).

2.1 Setting & Linking Known Treatment to Structural Breaks

To illustrate the overall motivation and the close link between structural breaks and treatment evaluation, consider a panel of N units over T time periods where one group is treated with a single treatment from time $t = q$ onwards. We initially consider the baseline case of known treatment assignment and timing, where the treatment indicator $d = 1$ for the treated group (or individual) and $d = 0$ for the untreated.

Borrowing the notation in Wooldridge (2025), we denote by $y_t(0)$ the outcome in the untreated control group, and $y_t(1)$ the outcome in the treated group at time t. The treatment effect at time t due to treatment occurring from time $t = q$ onwards is given by the difference $y_t(1) - y_t(0)$. As is convention in the literature, we focus on the average treatment effect on the treated τ_t:

$$\tau_t = E[y_t(1) - y_t(0)|d = 1] \tag{1}$$

To identify the average treatment effect we assume there is no anticipation of treatment, in other words, the potential outcome for a unit prior to treatment is identical to the untreated units:

$$E[y_t(1) - y_t(0)|d = 1] = 0, \text{ for t} < \text{q} \tag{2}$$

Further we rely on the common trend assumption which is standard in much of the treatment effects literature:

$$E[y_t(0) - y_{t=1}(0)|d] = E[y_t(0) - y_{t=1}(0)] = \theta_t, \text{ for } t = 2, ..., T \tag{3}$$

Finally, we also assume there is at least one untreated unit. We then write the observed outcome as:

$$y_t = y_t(0) + d[y_t(1) - y_t(0)] \tag{4}$$

The expected outcome conditional on treatment is:

$$E[y_t|d] = E[y_t(0)|d] + d \times \tau_t \tag{5}$$

We define the change in y_t over time in absence of treatment as:

$$g_t(0) = y_t(0) - y_{t=1}(0) \tag{6}$$

and under the common trend assumption we have that $E[g_t(0)|d] = E[g_t(0)] = \theta_t$. We thus have that:

$$E[y_{t=1}(0)|d] = \lambda + \xi d \tag{7}$$

where ξ denotes the average pre-treatment difference between the treated and untreated groups and λ denotes the average level of y for the untreated. Combining all above yields the expected value of y_t conditional on treatment as:

$$E[y_t|d] = \lambda + \xi d + \theta_t + d \times \tau_t \tag{8}$$

For illustration purposes, assume the treatment effect is constant over time, $\tau_t = \tau$. Under the assumption of no anticipation we have that:

$$\begin{aligned} E[y_t|d] &= \lambda + \xi d + \theta_t, \text{ for } t < q & (9) \\ &= \lambda + \xi d + \theta_t + d \times \tau, \text{ for } t \geq q & (10) \end{aligned}$$

This is a standard result in the treatment effects literature and the above model can be consistently estimated using a TWFE estimator (see e.g. Wooldridge 2025):

$$y_{i,t} = c_i + g_t + \tau w_{i,t} + u_{i,t} \tag{11}$$

with $w_{i,t} = d_i \times q_t$ where d_i is an indicator for whether the individual is treated, q_t an indicator for the post-treatment period, c_i denote individual fixed effects, and g_t time fixed effects. Note that Wooldridge

(2025) groups the untreated mean into a single intercept, however, the treatment effect estimates are unaffected by whether we include a common intercept or allow for unit-specific intercepts (i.e. fixed effects). The above model shows the close link between treatment effects and structural breaks: we identify the average treatment effect as a step-shift of magnitude τ at time q in the treated unit's intercept:

$$\begin{aligned} E[y_{i,t}|d_i = 1] &= c_i + \tau \times 1_{\{t \geq q\}} + g_t \quad (12) \\ &= c_{i,t} + g_t \\ \text{where } c_{i,t} &= \begin{cases} c_i & \text{for } t < q \\ c_i + \tau & \text{for } t \geq q \end{cases} \quad (13) \end{aligned}$$

Figure 1 (column A2, left) shows this in a stylised example illustrating how a constant treatment effect corresponds to a simple step-shift in the individual-specific intercept.

Time-Varying Treatment Effects Allowing for time-varying treatment effects τ_t we can write the expected outcome conditional on treatment as:

$$\begin{aligned} E[y_t|d] &= \lambda + \xi d + \theta_t, t < q \\ &= \lambda + \xi d + \theta_t + d \times \tau_t, t \geq q \quad (14) \end{aligned}$$

If treatment assignment and timing were known, the above could be consistently estimated using interactions in a TWFE estimator (see Wooldridge 2025 for the 'known treatment' case), where again we here allow for unit-specific intercepts:

$$y_{i,t} = c_i + g_t + \sum_{s=q}^{T} \tau_s (d_i \cdot 1_{\{t=s\}}) + u_{i,t} \quad (15)$$

This is equivalent to a set of step-shifts of duration 1 (with common coefficient over i in H) at times $q, q+1, ..., T$, where each time step is denoted as an index s, which ranges from $s_1, s_2, ..., S$. We now relax the assumption of having common coefficients over i, in other words, we allow for heterogeneity in treatment effects over i. Let $H = \{m_1, m_2, ..., m_M\}$ denote the set of M treated units. For example if units $i = 2$ and $i = 3$ are treated, then there are two treated units $M = 2$, and $m_1 = 2$ and $m_2 = 3$. The above model (15) can then be written in a general specification allowing for unit-heterogeneous treatment effects at every time as:

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \sum_{s=q}^{T} \tau_{j,s} 1_{\{i=j,t=s\}} + u_{i,t} \quad (16)$$

where $1_{\{i=j,t=s\}}$ denotes an indicator function equal to one for all treated i in the set of treated units H and $t = s$ in the post-treatment period $s \geq q$, and zero otherwise. This specification relaxes the restriction of homogeneous treatment effects across treated units. Figure 1 (column A1, left) shows a stylised example of individual impulses capturing a treatment effect. Specifically, each treated post-treatment observation is captured by a single time-period dummy. While these cannot be estimated consistently because they capture single observations, such dummy variables can be estimated unbiasedly (see Hendry & Santos 2005) and, as we showed here, identify unit- and time-specific treatment effects.

To relate the known case to the unknown treatment setting, we further refine our notation. We define

an index of the timing of non-zero treatment effects for each treated unit $j \in H$ denoted as $R_j = \{q_{j,s=1}, q_{j,s=2}, ..., q_{j,S_j}\}$ where S_j denotes the number of treatment indicators for unit j. For example, suppose that in a 3-unit panel with $T = 20$ observations units $i = 2$ and $i = 3$ are treated ($m_1 = 2, m_2 = 3$) with non-zero treatment effects from $t = q, ..., T$. Then $H = \{2, 3\}$ with corresponding treatment effects at $R_2 = \{q_{2,1} = q, q_{2,2} = q+1, ..., q_{2,S_2} = T\}$ and $R_3 = \{q_{3,1} = q; q_{3,2} = q+1, ..., q_{3,S_3} = T\}$. With common treatment timing and effects this implies that $R_2 = R_3$. Thus the known-treatment and known-timing baseline in (16) can be written as:

$$y_{i,t} = c_i + g_t + \sum_{j=m_1}^{m_M} \sum_{s=q_{j,1}}^{q_{j,S_j}} \tau_{j,s} 1_{\{i=j,t=s\}} + u_{i,t} \tag{17}$$

or by simplifying notation as:

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \sum_{s \in R_j} \tau_{j,s} 1_{\{i=j,t=s\}} + u_{i,t} \tag{18}$$

Our proposed approach will be to recover this known treatment model without prior knowledge of treatment timing and assignment.

2.2 Unknown Treatment Assignment & Timing for a Single Treatment

Now what if treatment assignment and timing are unknown? The above model in (18) constitutes the 'known' intervention baseline, i.e. the target of model selection/break detection. In 2.2.1 we now consider the detection of treatment assignment and timing allowing for unit-heterogeneous and time-varying treatment effects as in (18) which we will show is matched by a saturating set of unit-time-specific impulse dummies. We then consider treatment detection allowing for unit-heterogeneous but piece-wise constant treatment effects over time, which we will show is nested by a saturating set of unit-specific step-shift breaks.

2.2.1 Detecting Unknown Treatment When Treatment Effects are Fully-Time Varying

If treatment assignment and treatment timing is unknown, we propose that the 'known' treatment model (18) can be embedded in a general model allowing for potential treatment of any unit at any point. The most flexible specification that nests the 'known' treatment specification (18) as a special case is a fully-saturated model allowing for a treatment dummy for each individual at every point in time:

$$y_{i,t} = c_i + g_t + \sum_{j=1}^{N} \sum_{s=1}^{T} \tau_{j,s} 1_{\{i=j,t=s\}} + u_{i,t}. \tag{19}$$

The model in (19), which identifies unit-specific treatment effects for each unit for each time period, however, cannot be estimated as such because the number of parameters matches (or exceeds) the number of observations. Effectively there are NT possible indicator variables added to the balanced panel. Figure 1 (column B1, right) shows the full set of these impulse indicators, a subset of which identify the true treatment effect shown in panels on the right.

The aim is then to reduce the general model (19) to a sparse model, ideally coinciding with the underlying target of the known baseline (18). Thus, we require the additional assumption that treatment effects are sparse, we have at least one untreated unit and some untreated time-periods for treated units – assumptions that are very common in the wider treatment evaluation literature. Starting with this general

model, we then apply machine learning/model selection to remove all but 'relevant' dummy variables using selection algorithms capable of handling more variables than observations. We discuss two possible machine learning algorithms in more detail in section 3.1. In fact, the dummy-saturated model in (19) is equivalent to an outlier-robust Huber-skip estimator, where the retained impulse dummies detecting 'outliers' relative to the model capture the time-varying unit-specific treatment effects (see e.g. Jiao et al. 2021, Hendry et al. 2008, Johansen & Nielsen 2009, Johansen & Nielsen 2016a). We write the sparse final selected model as:

$$y_{i,t} = \hat{c}_i + \hat{g}_t + \sum_{j \in \widehat{H}} \sum_{s \in \widehat{R}_j} \hat{\tau}_{j,s} 1_{\{i=j, t=s\}} \tag{20}$$

where we effectively estimate treatment assignment $\widehat{H} = \{\hat{m}_1, \hat{m}_2, ..., \hat{m}_{\hat{M}}\}$, together with the index of treatment occurrence $\widehat{R}_j = \{\hat{q}_{j,1}, \hat{q}_{j,2}, ..., \hat{q}_{j,\hat{S}_j}\}$ where $\hat{S}_j$ denotes the number of treatment indicators for unit j, and the time-varying and unit-specific treatment effects $\hat{\tau}_{j,s}$ conditional on having non-zero treatment effects. Note that we detect treatment when it has an effect, i.e. we detect treatment effects conditional on them being non-zero (and depending on the selection algorithm, large enough to be detectable). Using the resulting estimated treatment effects $\tau_{j,s}$ we can compute the average treatment effects for the treated (ATTs) for specific units or time periods. As impulse indicators are orthogonal (and properties of Huber-skip estimators are reasonably well-understood), we can also compute the standard error for the resulting ATTs. For example, we can compute the estimated ATT for individual j over the entire period of non-zero detected treatment effects as:

$$\widehat{ATT}_j = \frac{1}{\hat{S}_j} \sum_{s=1}^{\hat{S}_j} \hat{\tau}_{j,s}, \text{ with standard error } se\left(\widehat{ATT}_j\right) = \sqrt{\frac{1}{\hat{S}_j} \sum_{s=1}^{\hat{S}_j} se\left(\hat{\tau}_{j,s}\right)^2} \tag{21}$$

If we are interested in a subset of treated periods we could simply restrict the ATT to those relevant time periods (or units). A remaining issue, however, is that detecting individual impulses may suffer from low power if treatment effects are small and actually constant over some period of time. If we are interested in ATTs over time for some treated units, allowing for piece-wise constant treatment effects may yield higher power of detection which we discuss in the next section 2.2.2.

2.2.2 Detecting Unknown Treatment With Piece-Wise Constant Treatment Effects

While treatment effects may be heterogeneous over individuals i, they may be constant for some time periods. Such constancy over time can lead to higher power to detect treatment. Consider treatment effects in (14) that are constant over time following treatment from $t = q$ onwards, but allowed to vary over treated individuals:

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \tau_j 1_{\{i = j,\ t \geq q\}} + u_{i,t} \tag{22}$$

Then for each treated unit in H (where $d_i = 1$) with time-invariant treatment effect $\tau_{i,t} = \tau_i,$ for all t, the change from pre-treatment to post-treatment is given by a step-shift change in the unit-specific intercept of magnitude τ_i. For example, for treated unit i with time-invariant treatment effect, the expected

outcome is given by:

$$\begin{aligned} E[y_{i,t}|d_i = 1] &= c_i + g_t + \tau_i \times 1_{\{t \geq q\}} \\ &= c_{i,t} + g_t \\ \text{where } c_{i,t} &= \begin{cases} c_i & \text{for } t < q \\ c_i + \tau_i & \text{for } t \geq q \end{cases} \end{aligned} \tag{23}$$

which is just a step-shift in the unit-specific intercept (i.e. fixed effect), equal to c_i prior to treatment, and $c_i + \tau_i$ post-treatment. Estimates of τ_i then correspond to the unit-specific average treatment effect over time. If treatment timing and assignment are unknown, we can generalise the impulse-dummy approach to nest known treatment as a special case in a general model now allowing for step-shifts at any point in time as:

$$y_{i,t} = c_i + g_t + \sum_{j=1}^{N} \sum_{s=2}^{T} \tau_{j,s} 1_{\{i=j, t \geq s\}} + u_{i,t} \tag{24}$$

where a subset of the step-functions $1_{\{i=j,\ t \geq s\}}$ correspond to the actual treatment effects model in (22). This allows for any unit to be potentially treated at any point in time – with s starting at 2 rather than 1, so as not to coincide with the fixed effect in c_i. We then aim to remove treatment indicators such that we only retain the subset of truly treated units and time periods. Under sparsity of treatment effects i.e., there remain units and time periods without treatment, we write the final selected model as:

$$y_{i,t} = \hat{c}_i + \hat{g}_t + \sum_{j \in \widehat{H}} \sum_{s \in \widehat{R}_j} \hat{\tau}_{j,s} 1_{\{i=j, t \geq s\}} \tag{25}$$

where we estimate treatment assignment by detecting those units i that have at least one break indicator retained, and break times are estimated by the starting date of each retained break function. Figure 1 (column B2, right) shows the full set of step-functions, a subset of which identify the true treatment effect.

Identification Relative to Fixed Effects It is crucial to recognize that structural breaks in this framework are identified relative to the unit and time fixed effects. In a two-way fixed effects model, a detected break represents a unit-specific deviation from the common time trend. In panels with a small number of units, this implies that breaks are technically identified only relative to the other units in the sample.

Consider the limiting case of $N = 2$ with one treated unit ($i = 1$) experiencing a break of magnitude τ at time q, and one control unit ($i = 2$). The standard representation is $y_{1,t} = c_1 + g_t + \tau 1_{\{t \geq q\}}$. However, this is observationally equivalent to a specification where the common time trend absorbs the shift (i.e., $g'_t = g_t + \tau \times 1_{\{t \geq q\}}$) and the control unit experiences a break of opposite magnitude $-\tau$ (i.e., $y_{2,t} = c_2 + g'_t - \tau \times 1_{\{t \geq q\}}$). Thus, in small samples, we technically identify the relative difference in breaks between units rather than an absolute structural shift. This ambiguity disappears as the number of untreated units increases; with a sufficiently large control group, the time fixed effects g_t are pinned down by the average outcome of the untreated majority. Consequently, the deviation is correctly attributed to the treated unit diverging from the stable common trend.

Time-varying Treatment Effects as Linear Trends Note that this setup does not impose that treatment effects have to be strictly constant over time post-treatment, as a linear combination of step-functions can capture time-varying treatment effects. In fact, our general framework also allows for treatment effects which are non-constant across the time dimension. This could for example take the

form of broken linear trend functions – so called Trend Indicator Saturation (TIS) – as developed by Castle et al. (2025). Based on equation (24), we can therefore replace the step-indicators with trend-indicators in the form:

$$y_{i,t} = c_i + g_t + \sum_{j=1}^{N} \sum_{s=2}^{T} \tau_{j,s}(t-s)1_{\{i=j,t\geq s\}} + u_{i,t} \tag{26}$$

These variations of different saturation methods allow our framework to substantially expand the type of heterogeneous treatment effects that can be considered in a standard causal inference set-up.

2.2.3 Unknown Treatment Assignment and Timing For Multiple Treatments

If there is a single underlying treatment and break detection identifies a single intervention then the interpretation and attribution of detected effects is straightforward. However, in practice there may be multiple treatments detected as breaks at different times for multiple different units. Irrespective of the selection algorithm employed (see section 3.1), consider the following final retained model with a range of detected treatment impulse dummies:

$$y_{i,t} = \hat{c}_i + \hat{g}_t + \sum_{j\in\widehat{H}} \sum_{s\in\widehat{R}_j} \hat{\tau}_{j,s}1_{\{i=j,t=s\}} \tag{27}$$

or step-functions:

$$y_{i,t} = \hat{c}_i + \hat{g}_t + \sum_{j\in\widehat{H}} \sum_{s\in\widehat{R}_j} \hat{\tau}_{j,s}1_{\{i=j,t\geq s\}} \tag{28}$$

What is identified if the detected treatment time varies across units in the panel? For example, what if we find both $j = 1$ and $j = 2$ to be in the treated group, but their treatment timing differs, i.e. $R_{j=1} \neq R_{j=2}$ for both $j = 1$ and $j = 2$? We show that the final retained models with heterogeneous treatment dummy variables (27) and (28) are equivalent to staggered treatment with heterogeneous effects where heterogeneity and staggered adoption are captured through interactions. In other words, the impulse indicator estimator identifies unit and time-specific staggered treatment effects conditional on the treatment effect being non-zero. If treatment effects are constant over time, then a saturating set of step-functions nests the known-treatment assignment and timing model as a special case even when treatment is staggered. To illustrate this equivalence, we follow the discussion in a known-treatment setting by Wooldridge (2025) on how interaction terms identify treatment effects in a staggered treatment setting. Subsequently we show that this is nested by the IIS and SIS break detection estimators in an unknown treatment setting, establishing that detected breaks identify unit- and time-specific treatment effects.

For exposition, consider a staggered treatment DGP where we denote the time of the first intervention by q. We define treatment cohort dummies as $d_q, ..., d_Q$ where Q denotes the final time of intervention, which would be equal to T when treatment lasts until the end of the sample. We refer to the time of each intervention as $r \in \{q, q+1, ..., Q\}$. The potential outcome at time t for unit treated at time r is given by $y_t(r)$, with the outcome for the never treated unit referred to as $y_t(\infty)$ i.e. treated at no point in time. The quantities of interest are the treatment effects of each unit first receiving treatment at time r given by the difference in outcomes $y_t(r) - y_t(\infty), r = q, ..., Q$. In a staggered setting we hope to identify the average treatment effects on the treated ATT for each intervention (given by different cohorts which we

will relax to different individuals):

$$\tau_{r,t} = E[y_t(r) - y_t(\infty)|d_r = 1], r = q, ..., Q; t = r, ..., T. \tag{29}$$

Under no anticipation and common trends in a standard known-treatment setting, Wooldridge (2025) demonstrates that heterogeneous treatment effects can be consistently estimated in a staggered treatment setting using interactions in a TWFE estimator (we replicate the derivation in Supplementary Section 6.2). Specifically, the expected outcome in a staggered treatment setting can be written as

$$\begin{aligned} E[y_t|\mathbf{d}] &= \eta + \lambda_q d_q + ... + \lambda_Q d_Q + \theta_t \text{ (pre-treatment } t < q) \\ &= \eta + \lambda_q d_q + ... + \lambda_Q d_Q + \theta_t + \tau_{q,t} d_q + ... + \\ &\quad + \tau_{Q,t} d_Q \text{ (post-treatment } t \geq q) \end{aligned} \tag{30}$$

where η is the average level of y for the untreated group and λ_q refers to the average level of y for the treated cohorts pre-treatment. This can be consistently estimated using a TWFE estimator with time-cohort interactions as:

$$y_{i,t} = c_i + g_t + \sum_{r=q}^{Q} \sum_{s=r}^{T} \tau_{r,s}(d_{i,r} \cdot 1_{\{t=s\}}) + u_{i,t} \tag{31}$$

In the above equation each cohort has a set of time-varying treatment effect estimates. Now consider each treated unit in the panel being allowed its own treatment effects (i.e. each unit in each cohort receives its own treatment effect or each cohort is of size one). As before, consider $H = \{m_1, m_2, ..., m_M\}$ as the set of i that are treated at some time, where treatment timing is not exclusive. In other words, m_1 and m_2 may be treated at the same time (but may also be treated at different times). Then relaxing the above assumption that each treatment cohort has the same treatment effect, the above model (31) can be written as:[2]

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \sum_{s=r}^{T} \tau_{j,s} 1_{\{i=j, t=s\}} + u_{i,t} \tag{33}$$

This is identical to the interaction of the treatment dummy $d_{i,r}$ and time dummies $1_{\{t=s\}}$ above, except we disaggregate treatment cohorts into individual units. Using simplifying notation, we can write (33) as:

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \sum_{s \in R_j} \tau_{j,s} 1_{\{i=j, t=s\}} + u_{i,t} \tag{34}$$

This matches the impulse-dummy saturated final specification where treatment assignment and timing is estimated in (27). Similarly, if treatment effects are piece-wise constant over time we can write (33) as:

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \sum_{s \in R_j} \tau_{j,s} 1_{\{i=j, t \geq s\}} + u_{i,t} \tag{35}$$

Estimating this heterogeneous staggered-treatment model matches the post-selection step-function model in (28). Thus, detecting multiple treatments through impulses or step-indicators is equivalent to the estimation of treatment effects in a staggered-intervention setting when heterogeneous treatment effects are

[2]To recover (31) we could restrict equation (33) as:

$$\tau_{m_l,s} = \tau_{m_k,s} = \tau_{r,s}, \text{ for } k \neq l \text{ and } (m_l, m_k) \in \text{the same treatment cohort } r \tag{32}$$

identified using interactions. We identify average treatment effects (over time) for each treated unit relative to the never treated cases, conditional on the treatment effect being large enough to be detected. If a single unit experiences more than one treatment – then this can be interpreted as time-varying treatment (where the sum of effects is the treatment effect relative to the never treated), or a separate treatment event relative to treatment received earlier.

2.3 Challenges

Naturally there are challenges to our proposed approach to detect treatment, and properties of the final identified model will depend on the machine learning/model selection algorithms employed. In section 3, we briefly discuss the general properties of using 'gets' (through impulse- and step-indicator saturation, IIS and SIS respectively) and the (adaptive) LASSO (Tibshirani 1996, Zou 2006) and how they relate to the power of identifying treatment correctly, controlling the false-positive rate of retained break variables, and conducting valid inference.

First, we may miss that treatment occurred (a relevant break variable is not retained). However, our approach allows us to embed a known or suspected treatment just like in a difference-in-difference treatment evaluation setting (see section 3.2). Additionally, varying the acceptable false-positive rate can also result in identifying more potential treatments. Second, we may detect spurious treatment as false-positives by retaining irrelevant break variables. Though, the 'gets' approach described in section 3.1.1 allows an explicit control of the false-positive rate. Third, we may face challenges with post-selection inference (effects may be biased as large breaks are more likely to be retained than small ones). Some of these concerns can be mitigated through bias-correction and adjustment for post-selection inference.[3] Fourth, if treatment affected all units under analysis, then the treatment effect will be subsumed into the year fixed effects g_t and not detectable as such – but again such a treatment would also not be identified with comparable treatment evaluation methods.

3 Operationalising the Detection of Treatment Assignment and Timing

3.1 Detection Methods and Their Approximate Properties

The idea of detecting structural breaks to identify treatment can be operationalised by applying break-detection in a panel setting, starting with the general saturated models for impulses (19), steps (24), or trends (26). We emphasise that the idea of detecting treatment by detecting breaks is separate from the method of implementation – there are numerous possible machine learning detection/selection methods available and their properties will determine the effectiveness of detecting previously unknown treatment. Here we briefly consider two model selection approaches: general-to-specific selection using impulse- (interpreted as an outlier-robust estimator), step-, and trend-indicators; and the (adaptive) LASSO, though we are not limited to these in practice. For example, the group-specific break detection approach in Okui and Wang (2021) could be a promising future avenue of detecting treated groups rather than individuals. Note that in the outlier-robust/general-to-specific setting, the problem of variable selection is generally studied under the null of no treatment – i.e. we focus on the false positive rate of detection which is also the main calibration parameter. In turn, in the shrinkage-based model selection literature (e.g. the LASSO) the focus has been on consistent selection, with less attention being paid to the false positive rate. Further, the Bayesian Indicator Saturated Model (BISAM) proposed by Konrad et al. (2026) provides a viable alternative framework.

[3] See e.g., the coefficient bias correction function in the 'gets' package.

3.1.1 Treatment Detection using 'gets' and Impulse- or Step- Indicator Saturation

The impulse-indicator saturated model (19) is equivalent to impulse indicator saturation (IIS, see Hendry, Johansen and Santos, 2008) in a panel and can be interpreted as a Huber-skip outlier-robust estimator (see e.g. Jiao et al. 2021, Johansen & Nielsen 2009, Johansen & Nielsen 2016a). Coefficients on dummies that are used to determine outliers correspond thus to individual- and time-specific treatment effects. IIS has well-established properties under the null of no outliers (here interpreted as no treatment/zero treatment effects) where the false positive rate can be easily controlled by specifying the relevant tuning parameter. IIS corresponds to a robust Huber-skip estimator and targets the false-positive rate of detection by removing impulse indicators up to the chosen level of significance γ_c. For example, under a normal reference distribution, choosing $c = 1.96$ would correspond to a target level of significance of $\gamma_{1.96} = 0.05$. We denote the observed false positive rate $\hat{\gamma}_c$ as the proportion of spuriously retained indicators at the chosen cut-off c out of all possible break variables considered:

$$\hat{\gamma}_c = \frac{L_c}{L} \tag{36}$$

where $L_c = \sum_{j=1}^{\hat{M}} \hat{S}_j$ is the number of retained indicator variables at cut-off c and L denotes the total number of potential treatment variables selected over, usually equal to the total sample size $L = n = NT$ in a balanced panel allowing for treatment at any point in time for every unit. The asymptotic properties of IIS under the null of no breaks as the total sample size $n \to \infty$ are explored in Hendry et al. (2008) and Johansen and Nielsen (2009; 2016b), who show that when there are no breaks (and accounting for multiple testing), the false positive rate of retained breaks (i.e. the number of retained indicator dummies L_c relative to all possible indicators L) converges to the chosen nominal level of significance of selection γ_c:

$$\hat{\gamma}_c = \frac{L_c}{L} \to \gamma_c, \text{ as } n \to \infty \tag{37}$$

where n denotes the sample size (in a balanced panel $n = NT$). In other words, if there is no treatment effect (i.e. if there are no true underlying breaks), then the proportion of spuriously detected indicators converges to the chosen level of significance, e.g. 1% for $\gamma_c = 0.01$. In the present context of detecting treatment at any point in time for any unit in a balanced panel, selecting at $\gamma_c = 0.01$ yields an expected number of $0.01 \times NT$ spuriously retained indicators. Thus, IIS in a Huber-skip robust interpretation makes it straightforward to control the false discovery rate of breaks (and thereby treated units) by varying γ_c.

We can estimate the set of treated units $\widehat{H}$ as those that have at least one treatment indicator (i.e. impulse dummy) retained:

$$i \in \widehat{H} \text{ if } \hat{Q}_i > 0 \tag{38}$$

For practical purposes, this definition could also be made more stringent to differentiate between 'outliers' and actual treatment that can be attributed to potential causes. In other words, we could restrict identification of treatment to some minimum of consecutive impulse dummies. The number of estimated treatment breaks for unit j is given by $\hat{S}_j$ (with $E[\hat{S}_j] = \gamma_c \times T$). The probability of a particular unit in the panel being falsely-classified as ever-treated depends on the number of time series observations for each unit. Consider a panel of N individuals over T time periods. IIS adds NT dummies, with an expected number of retained dummies of $\gamma_c \times NT$. The probability of at least one break per individual will depend on the number of time periods in the sample and the cut-off γ_c. The probability of a particular unit i being spuriously classified as ever-treated is given the probability of at least one observation

of unit i being falsely-classified as treated:

$$P(i \in \widehat{H}|d_i = 0) = 1 - (1 - \gamma_c)^T \tag{39}$$

which increases with T because for larger samples (and fixed γ_c) the probability of retaining an indicator spuriously increases as the number of indicators increases with T. Under the null of no treatment (when in fact no unit is treated), the expected number of falsely detected treated units is then given by:

$$E[\hat{M}] = P(i \in \widehat{H}|d_i = 0) \times N = (1 - (1 - \gamma_c)^T)N \tag{40}$$

If we are worried about the false-positive *rate* of treated units specifically (rather than the false-positive rate γ_c of treatment at any point in time for any unit), it is possible to scale γ_c to ensure a stable false-positive rate of classifying the treatment group. Let p_H denote the target false positive rate of a unit being incorrectly-classified as treated. Then for any target false positive rate p_H, we can choose γ_c as:

$$\gamma_c = 1 - (1 - p_H)^{\frac{1}{T}} \tag{41}$$

This controls the false positive rate of being assigned to the ever treated group to p_H in expectation. For example if $T = 50$, and we aim for a false-positive rate of a single unit incorrectly being classified as treated of 5% i.e. $p_H = 0.05$, then we should set the nominal level of selection to $\gamma_c = 0.001 = 1 - (1 - 0.05)^{\frac{1}{50}}$. Similarly, we could set the target level of significance to maintain a stable *number* of false-positive treated units. If on average we are willing to accept a total of $N_0 = E[\hat{M}]$ false-positive treated units in expectation (where $\hat{M}$ is the estimated number of treated units), the above results imply that:

$$N_0 = (1 - (1 - \gamma_c)^T)N \tag{42}$$

which can be targeted by setting γ_c to:

$$\gamma_c = 1 - (1 - \frac{N_0}{N})^{\frac{1}{T}} \tag{43}$$

and which will yield N_0 expected treated units in expectation when there are in fact no treated units in the true underlying DGP. For example, if we have a panel of $N = 20,\ T = 50$, and we are willing to accept one unit to be falsely-classified as treated on average ($N_0 = 1$), then we can set $\gamma_c = 0.001 \approx 1 - (1 - \frac{1}{20})^{\frac{1}{50}}$. Thus, if we are concerned about the false positive rate of treatment classification, then treatment detection in a panel perhaps warrants tighter target significance levels γ_c than conventionally used in the indicator saturation literature.

If we consider the piece-wise constant treatment effects model matched by step indicators (24), then selecting over treatment variables using the tree search 'gets' is equivalent to applying step-indicator saturation (SIS, Castle et al., 2015) in a fixed effects panel where blocks of steps are included for each individual. SIS uses a near exhaustive tree-search based on a specified level of significance γ_c up to which individual step-functions are removed. The properties of SIS are reasonably well-understood (see Castle et al., 2015; Nielsen & Qian, 2025), and transfer to the panel setting when interpreted as a least-squares dummy variable estimator. The asymptotic properties of SIS under the null of no breaks as $n \to \infty$ are explored in Nielsen & Qian (2025), who show that when there are no breaks (and accounting for multiple testing), the false positive rate of retained breaks (i.e. the number of detected break indicators L_c relative to all possible break variables L) converges to the chosen nominal level of significance of selection γ_c:

$$\hat{\gamma}_c = \frac{L_c}{L} \to \gamma_c, \text{ as } n \to \infty \tag{44}$$

Specifically, if we allow for possible treatment of each unit at every point in time, then – in absence of treatment – the expected value of detected breaks is $\gamma_c \times N(T-1)$ in a balanced panel.[4] This again translates into a probability of being classified as treated as above, albeit with an exponent of (T-1):

$$P(i \in \widehat{H} | d_i = 0) = 1 - (1 - \gamma_c)^{(T-1)} \tag{45}$$

Then for any target false positive rate of being classified as treated p_H, we could choose γ_c as:

$$\gamma_c = 1 - (1 - p_H)^{(1/(T-1))} \tag{46}$$

This matches the properties of IIS except we are searching over $N(T-1)$ rather than NT possible indicators in an exhaustive search since the first indicator would coincide with the fixed effects.

Under the alternative (i.e. in the presence of actual treatment), for simple cases (where the number of variables does not exceed the number of observations), 'gets' has been shown to be a consistent model selection procedure retaining all relevant variables with probability equal to one as $n \to \infty$ (see e.g. Campos et al., 2003). In our setting where the number of variables can exceed the number of observations, we investigate the performance under the alternative (in the presence of structural breaks/treatment) using a range of simulations (see section 6.3). As the selection rule is pre-specified, coefficients on impulse and step-indicators could be bias-corrected to address concerns about post selection inference (see Pretis et al. 2018 for an implementation of bias correction in SIS).

3.1.2 Treatment Detection using the (adaptive) LASSO

As an alternative to the indicator saturation approach, we also propose variants of the Adaptive LASSO (Zou 2006) to identify unknown treatment effects. Unlike the standard LASSO, which applies a uniform penalty to all coefficients and can suffer from selection bias (selectively retaining variables but shrinking their coefficients excessively), the Adaptive LASSO applies data-dependent weights to the penalty term. These weights allow the estimator to enjoy "oracle properties" Huang et al. (2008): asymptotically, the procedure performs as if the true underlying subset of non-zero parameters were known, selecting the correct non-zero coefficients with probability approaching one and estimating them with the same asymptotic distribution as the OLS estimator on the true subset.

We apply the Adaptive LASSO to the fixed effects panel setting following Kock (2013). We partition the parameter set into unpenalized control variables (the unit and time fixed effects, c_i and g_t) and the penalized candidate treatment indicators ($\tau_{j,s}$). The Adaptive LASSO estimator $(\hat{c}, \hat{g}, \hat{\tau})$ is defined as the minimizer of the following objective function:

$$(\hat{c}, \hat{g}, \hat{\tau}) = \underset{c,g,\tau}{\arg\min} \left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{i,t} - c_i - g_t - \sum_{j=1}^{N} \sum_{s=2}^{T} \tau_{j,s} 1_{\{i=j, t \geq s\}} \right)^2 + \lambda \sum_{j=1}^{N} \sum_{s=2}^{T} \hat{w}_{j,s} |\tau_{j,s}| \right]$$

where λ is a non-negative tuning parameter regulating the sparsity of the solution. Crucially, $\hat{w}_{j,s}$ are the adaptive weights derived from an initial consistent estimator, $\tilde{\tau}_{j,s}$. Typically, $\tilde{\tau}$ is obtained via a standard LASSO (or Ridge) regression in a first step ((Kock, 2016)). The weights are constructed as $\hat{w}_{j,s} = 1/|\tilde{\tau}_{j,s}|^{\gamma}$ for some $\gamma > 0$ (commonly $\gamma = 1$). This weighting scheme ensures that candidate

[4] Albeit simulation results show a higher false-positive rate for SIS in small samples, warranting perhaps a more conservative choice of γ_c.

treatments with small coefficients in the initial step receive larger penalties, driving them to zero, while strong signals receive smaller penalties, reducing the estimation bias.

The choice of the penalty parameter λ is critical. While K-fold Cross-Validation (CV) is widely used to minimize prediction error, it is known to be inconsistent for variable selection, often retaining too many irrelevant variables (over-selection) in finite samples. In the context of identifying specific policy interventions or treatment episodes, our primary goal is structural discovery (consistent selection) rather than pure prediction. Therefore, we recommend the use of the Bayesian Information Criterion (BIC) to select λ. As shown in Zou (2006) and Wang et al. (2009), minimizing the BIC in the Adaptive LASSO framework leads to consistent variable selection.

A further promising approach in our setting would be to use the thresholded LASSO (Zhou 2010) to further reduce the false discovery rate. The threshold LASSO uses an initial estimator (e.g. LASSO or adaptive LASSO), then apply a threshold post-selection where only coefficients greater than some specified threshold are retained. Subsequently the model is re-estimated (using either OLS or the LASSO again).

A major challenge in penalized regression is conducting valid inference on the selected parameters. Standard errors calculated directly from the LASSO estimates are biased due to the shrinkage and selection effects. To address this, we adopt the 'post-model-selection estimator' or 'naive' approach: we first use the Adaptive LASSO to select the set of non-zero treatment indicators, and then re-estimate the selected model using standard OLS (or the standard Within estimator).

While naive post-selection inference is generally cautioned against (see e.g. (Leeb & Pötscher, 2008)), the specific properties of the Adaptive LASSO provide a theoretical justification for this approach in our context. Conditional on the oracle property holding—which implies the probability of selecting the exact true model converges to one—the post-selection OLS estimator is asymptotically efficient Huang et al. (2008). Furthermore, in finite samples, Zhao et al. (2017) demonstrate that this 'naive' approach of refitting OLS to the selected support often outperforms complex inference adjustments. This two-step procedure allows us to report standard errors and confidence intervals for the detected treatment effects that are easily interpretable by applied researchers.

3.2 Embedding Known Interventions

There are two ways in which break detection to identify treatment can be implemented: either as a fully agnostic way to detect entirely unknown treatment assignment and timing, or as a robustness check embedding known treatment and searching for additional previously-unknown interventions. Above we outlined the case where we detect treatment as a purely agnostic data-driven approach to identify interventions without any prior knowledge of their occurrence. While the approach is agnostic and any unit may be treated at any point in time, a potential downside is a loss in power if treatment assignment and timing is known and there are multiple treated units with a homogeneous treatment effects, since each treated unit would have to be identified individually.

If treatment assignment and timing is known for a particular intervention then break detection can be adapted as a robustness check for additional unknown treatment in conventional TWFE difference-in-differences models. In this case we force the known treatment dummy (or dummies for interactions) to be included in the model, and select over additional treatment indicators. This corresponds to the theory-embedding approach of Hendry & Johansen (2015) where fixed regressors are embedded in a wider information set that we select over.

Then selection takes place over the break variables to detect additional treatment (omitting the break variables perfectly coinciding with known treatment dummies), known treatment dummies remain in the model without being selected over. This allows additional unknown treatment to be detected, while the coefficient on the forced (not-selected-over) break variable yields an estimate of the conventional treatment effect in a TWFE panel. For example, Stechemesser et al. (2024) embed the introduction of the EU emissions trading scheme as a known intervention for multiple countries and search for additional breaks (ie. treatment effects) in CO_2 emissions.

It is worth highlighting that we do not necessarily need to allow for a break at every point in time (or individual). If there is a strong reason that the break should be localised in particular time periods (or among particular individuals), then only those could be included in the candidate set of break variables selected-over.

3.3 Ex-Post Attribution of Detected Events

Having identified treatment as structural breaks, the second stage of our framework is the ex-post attribution of these detected shifts to potential causes. This step transforms policy evaluation from testing a single known hypothesis into a process of generating new hypotheses about what mattered. While the detection phase is purely data-driven with the properties of selection algorithms reasonably well-understood, attribution requires contextual knowledge to link statistical breaks to real-world events. The rigor and methodology of this attribution process have evolved significantly in recent applications.

Early applications (notably in climate policy evaluation) of this framework relied on unstructured attribution. For example, Pretis (2022) utilized a combination of subject-specific knowledge and general archival searches to attribute detected breaks in British Columbia's carbon emissions to the introduction of the provincial carbon tax. This approach relies on the researcher's domain expertise to identify plausibly exogenous events that coincide temporally with the detected break. While effective for single-unit case studies, unstructured attribution can be difficult to scale and may be susceptible to confirmation bias if search criteria are not transparent.

To address these limitations, subsequent work has adopted structured attribution protocols. Koch et al. (2022) systematized the search process in their analysis of road transport emissions. They established ex-ante search criteria—specifically looking for major policy reforms or fuel price shocks within a defined temporal window of the detected break—thereby reducing researcher degrees of freedom. This moves the analysis toward a more falsifiable standard where potential causes must meet pre-defined relevance criteria.

Most recently, the method has advanced to systematic database matching. Stechemesser et al. (2024) demonstrated this by constructing a comprehensive database of climate policies from the OECD and other international bodies. By algorithmically matching detected emission breaks across 41 countries to known policy implementation dates in the database, they were able to conduct "success attribution" on a global scale. This approach allows for the identification of successful policy mixes without prior assumptions about which specific policies would work, effectively using the break detection algorithm to query the policy database for effective interventions.

Looking forward, Large Language Models (LLMs) may offer useful tools to further automate and scale the unstructured search process. LLMs can scan vast corpora of news, legislative records, and historical documents to propose candidate events that coincide with detected breaks, particularly in contexts where structured policy databases do not yet exist.

However, we emphasize that regardless of the method—whether manual, database-driven, or AI-assisted, the attribution stage remains hypothesis generating. The detected break provides the empirical evidence that a structural change occurred; the attribution proposes a hypothesis for why. This generated hypothesis can then be subjected to further testing, for example, by treating the attributed event as a "known" intervention in a standard Difference-in-Differences framework to verify its robustness and causal validity.

4 Illustrative Applications

We illustrate our method with two distinct applications. We purposely choose two well-known examples to illustrate our methods, and specifically show that we can recover known treatment effects without prior knowledge.[5] First we turn to climate policy evaluation where the "reverse causal" question - what actually reduced emissions? - is often more pertinent than testing the effect of a single known measure. We demonstrate the method's utility for climate econometrics by replicating and expanding the analysis of the Swedish carbon tax on transport emissions (Andersson 2019). We show that our data-driven framework recovers the correct treatment start date and effect magnitude without prior knowledge, validating its potential to "discover" effective climate policies in settings where they are unknown.

Second, as a validation exercise against a canonical benchmark in the causal inference literature, we apply our framework to the economic impact of ETA terrorism in Spain (Abadie & Gardeazabal 2003), demonstrating that we can recover the standard treatment effect estimates in a purely agnostic manner.

4.1 Application I: Detecting the Effect of Carbon Taxes in Sweden

Evaluating climate policy effectiveness is challenging because we are regularly faced with a myriad of policies introduced across many countries at different times with uncertain effects (see e.g. Stechemesser et al. 2024). Traditional "forward causal" approaches require the researcher to pre-specify the exact treatment date and unit. However, macro-economic outcomes such as emissions are often affected by a wide range of possible interventions.

We apply our break detection approach to the well-known case of the Swedish carbon tax. Andersson (2019) used synthetic controls to show that the introduction of a carbon tax in Sweden in 1991 led to a significant reduction in CO_2 emissions from transport, estimated at approximately 0.29 metric tons per capita. Here, we ask the reverse question: Could we have identified this reduction—and attributed it to the correct year—without knowing *a-priori* that a tax was introduced?

We model annual CO_2 emissions (in metric tons) per capita from transport for OECD countries, consistent with the data used in Andersson (2019), covering the period 1960–2005 (ending before the introduction of the European Emissions Trading scheme).

4.1.1 Detecting "What Mattered" without Prior Knowledge

To illustrate the proposed methods starting with the simplest setup, we begin with a two-country panel ($N = 2$) over 46 years ($T = 46$) consisting of Sweden (the treated unit) and Belgium, the latter selected as one of the primary controls following Andersson's optimal synthetic control weights. We first estimate the 'unknown treatment' model using step-indicator saturation (SIS) via the getspanel

[5] We provide policy-focused application with novel data in our closely-related papers in Pretis (2022), Koch et al. (2022) and Stechemesser et al. (2024).

package, selecting breaks at a tight target significance level of $\gamma_\alpha = 0.001(0.1\%)$ as well as the adaptive LASSO. The general model takes the form:

$$CO_{2i,t}pc = c_i + g_t + \sum_{j \in \{\text{SWE, BEL}\}} \sum_{s=1961}^{2005} \tau_{j,s} 1_{\{i=j, t \geq s\}} + \beta X_{it} + \epsilon_{it} \tag{47}$$

where we allow both Sweden and Belgium to be potentially-treated at every point in time. The vector of control variables $X_{i,t}$ includes GDP per capita, the share of urban population, and population density. For getspanel, setting $\gamma_\alpha = 0.001(0.1\%)$ implies that we expect 0.1% of the observations to be spuriously labelled as a break by chance under the null hypothesis of no break. For the small two-country panel this means we expect $0.001 \times (92 - 2) = 0.090 < 1$ expected breaks.

We compare this unknown treatment setting to the benchmark of a standard TWFE difference-in-differences specification where the treatment is known and imposed:

$$CO_{2i,t}pc = c_i + g_t + \tau_{\text{known}} 1_{\{i=\text{SWE}, t \geq 1991\}} + \beta X_{it} + \epsilon_{it} \tag{48}$$

Without imposing any known intervention date, we detect a significant negative structural break in Sweden's fixed effect starting in 1991 (see Figure 2 and Table 1). The estimated coefficient on the detected step-indicator is -0.247 (se = 0.046) metric tons. The resulting selected model is *identical* (in absence of any bias-correction due to selection) to the TWFE estimator with the known treatment intervention imposed (i.e. with a difference-in-difference TWFE specification for the introduction of the Swedish carbon tax), with the estimated coefficient on the retained break variable of -0.247 (se = 0.046) matching the estimated treatment effect in the TWFE difference-in-differences model (see Table 1). This estimate is also statistically indistinguishable from the average reduction of -0.29 metric tons of CO_2 per capita reported by Andersson (2019, albeit using synthetic controls in his application).

Crucially, the break date is identified directly from the data. The Swedish carbon tax was legislated in 1990 and implemented in January 1991. Our method detects the shift exactly at this juncture, validating the capability of the reverse-causal approach to correctly pinpoint the timing of effective climate policy. This "discovery" suggests that in settings where policy portfolios are complex, break detection can serve as a powerful first step to identify which interventions within a mix actually coincided with structural shifts in emissions.

When we repeat the analysis using the adaptive LASSO where we penalise the possible treatment coefficients τ, with penalty weights chosen using the simple LASSO as an initial estimator, and the tuning parameter selected using BIC. The adaptive LASSO estimates are reported using the 'naive' approach of re-estimating the selected model using OLS (see e.g. Zhao et al., 2021) and shown in Table 1.

Using the LASSO we also detect the large negative break in Sweden in 1990. However, the adaptive LASSO further retains many additional breaks in the series, highlighting the difficulty of controlling the false-positive rate when using this method compared to the indicator saturation approach (gets) where the significance level can be explicitly targeted. Nevertheless, even in the LASSO specification, the 1990 break in Sweden remains the largest detected shift, again allowing for the correct identification of the policy effect (particularly if we imposed a minimum effect threshold as suggested in the threshold adaptive LASSO – Zhou (2010)).

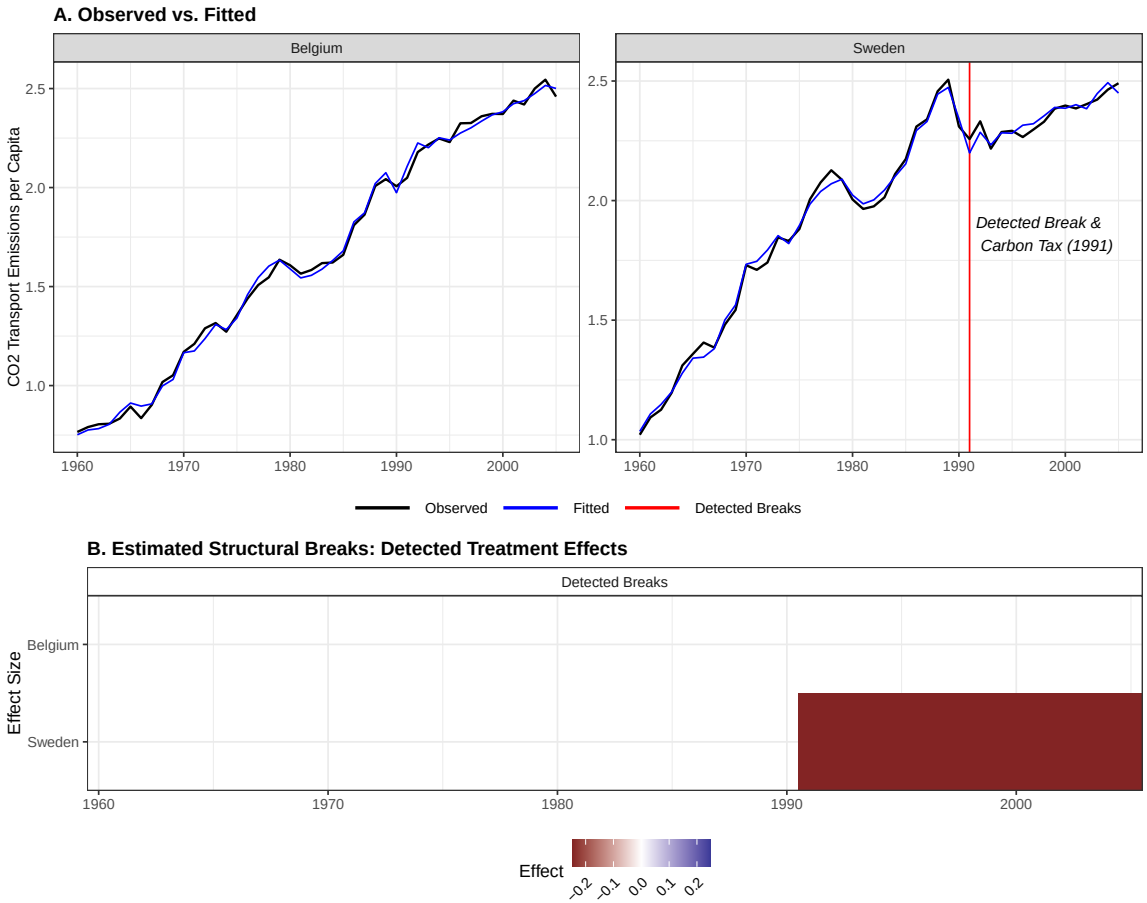

Figure 2: Top: Transport CO2 Emissions per Capita, Belgium & Sweden – Treatment detected using step-indicators in 'getspanel' for a target significance level of $\gamma_c = 0.001$. Red vertical lines denote detected impulses (identifying time-varying treatment effects). Bottom: detected structural breaks and estimated treatment effect (coefficient on break indicators).

Table 1: Table 1: Carbon Tax Effects

	Two Countries (Sweden, Belgium)			Multi-Country (7 Countries)	
	GETS	Lasso	Known	GETS	Known
Sweden (1973, Step)		0.020			
		(0.042)			
Sweden (1976, Step)		-0.001			
		(0.041)			
Sweden (1990, Step)		-0.168***		-0.241***	
		(0.049)		(0.035)	
Sweden (1991, Step)	-0.248***		-0.248***		-0.391***
	(0.046)		(0.046)		(0.059)
Sweden (1993, Step)		-0.139***			
		(0.049)			
Sweden (1996, Step)		-0.057			
		(0.042)			
UnitedStates (1968, Step)				0.645***	
				(0.058)	
UnitedStates (1972, Step)				0.403***	
				(0.058)	
UnitedStates (1980, Step)				-0.594***	
				(0.042)	
Greece (1990, Step)				0.232***	
				(0.031)	
NewZealand (1994, Step)				0.377***	
				(0.039)	
GDP per Capita	1.96e-05**	2.58e-05**	1.96e-05**	2.78e-05***	7.97e-06
	(8.86e-06)	(9.91e-06)	(8.86e-06)	(5.00e-06)	(6.69e-06)
Pop Density	0.015***	0.009	0.015***	0.001	-0.005*
	(0.003)	(0.006)	(0.003)	(0.002)	(0.003)
Urban Pop	0.068***	0.050***	0.068***	-0.008***	-0.011**
	(0.010)	(0.014)	(0.010)	(0.002)	(0.005)
Observations	92	92	92	322	322
AIC	-295.0	-318.2	-297.0	-601.4	-218.6
Country/Region FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Countries (N)	2	2	2	7	7
Time Periods (T)	46	46	46	46	46

* p ¡ 0.1, ** p ¡ 0.05, *** p ¡ 0.01

4.1.2 Robustness in a Multi-Country Panel

We now expand the panel to include a larger set of possible control countries ($N = 7$, $T = 46$): Denmark, Belgium, New Zealand, Greece, Switzerland, and the United States.[6] We again detect breaks using *getspanel* at $\gamma_c = 0.001$, this implies an expected false positive number of breaks of $0.001 \times (322 - 7) = 0.315 < 1$. We can also compute the expected number of incorrectly-classified units as treated as $P(i \in \hat{H}|d_i = 0) \times N = (1 - (1 - 0.001)^{(46-1)}) \times 7 = 0.31$, which suggestes that we expect to detect less than one untreated unit as treated.

Running *getspanel* on this expanded panel ($N = 7, T = 46$), we again identify a negative structural break for Sweden beginning in 1991 (see Figure 3). The estimated treatment effect remains robust, with a magnitude of -0.241 (se = 0.035) metric tons.

Our method also detects breaks in other countries during the sample period, highlighting the hypothesis-generating nature of the framework. For instance, in the United States, we detect a negative break in 1980, which plausibly corresponds to the combined effects of the 1979 oil crisis and the tightening of CAFE fuel efficiency standards. Conversely, positive breaks detected in the late 1960s align with the post-war boom in vehicle ownership. These results underscore the utility of the framework not just for validating known taxes, but for generating plausible hypotheses about structural shifts in global

[6]With countries chosen as those who received a weight of more than 5% in Andersson's synthetic control analysis.

emissions trajectories.

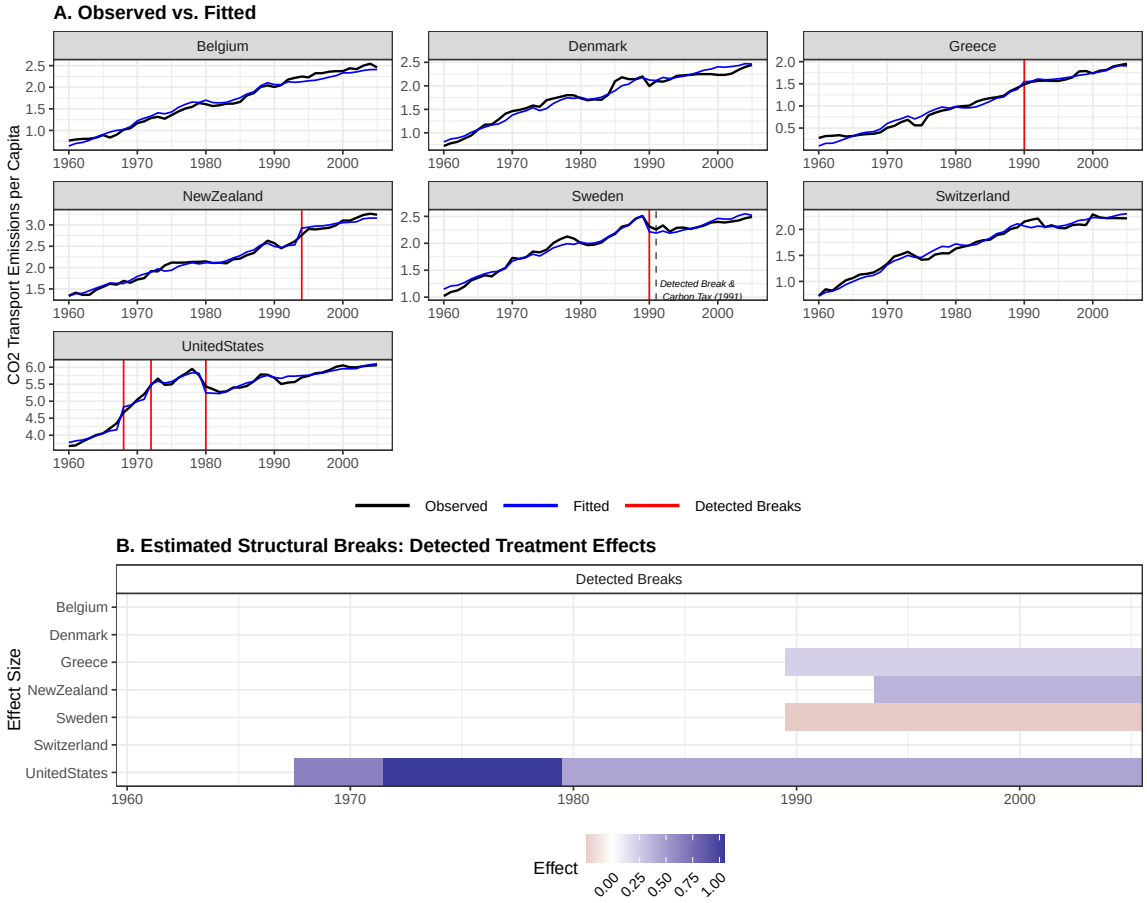

Figure 3: Top: Transport CO2 Emissions per Capita with multiple potential control countries – Treatment detected using step-indicators in 'getspanel' for a target significance level of $\gamma_c = 0.001$. Red vertical lines denote detected impulses (identifying time-varying treatment effects). Bottom: detected structural breaks and estimated treatment effect (coefficient on break indicators).

4.2 Application II: Detecting the Impacts of Terrorism on GDP per Capita

To further validate that our agnostic detection method recovers estimates consistent with standard causal inference designs, we also replicate the analysis of Abadie & Gardeazabal (2003) on the economic impact of ETA terrorism. This application serves as a further "ground truth" test: we know the intervention (terrorism) occurred, and we wish to see if our algorithm can find it without being told.

The dataset for our illustrative application here spans all of mainland Spain's 15 regions (where we exclude the Canary and Balearic Islands) over 31 years from 1965 to 1995 for a total of 465 region-year observations. In their seminal paper, Abadie & Gardeazabal (2003) used a forward causal approach to study the effect of ETA terrorism on regional economic output. The authors find a substantial reduction in regional GDP in response to local terrorism introducing synthetic control methods. Here we ask the reverse causal question: what affected regional GDP per capita in the Basque Country (or wider Spain)? We show that the "treatment" taking the form of ETA terrorism (alongside a number of other previously unidentified treatments) can be detected without prior knowledge of its occurrence using our proposed break detection approach.

Again we first consider a simple TWFE panel setting with two regions (the Basque Country and Madrid) where we search for breaks to detect treatment in GDP per capita.[7] We then expand this into a multi-region panel of mainland Spain to assess breaks in a wider context. Our results show that we can detect the effect of ETA terrorism without prior knowledge of its occurrence and obtain treatment effect estimates that are near-identical to a known-treatment model. Our break detection approach also provides evidence that the treatment effects of GDP impacts of ETA terrorism were transitory and are no longer detectable post-1990. In addition, in the panel with more than two regions we also detect breaks which we attribute to an industrial crisis and increased autonomy following the Franco era in other regions.

4.2.1 Detecting Treatment in a Panel with Two Regions

We first consider a simple panel with two regions: the Basque country and Madrid ($N = 2, T = 31$, $NT = 62$). For comparison, we initially estimate the forward causal 'infeasible' model of log GDP per capita (controlling for log investments Inv similar to Abadie & Gardeazabal 2003) using a TWFE estimator with a known intervention of Basque terrorism to provide a baseline relative to our break detection approach. We then demonstrate that we can directly detect the terrorism 'treatment' without prior knowledge using our reverse causal approach.

As a baseline, consider a TWFE estimator with a 'known' intervention of ETA terrorism. We estimate baseline models first allowing for time-varying treatment effects using interactions in (49), then assuming time-invariant treatment effects in (50) specified as a dummy variable for the Basque region in the 'post-treatment' period, defined here as 1979 onwards, as Abadie & Gardeazabal (2003) found that the impact of terrorism was notable in GDP per capita from the end of the 1970s.

'Known' Treatment (fully time-varying treatment effects):

$$\begin{aligned} log(GDPpc_{i,t}) &= \alpha_i + \phi_t + \sum_{s=1979}^{1995} d_i \tau_s 1_{\{t=s\}} + \beta_1 log(Inv)_{i,t} + u_{i,t} \\ &\text{where } d_i = 1_{\{i=\text{Basque}\}} \end{aligned} \tag{49}$$

'Known' Treatment (time-constant treatment effects):

$$\begin{aligned} log(GDPpc_{i,t}) &= \alpha_i + \phi_t + d_{i,t}\tau + \beta_1 log(Inv)_{i,t} + u_{i,t} \\ &\text{where } d_{i,t} = 1_{\{i=\text{Basque}, t\geq 1979\}} \end{aligned} \tag{50}$$

Estimation results for the 'known' baseline models are shown in Tables 2 and 3, under the columns "Known TWFE". The results of the known baseline show an approximate 5% reduction in GDP per capita in the Basque country relative to Madrid in response to ETA terrorism in this simple two-region model. This result is similar across the time-varying model (see equation 19) (where the estimated ATT is given by the average of the impulse coefficients) as well as the piece-wise constant treatment effects model (see equation 24). Specifically, the ATT across impulses in the known baseline in (49) is -0.0496 (se=0.0197), and the time-constant estimate given by the coefficient in (50) on the known step-function is -0.0495 (se=0.006).

Now suppose the "treatment" of ETA terrorism in the Basque country was unknown, and we approached the data with our reverse causal question of 'what affected GDP per capita?

[7] For completeness we also show that a simple time series model of Basque GDP per capita is unable to identify ETA terrorism impacts due to the lack of control groups – see Supplementary Material 6.4).

Unknown Treatment with Fully Time-Varying Effects We now estimate a model allowing for the potential treatment of any unit at any point in time first using impulse dummies capturing time-varying treatment and select over them using the 'gets' selection algorithm (we consider the LASSO for the piece-wise constant setting below). The model is saturated with a full set of impulse dummies in (51) which are selected over at a target level of significance γ_c. We consider three different target significance levels, $\gamma_c = 0.05$ as well as 0.025 and 0.01 to illustrate the impact of the calibration choice on treatment detection.

'Unknown' Treatment:

$$log(GDPpc_{i,t}) = c_i + g_t + \sum_{j=1}^{2} \sum_{s=1966}^{1995} \tau_{j,s} 1_{\{i=j,t=s\}} + \beta_1 log(Inv)_{i,t} + u_{i,t} \tag{51}$$

The resulting detected impulses, which we interpret as unit-specific time-varying treatment effects, are shown in Figure 4 (for $\gamma_c = 0.05$) and Table 2 (for all three values of γ_c). We detect the treatment of Basque terrorism without prior knowledge of its occurrence as individual impulses in the Basque region from 1980 to 1990. Each coefficient provides an estimate of the unit- and time-specific treatment effect. We can easily compute our estimates of the ATT by taking the mean of the impulses over time. Standard errors for the ATT are also straight-forward to compute as impulses are orthogonal. Computing the ATT over the time period from 1980 to 1990 from the model with $\gamma_c = 0.05$ yields an estimate of the ATT of -0.059 (se=0.016) which is nearly identical (and not significantly different) to the known-treatment baseline estimate of -0.0496 (se=0.0197). The fact that the ATT using detected impulses is marginally larger than the known baseline ATT can be explained by the fact that the impulses are only retained up to 1990 while the 'known' baseline time-varying treatment considers treatment effects up until the end of the sample in 1995. Indeed, we only detect treatment breaks up until 1990, suggesting that the impacts of ETA terrorism on GDP were transitory and no longer detectable post-1990. This is consistent with the known-treatment baseline which finds predominantly insignificant time-varying treatment effects after 1990.

Varying γ_c, we successfully detect the intervention at relatively loose levels of significance $\gamma_c = 0.05$ or $\gamma_c = 0.025$. The loss of power for more conservative levels of the target false positive rate becomes apparent when we set $\gamma_c = 0.01$, where we do not detect any treatment as impulse dummies coinciding with ETA terrorism. However, this reduction in power can be tackled by specifying piece-wise constant treatment effects using step functions as we demonstrate in the following section 4.2.1. Note that in this N=2 panel, the treatment effects are relative to the single control region and one could achieve the same detected treatment if Basque country was selected as the 'control', in which case the treatment effects would be detected for Madrid and opposite-signed. We would then interpret them as the effect of the absence of terrorism.

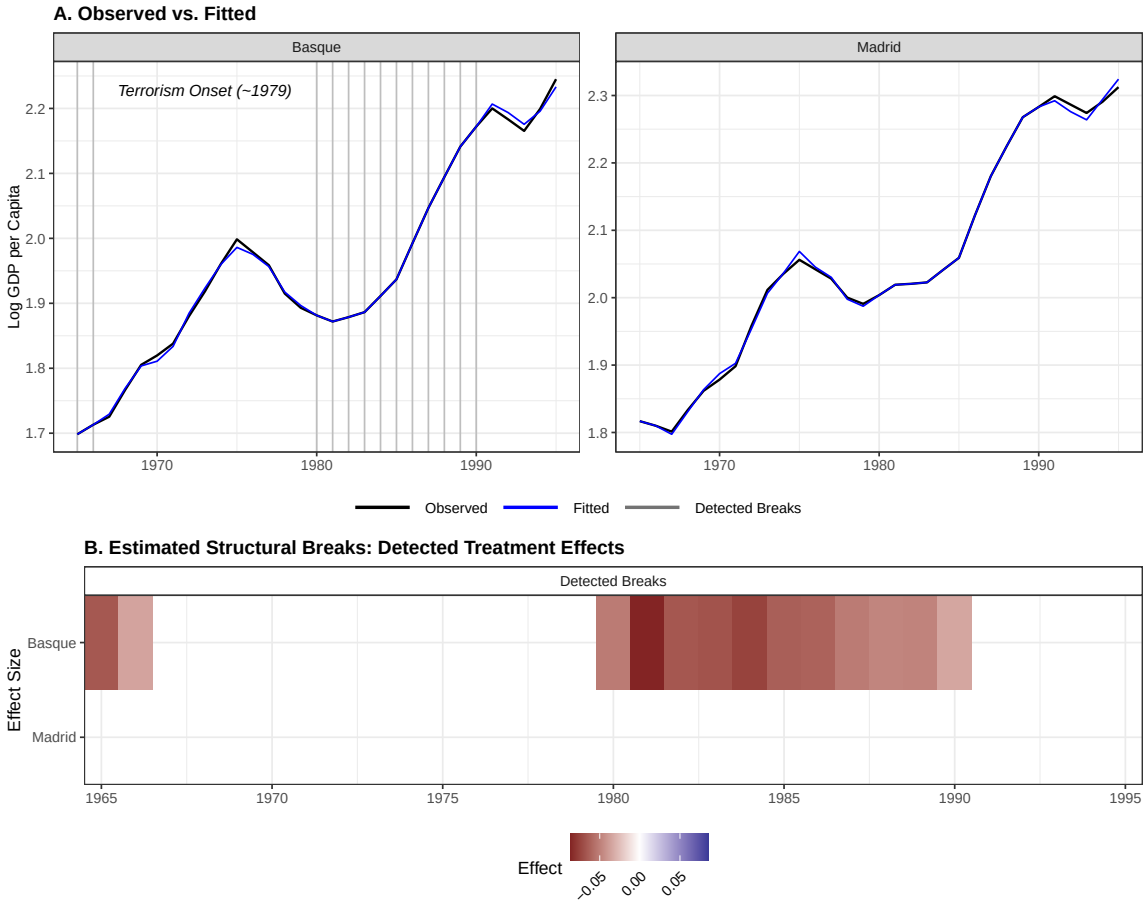

Figure 4: TWFE Panel: GDP per Capita, Basque Country & Madrid – Time-Varying Treatment detected using IIS in 'gets' for a target significance level of $\gamma_c = 0.05$. Red vertical lines denote detected impulses (identifying time-varying treatment effects).

Table 2: Table A: Impulse Indicators (Basque)

	Two Regions			
	GETS 0.05	GETS 0.025	GETS 0.01	Known
Basque (1965, Impulse)	-0.066***	-0.056***		
	(0.016)	(0.019)		
Basque (1966, Impulse)	-0.035**			
	(0.015)			
Basque (1979, Impulse)				-0.035
				(0.021)
Basque (1980, Impulse)	-0.051***	-0.046**		-0.047**
	(0.014)	(0.017)		(0.018)
Basque (1981, Impulse)	-0.086***	-0.078***		-0.066***
	(0.015)	(0.018)		(0.019)
Basque (1982, Impulse)	-0.066***	-0.062***		-0.070***
	(0.014)	(0.017)		(0.019)
Basque (1983, Impulse)	-0.068***	-0.062***		-0.060***
	(0.014)	(0.017)		(0.018)
Basque (1984, Impulse)	-0.074***	-0.065***		-0.046**
	(0.016)	(0.019)		(0.020)
Basque (1985, Impulse)	-0.063***	-0.055***		-0.040*
	(0.015)	(0.018)		(0.019)
Basque (1986, Impulse)	-0.062***	-0.056***		-0.053**
	(0.014)	(0.017)		(0.018)
Basque (1987, Impulse)	-0.051***	-0.049**		-0.066***
	(0.014)	(0.017)		(0.019)
Basque (1988, Impulse)	-0.048***	-0.046**		-0.063***
	(0.014)	(0.017)		(0.020)
Basque (1989, Impulse)	-0.048***	-0.045**		-0.056**
	(0.014)	(0.017)		(0.019)
Basque (1990, Impulse)	-0.034**			-0.039*
	(0.014)			(0.019)
Basque (1991, Impulse)				-0.033
				(0.020)
Basque (1992, Impulse)				-0.035*
				(0.019)
Basque (1993, Impulse)				-0.044*
				(0.021)
Basque (1994, Impulse)				-0.033
				(0.024)
Basque (1995, Impulse)				-0.005
				(0.021)
Linvest	0.105***	0.078*	-0.033	-0.064
	(0.033)	(0.038)	(0.049)	(0.056)
Observations	62	62	62	62
AIC	-388.3	-360.6	-285.0	-369.2
Country/Region FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Countries (N)	2	2	2	2
Time Periods (T)	31	31	31	31

* p ¡ 0.1, ** p ¡ 0.05, *** p ¡ 0.01

Table 3: Table B: Step Indicators (Basque)

	Two Regions				Multi-Region	
	GETS 0.01	GETS 0.001	Lasso	Known	GETS 0.0001	Known
Basque (1967, Step)			0.032**			
			(0.013)			
Basque (1968, Step)			0.008			
			(0.011)			
Basque (1978, Step)			-0.017		-0.156***	
			(0.011)		(0.012)	
Basque (1979, Step)	-0.040***	-0.049***	-0.012	-0.049***		-0.155***
	(0.012)	(0.006)	(0.014)	(0.006)		(0.018)
Basque (1980, Step)			-0.035***			
			(0.012)			
Basque (1981, Step)	-0.018					
	(0.012)					
Basque (1990, Step)	0.028***		0.030***			
	(0.009)		(0.007)			
Basque (1995, Step)			0.035***			
			(0.011)			
Madrid (1970, Step)					-0.126***	
					(0.018)	
Castilla LaMancha (1972, Step)					0.117***	
					(0.014)	
Galicia (1976, Step)					0.098***	
					(0.012)	
Rioja La (1981, Step)					0.080***	
					(0.012)	
PrincipadoDeAsturias (1986, Step)					-0.122***	
					(0.012)	
Extremadura (1987, Step)					0.135***	
					(0.013)	
Madrid (1990, Step)					-0.090***	
					(0.015)	
Linvest	-0.054*	-0.107***	0.002	-0.107***	0.117***	0.138***
	(0.031)	(0.029)	(0.026)	(0.029)	(0.012)	(0.018)
iis125 (NA, Impulse)					0.091**	
					(0.036)	
Observations	62	62	62	62	465	465
AIC	-370.6	-355.8	-419.5	-357.8	-1864.7	-1468.4
Country/Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Countries (N)	2	2	2	2	15	15
Time Periods (T)	31	31	31	31	31	31

* p ¡ 0.1, ** p ¡ 0.05, *** p ¡ 0.01

Unknown Treatment with Piece-Wise Constant Effects Treatment effects may be piece-wise constant and thus detected with greater likelihood (due to the higher power of step-functions). To illustrate this, we estimate a TWFE panel (52) saturated with a full set of step-functions denoting potential treatment in either region (Basque or Madrid) at any point in time:

'Unknown' Treatment:

$$log(GDPpc_{i,t}) = c_i + g_t + \sum_{j=1}^{2} \sum_{s=1966}^{1995} \tau_{j,s} 1_{\{i=j, t \geq s\}} + \beta_1 log(Inv)_{i,t} + \epsilon_{i,t} \tag{52}$$

We select over treatment functions using 'gets' at two target levels of $\gamma_c = 0.001$ and $\gamma_c = 0.01$ as well as the adaptive LASSO as before.

Table 3 and Figure 5 shows the results of break detection. Detecting treatment using 'gets' at $\gamma_c = 0.001$ results in a single treatment indicator being retained for the Basque Country from 1979 onwards. In other words – without knowing that treatment occurred – we are able to detect the treatment intervention and estimate a model effectively *identical* to the known intervention panel. Similarly, the adaptive LASSO is able to identify treatment (detecting a negative intervention in Basque country in 1980), with the estimated 'naive' post-LASSO treatment effect near identical to the 'known' imposed intervention in 1979. The adaptive LASSO further detects additional earlier breaks which is unsurprising as it can be often less conservative than 'gets' with low levels of γ_c. Relaxing the target level of γ_c to a less conservative level of 0.01 results in additional breaks being detected which can be interpreted as time-varying treatment effects: the negative break in 1981 suggests that the initial impact of ETA terrorism became larger in the early 1980s, however, the opposite-signed break in 1990 provides evidence of the transitory nature of the impact. Consistent with our results from the fully-time-varying specification (and known baseline), treatment effects post-1990 are closer to zero (see section 4.2.1).

Overall, both 'gets' and the adaptive LASSO implementation of our proposed break detection approach detect the 'treatment' without prior knowledge of its occurrence. Break detection estimates to detect treatment suggest a roughly 5% reduction in GDP per capita in response to terrorism in the Basque region relative to Madrid as the control region, which is identical to the known intervention TWFE estimator. Further, it is worth noting that the break detection approach suggests a reduction in GDP per capita from around 1979/1980 onwards, which is consistent with Abadie and Gardeazabal's finding that GDP per capita reductions occurred with a lag relative to the onset of terrorism in the mid 1970s.

Thus, not only are we able to detect treatment without prior knowledge on which regions were treated and when treatment occurred, but the estimated break dates also provide insights into the lagged onset of the economic impacts of terrorism.

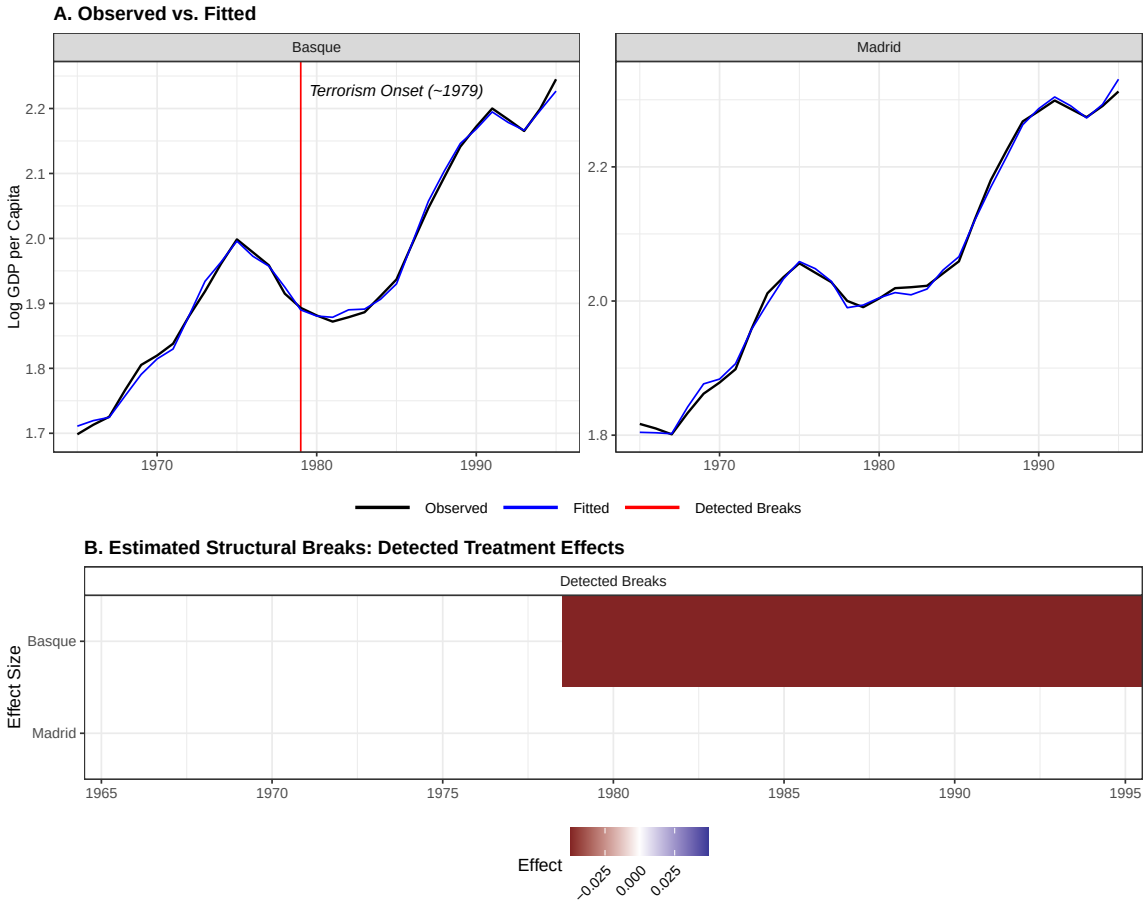

Figure 5: TWFE Panel: GDP per Capita, Basque Country & Madrid – Treatment detected using SIS in 'gets' at $\gamma_c = 0.001$. Red vertical lines denote detected step-shifts (identifying treatment effects).

4.2.2 Detecting Treatment in a Panel with Multiple Regions

We repeat the above analysis for a panel covering all of mainland Spain using 'getspanel'. We now include all $N = 15$ regions of mainland Spain over $T = 31$ years (for a total sample size of $NT = 465$). Just as before, we compare the detected treatment in this larger panel to the benchmark of a known intervention by imposing the 'treatment' as a dummy variable for the Basque region from 1979 onwards in a TWFE estimator. The 'known treatment' baseline yields an estimated treatment effect of -0.155 (se=0.018) relative to the control regions in wider Spain (see Table 5).[8]

Our break detection results using gets at $\gamma_c = 0.001$ show that even in this more general setting we are able to detect the treatment of ETA terrorism through the impacts on GDP in the Basque Country without prior knowledge of its occurrence (see Figure 6 and Table 5). The ETA treatment is detected in 1978 (close to the imposed intervention in the known TWFE estimator in 1979) with an estimated treatment effect of -0.156 (se=0.012) which is *near identical* to the 'known treatment' benchmark.

In addition to the ETA break in 1978, we also detect a small number of possible treatment effects through breaks in the fixed effects of other regions.[9] It is worth noting though that the break associated with the ETA 'treatment' is the single largest break in magnitude compared to all detected breaks. Given the set of detected effects (captured through breaks) for some of the regions, the next step of our approach (see section 3.3) is to investigate the relevant literature for potential causes.

A brief review of the literature on Spanish economic history suggests that the positive breaks (i.e. positive treatment effects on GDP per capita) in Extremadura, Galicia, and Rioja, may correspond to the increased autonomy of the regions awarded in the post-Franco era. The negative break in Madrid in 1970 coincides with an industrial crisis that hit Madrid disproportionally relative to other regions (Rodríguez-Pose & Hardy 2021, and Tobío 1989).

The fact that ex-post attribution is not always straightforward is highlighted by the fact that we have yet to identify likely causes for the positive break in Castilla-La Mancha in 1972 (though it is worth noting that the film adaptation of the highly popular musical "Man of la Mancha" was released in that year), and the negative breaks in Asturias (in 1986) and Madrid (in 1990).

[8] This estimate in the known benchmark and the detected break setting is larger than the two-region panel because the control group is different. The two-region panel only included Madrid as a control region)

[9] To control for outlying observations we also combine our selection over step functions with selection over impulse dummies, where impulses could capture outliers or can also be interpreted as single-period time-varying treatment indicators. Only a single outlying observation is identified: Madrid, 1965.

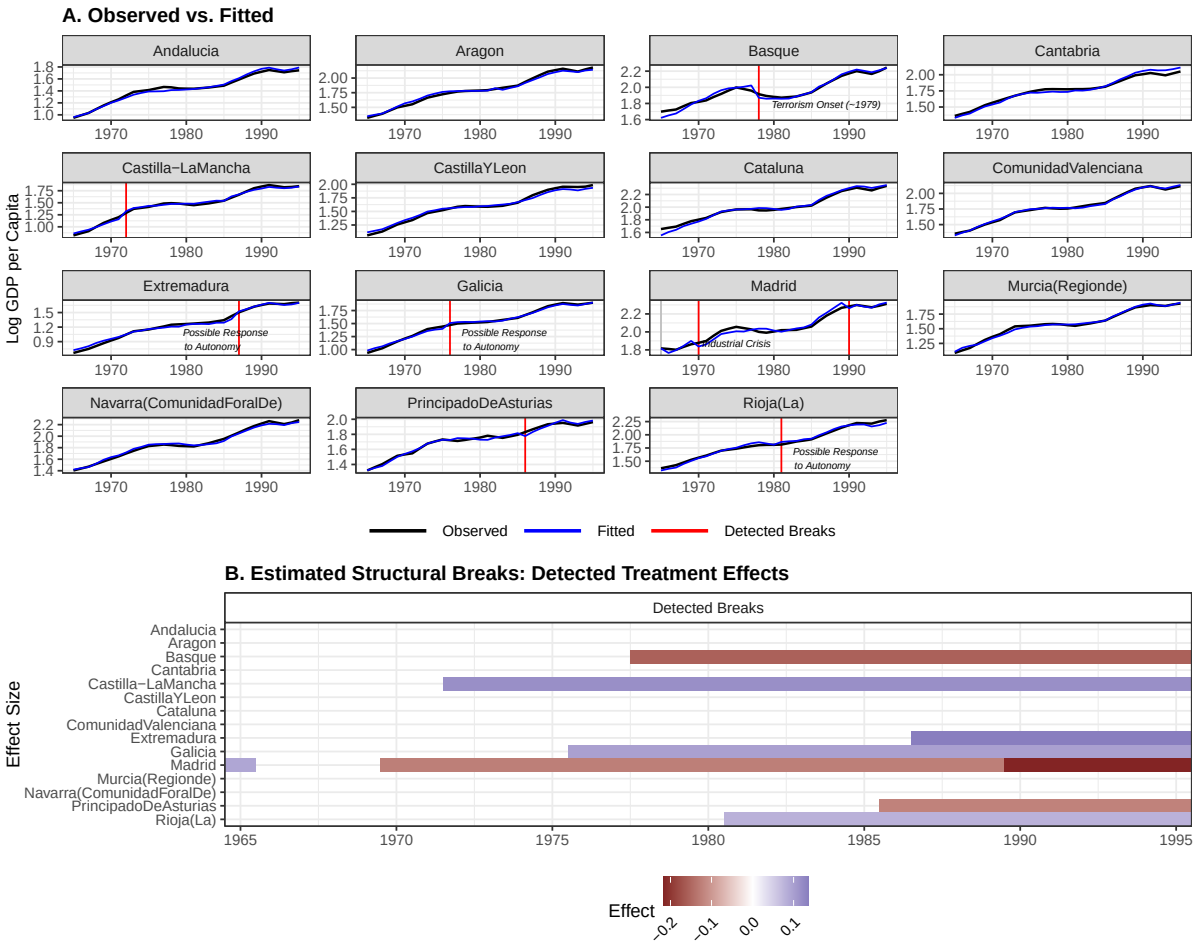

Figure 6: TWFE Panel: GDP per Capita in 15 Regions of Mainland Spain – Treatment detected using 'gets' and $\gamma_c = 0.001$. Red vertical lines denote detected step-shifts (identifying treatment effects).

5 Conclusion

We operationalise the discovery of treatment assignment and timing by searching for structural breaks in fixed effects panel models identifying previously unknown treatment effects which can subsequently be attributed to potential causes. We show that the two-way fixed effects estimator, which identifies heterogeneous treatment effects through interactions, can be nested as a special case of impulse- or step-dummy saturated models – a subset of which identifies underlying treatment effects.

We demonstrate the feasibility of detecting previously unknown treatment assignment and timing by using two machine learning methods suitable for selection over more candidate variables than observations here using 'gets' and the adaptive LASSO, though many other approaches such as bayesian model selection would also be feasible.

Our applications demonstrate the broad utility of this approach. First, we show that we can detect the effects of 'treatment' (taking the form of a carbon tax) on per capita transport CO_2 emissions in Sweden without prior knowledge of its implementation. Second, we replicate the analysis of the effects of ETA terrorism on Spanish regional GDP per capita, recovering treatment effects consistent with the established literature. The ability to detect such interventions without prior knowledge is particularly valuable for climate econometrics, where the timing and effectiveness of policy mixes are often uncertain. Broadly, our proposed approach is modular and allows for the detection of structural breaks in fixed effects panels with flexible choices for the machine learning algorithms employed, available via the R-packages 'gets' and 'getspanel'.

Ultimately, this framework formalizes the exploratory phase of causal inference. While statistical algorithms can identify where and when the data generating process shifted, they cannot explain why without context. The 'discovery' approach we propose should therefore be viewed as a hypothesis-generating machine: it produces candidate treatment episodes characterized by precise timing and magnitude. These candidates must then be subjected to 'ex-post attribution' — a falsifiable process of matching statistical breaks to historical records or policy databases. By rigorously separating detection (statistical) from attribution (contextual), we provide empirical researchers with a disciplined method to answer the reverse causal question: what mattered?

References

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

Andersson, J. J. (2019). Carbon taxes and co2 emissions: Sweden as a case study. *American Economic Journal: Economic Policy*, *11*(4), 1–30.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, 821–856. (Publisher: JSTOR)

Apergis, N., & Lau, M. C. K. (2015). Structural breaks and electricity prices: Further evidence on the role of climate policy uncertainties in the Australian electricity market. *Energy Economics*, *52*, 176–182. (Publisher: Elsevier)

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric theory*, 315–352. (Publisher: JSTOR)

Baltagi, B. H., Feng, Q., & Kao, C. (2016). Estimation of heterogeneous panels with structural breaks. *Journal of Econometrics*, *191*(1), 176–195. (Publisher: Elsevier)

Bazinas, V., & Nielsen, B. (2015). *Causal transmission in reduced-form models*. Citeseer.

Callaway, B., & Sant'Anna, P. H. (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*. (Publisher: Elsevier)

Campos, J., Hendry, D. F., & Krolzig, H.-M. (2003). Consistent model selection by an automatic Gets approach. *Oxford Bulletin of Economics and Statistics*, *65*, 803–819. (Publisher: Wiley Online Library)

Castle, J., Doornik, J., Hendry, D., & Pretis, F. (2015). Detecting location shifts during model selection by step-indicator saturation. *Econometrics*, *3*(2), 240–264. doi: 10.3390/econometrics3020240

Castle, J., Doornik, J., Hendry, D. F., & Pretis, F. (2025). Detecting breaks in trends by trend-indicator saturation. *Available at SSRN 5106350*.

Chan, F., Mancini-griffoli, T., Pauwels, L. L., & others. (2008). Testing structural stability in heterogeneous panel data. Christchurch, New Zealand. (Publisher: Citeseer)

Engle, R. F., Hendry, D. F., & Richard, J.-F. (1983). Exogeneity. *Econometrica*, *51*(2), 277–304. Retrieved 2023-02-19, from http://www.jstor.org/stable/1911990

Estrada, F., Perron, P., & Martínez-López, B. (2013). Statistically derived contributions of diverse human influences to twentieth-century temperature changes. *Nature Geoscience*, *6*(12), 1050–1055. (Publisher: Nature Publishing Group)

Gelman, A. (2011). *Causality and statistical learning*. University of Chicago Press Chicago, IL.

Gelman, A., & Imbens, G. (2013, November). *Why ask Why? Forward Causal Inference and Reverse Causal Questions* (Working Paper No. 19614). National Bureau of Economic Research. Retrieved 2021-10-25, from https://www.nber.org/papers/w19614 (Series: Working Paper Series) doi: 10.3386/w19614

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*. (Publisher: Elsevier)

Hendry, D. F. (2020). *First in, first out: Econometric modelling of UK annual CO2 emissions, 1860–2017* (Tech. Rep.). Oxford: Economics Group, Nuffield College, University of Oxford. Retrieved from https://ideas.repec.org/p/nuf/econwp/2002.html

Hendry, D. F., & Johansen, S. (2015). Model discovery and trygve haavelmo's legacy. *Econometric Theory*, *31*(1), 93–114. doi: 10.1017/S0266466614000218

Hendry, D. F., Johansen, S., & Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, *23*(2), 317–335. (Publisher: Springer)

Hendry, D. F., & Santos, C. (2005). Regression models with data-based indicator variables. *Oxford Bulletin of Economics and statistics*, *67*(5), 571–595. (Publisher: Wiley Online Library)

Huang, J., Ma, S., & Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 1603–1618. (Publisher: JSTOR)

Jiao, X., & Pretis, F. (2020). Testing the presence of outliers in regression models. *Working Paper*.

Jiao, X., Pretis, F., & Schwarz, M. (2021, August). *Testing for Coefficient Distortion due to Outliers with an Application to the Economic Impacts of Climate Change* (SSRN Scholarly Paper No. ID 3915040). Rochester, NY: Social Science Research Network. Retrieved 2021-11-03, from https://papers.ssrn.com/abstract=3915040 doi: 10.2139/ssrn.3915040

Johansen, S., & Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. *Castle, and Shephard (2009)*, *1*, 1–36.

Johansen, S., & Nielsen, B. (2016a). Analysis of the Forward Search using some new results for martingales and empirical processes. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, *22*(2), 1131–1183. (Publisher: Bernoulli Society for Mathematical

Statistics and Probability)

Johansen, S., & Nielsen, B. (2016b). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, *43*(2), 321–348. doi: 10.1111/sjos.12174

Koch, N., Naumann, L., Pretis, F., Ritter, N., & Schwarz, M. (2022). What reduces road CO_2 emissions? Policy attribution using break detection. *Working Paper*.

Kock, A. B. (2013). Oracle efficient variable selection in random and fixed effects panel data models. *Econometric Theory*, 115–152. (Publisher: JSTOR)

Kock, A. B. (2016). Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. *Econometric Theory*, *32*(1), 243. (Publisher: Cambridge University Press)

Konrad, L., Vashold, L., & Crespo Cuaresma, J. (2026). The bayesian indicator saturated model (bisam). *Working Paper*.

Leeb, H., & Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, *24*(2), 338–376.

Li, D., Qian, J., & Su, L. (2016). Panel data models with interactive fixed effects and multiple structural breaks. *Journal of the American Statistical Association*, *111*(516), 1804–1819. (Publisher: Taylor & Francis)

Martinez, A. B. (2020). Forecast accuracy matters for hurricane damage. *Econometrics*, *8*(2), 18.

Mill, J. S. (1843). *A system of logic* (Vol. 1). London: Parker.

Mukanjari, S., & Sterner, T. (2018). Do markets trump politics? evidence from fossil market reactions to the paris agreement and the us election.

Nielsen, B., & Qian, M. (2025). Asymptotic properties of the gauge and power of step-indicator saturation. *Econometric Theory*.

Perron, P. (1989). The great crash, the oil price shock, and the unit root hypothesis. *Econometrica: journal of the Econometric Society*, 1361–1401. (Publisher: JSTOR)

Perron, P. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics*, *1*(2), 278–352.

Piehl, A. M., Cooper, S. J., Braga, A. A., & Kennedy, D. M. (2003). Testing for structural breaks in the evaluation of programs. *Review of Economics and Statistics*, *85*(3), 550–558. (Publisher: MIT Press)

Pretis, F. (2021). Exogeneity in climate econometrics. *Energy Economics*, *96*, 105122. Retrieved from https://www.sciencedirect.com/science/article/pii/S014098832100027X doi: https://doi.org/10.1016/j.eneco.2021.105122

Pretis, F. (2022). Does a carbon tax reduce co2 emissions? evidence from british columbia. *Environmental and Resource Economics*, *83*(1), 115–144.

Pretis, F., Reade, J., & Sucarrat, G. (2018). Automated General-to-Specific (GETS) regression modeling and indicator saturation methods for the detection of outliers and structural breaks. *Journal of Statistical Software*, *86*(3). (Publisher: Foundation for Open Access Statistics)

Qian, J., & Su, L. (2016). Shrinkage estimation of common breaks in panel data models via adaptive group fused lasso. *Journal of Econometrics*, *191*(1), 86–109. (Publisher: Elsevier)

Rodríguez-Pose, A., & Hardy, D. (2021). Reversal of economic fortunes: Institutions and the changing ascendancy of Barcelona and Madrid as economic hubs. *Growth and Change*, *52*(1), 48–70. (Publisher: Wiley Online Library)

Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences?

a synthesis of the recent econometrics literature. *Journal of Econometrics*, *235*(2), 2218–2244.

Schwarz, M., & Pretis, F. (2026). *getspanel.* GitHub. Retrieved from https://github.com/moritzpschwarz/getspanel

Stechemesser, A., Koch, N., Mark, E., Dilger, E., Klösel, P., Menicacci, L., ... others (2024). Climate policies that achieved major emission reductions: Global evidence from two decades. *Science*, *385*(6711), 884–892.

Tebecis, T. (2023). *Have climate policies been effective in austria?: A reverse causal analysis*. Vienna University of Economics and Business.

Tebecis, T., & Crespo Cuaresma, J. (2025). A dataset of structural breaks in greenhouse gas emissions for climate policy evaluation. *Scientific Data*, *12*(1), 42.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. (Publisher: Wiley Online Library)

Tobío, C. (1989). Economic and social restructuring in the Metropolitan Area of Madrid (1970–85). *International journal of urban and regional research*, *13*(2), 324–338. (Publisher: Wiley Online Library)

Wang, H., Li, B., & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *71*(3), 671–683.

Wooldridge, J. M. (2025). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Empirical Economics*, 1–43.

Zhao, S., Witten, D., & Shojaie, A. (2017). In defense of the indefensible: A very naive approach to high-dimensional inference. *arXiv preprint arXiv:1705.05543*.

Zhao, S., Witten, D., & Shojaie, A. (2021). In defense of the indefensible: A very naive approach to high-dimensional inference. *Statistical Science*, *36*(4), 562–577. (Publisher: Institute of Mathematical Statistics)

Zhou, S. (2010). Thresholded lasso for high dimensional variable selection and statistical estimation. *arXiv preprint arXiv:1002.1583*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

6 Supplementary Material

6.1 Known Treatment Assignment with Unknown Timing

Now suppose we know which units are treated, but the timing of treatment is unknown. This may be the case when we suspect some intervention or event took place in some regions/countries, but the actual date of the intervention is uncertain. Let H denote the set of treated individuals and $1_{\{i \in H,\, t \geq q\}}$ an indicator function equal to one when i is part of the treated group and t falls in the post-treatment period. When treatment timing is unknown, we can interpret the identification of treatment effects as a break detection problem where we detect a structural break in the treated unit's specific intercept conditional on there being a non-zero treatment effect:

$$y_{i,t} = c_i + \tau \times 1_{\{i \in H,\, t \geq q\}} + g_t + u_{i,t} \tag{53}$$

When treatment is known, the above model (53) corresponds to a partial structural change model (see e.g. Perron 2006) with $c_{i,t}$ being allowed to break for treated individuals in the sample, and we estimate the break date q as well as treatment effect τ. If there is only a single treated unit and we detect a structural break in its intercept at the time of treatment, then the resulting model with a structural break is *identical* to the treatment effect model (11). There is thus a close link between break detection and the estimation of treatment effects in TWFE estimators.

However, the above model (53) may be overly restrictive as it assumes a single treatment with known-assignment and unknown-timing. In practice there may exist a myriad of possibly unknown interventions and we may face uncertainty around both treatment assignment as well as timing. In other words, we may not know which (if any) units are treated, and at what time such treatment may have occurred. In addition, treatment effects may also be heterogeneous over treated units as well as over time.

6.2 Identifying Heterogeneous Treatment Effects in Staggered Treatment

Here we briefly summarise the results from Wooldridge (2025), deriving equation (30) to identify treatment effects when treatment is staggered. We assume there is no anticipation of treatment for each $r = q, q+1, ..., Q$:

$$E[y_t(r) - y_t(\infty)|\mathbf{d}] = 0, \text{ for } t < r. \tag{54}$$

We also require a common trend assumption that the trend in absence of treatment is common regardless of state of treatment:

$$E[y_t(\infty) - y_1(\infty)|d_q, ...d_Q] = E[y_t(\infty) - y_1(\infty)] = \theta_t, \text{ for } t = 2, ..., T \tag{55}$$

and we assume at least one untreated group. The observed outcome in any period is given by:

$$y_t = y_t(\infty) + d_q te_t(q) + ... + d_Q te_t(Q) \tag{56}$$

where no anticipation implies that for the pre-treatment period ($t < q$):

$$E[y_t|\mathbf{d}] = E[y_t(\infty)|\mathbf{d}] \tag{57}$$

and for $t \geq q$:

$$E[y_t|\mathbf{d}] = E[y_t(\infty)|\mathbf{d}] + d_q\tau_{q,t} + ... + d_Q\tau_{Q,t} \tag{58}$$

We then write the never treated outcome $y_t(\infty)$ as an initial outcome and change relative to the initial period:

$$y_t(\infty) = y_1(\infty) + g_t(\infty) \tag{59}$$

By the common trend assumption $E[g_t(\infty)|\mathbf{d}] = \theta_t$:

$$E[y_t(\infty)|\mathbf{d}] = E[y_1(\infty)|\mathbf{d}] + E[g_t(\infty)|\mathbf{d}] = \eta + \lambda_q d_q + ... + \lambda_Q d_Q + \theta_t \tag{60}$$

which subsequently allows us to write the expected outcome as equation (30).

If we are interested in treatment effects of units treated at one point relative to those treated at a different point in time, as in Wooldridge (2025), we can define a sub-group ATT for those treated at r compared to for example one period later at $r + 1$ as

$$\tau_{(r:r+1)} = E[y_t(r) - y_t(r + 1)|d_r = 1] \tag{61}$$

This can be expressed as the difference in treatment effects relative to the untreated group:

$$y_t(r) - y_t(r + 1) = [y_t(r) - y_t(\infty)] - [y_t(r + 1) - y_t(\infty)] \tag{62}$$

Thus

$$\tau_{(r:r+1)} = \tau_{r,t} - E[y_t(r + 1) - y_t(\infty)|d_r = 1] \tag{63}$$

which under no anticipation and parallel trends simplifies to:

$$\tau_{(r:r+1)} = \tau_{r,t} - \tau_{r+1,t} \tag{64}$$

and which is matched by the difference in coefficients $\hat{\tau}$ obtained post-break detection on treatment dummies (step-functions or impulses).

6.3 Simulation Study

Here we investigate the properties of detecting treatment in our reverse causal setting using 'gets' and the adaptive LASSO. For the simulations we focus on detecting piece-wise constant treatment in the form of step-functions. Future work will expand simulations to also include fully-time-varying effects through impulse indicators.

We vary the treatment effect size σ as well as the number of treated units n. We compare the detection of unknown treatments against the 'known treatment' standard TWFE estimator for a single treated unit as well as multiple treated units. We then consider the case where we impose a known treatment while searching for additional treatment as described in section 3.2.

We simulate the DGP in (22) with errors drawn from the standard normal distribution and evaluate the performance of treatment detection as follows. For 'gets' we select over the full set of break functions using varying target levels of significance (γ_c). We use cross-validation to determine the penalty level for the adaptive LASSO. To measure the false positive rate of detection we compute the proportion of spuriously retained breaks (out of all possible spurious breaks). To measure whether we correctly identify treatment, we classify the proportion of correctly identified treated observations as those for which the detected breaks include the true treatment effect within a $(1 - \gamma_c)$ confidence interval.

Figure 7 shows the false positive rate together with the correctly classified proportion of treated observations for a single treated unit when varying the treatment magnitude (as a function of the standard deviation of the error term). Note that for a treatment effect size of 0 no treatment is present and hence no treatment should be identified – in this case therefore the rejection frequency yields a measure of the false-positive rate. Results show that treatment detection using 'gets' (red, solid) is close to the benchmark of a known treatment estimated using a conventional TWFE estimator (blue solid). The false positive rate is stable around the chosen level of significance of selection (red dashed). The adaptive LASSO (green solid) using cross-validation to choose the penalty factor also achieves a high level of accurate classification, however, is consistently lower than 'gets' for all significance levels considered. The adaptive LASSO using cross-validation also exhibits an erratic false positive rate (green dashed).

We increase the number of treated units from one to two and then five in our simulations (with identical treatment timing and homogeneous treatment effects) with results shown in Figures 8 and 9. As expected, as we increase the number of treated units, the correct classification (detection of treatment) falls relative to the known treatment case (using a single dummy variable) as our treatment detection approach has to identify a separate treatment dummy per treated unit. Nevertheless, the correct rejection frequency remains high given that no prior information about treatment assignment or timing was used.

Finally, we consider the costs of searching for additional treatment when there is a single known treatment that has been imposed from the outset (i.e. forced in the model and not selected over; see section 3.2).

Figure 10 shows the root-mean-squared error (RMSE) of the estimated treatment effect on the known treatment dummy when selecting over additional break variables relative to the simple TWFE estimator (without selection), together with the false-positive rate of detected treatment (gauge). The DGP only contains the single known treatment, with no other unknown treatment occurring. Thus this provides an assessment of the costs of searching for additional breaks when a known treatment is embedded. The results in Figure 10 show that searching for additional treatment when a known treatment is imposed, increases the RMSE on the estimated treatment effect for known treatment, however, for increasingly conservative selection significance levels this cost shrinks close to zero. This can be seen as an insurance cost – controlling for possible treatment (or breaks) that have been omitted from a standard model increases the RMSE of the known treatment indicator while providing robustness against omitted breaks. In other words, searching for additional breaks (i.e. treatment) lowers the precision on a known forced treatment somewhat, but the degree to which the RMSE increases can be easily controlled by choosing conservative levels of selection when using 'gets'. For 'gets', as Figure 10 shows, the false positive rate

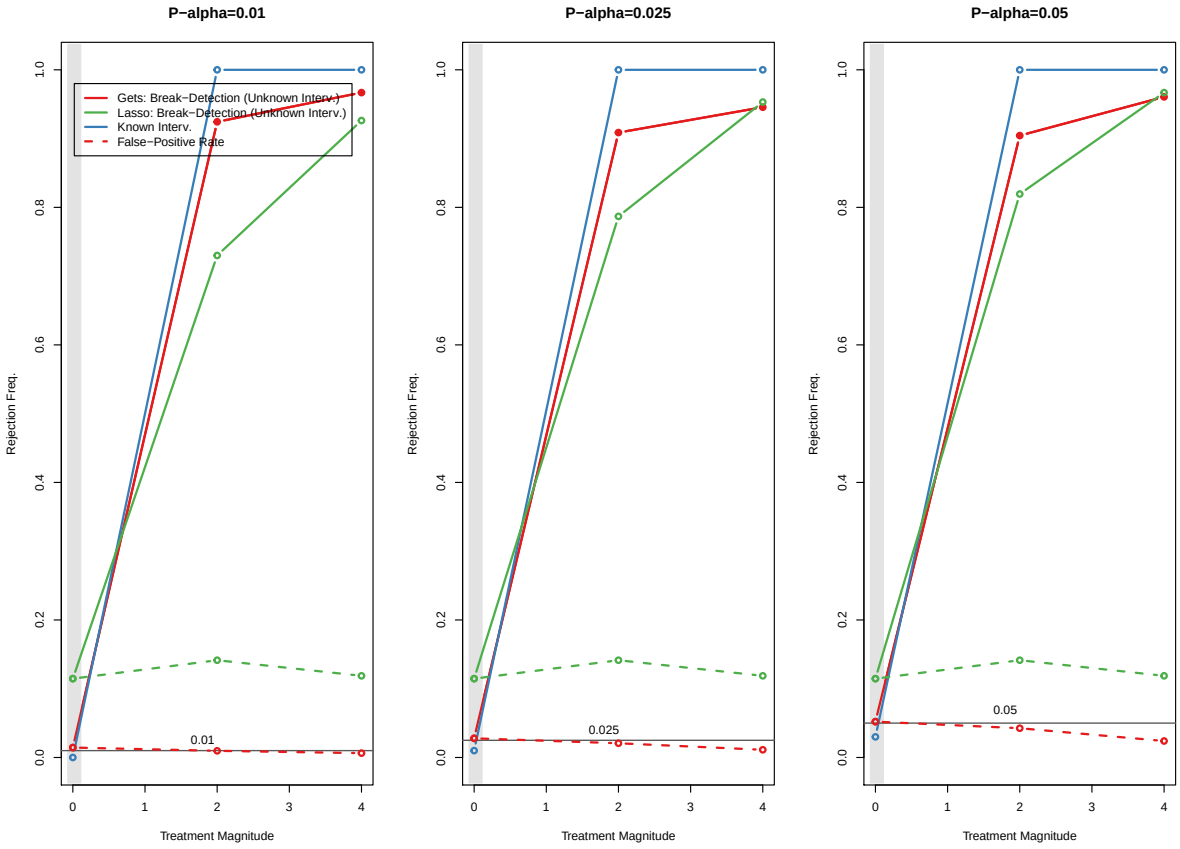

Figure 7: Simulation: Detecting treatment with **one** unknown treated units using 'gets' (SIS) and the adaptive LASSO compared to a 'known' treatment with $N = 10$

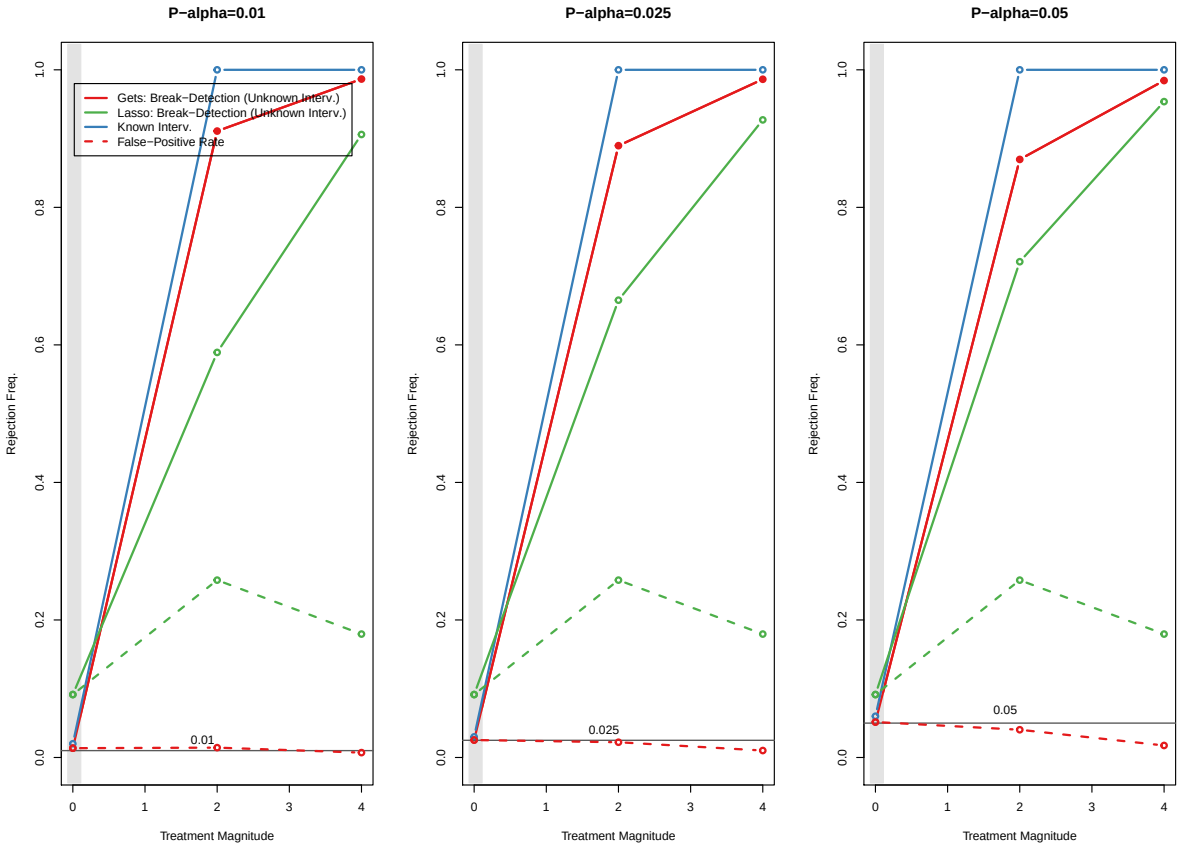

Figure 8: Simulation: Detecting treatment with **two** unknown treated units using 'gets' (SIS) and the adaptive LASSO compared to a 'known' treatment with $N = 10$

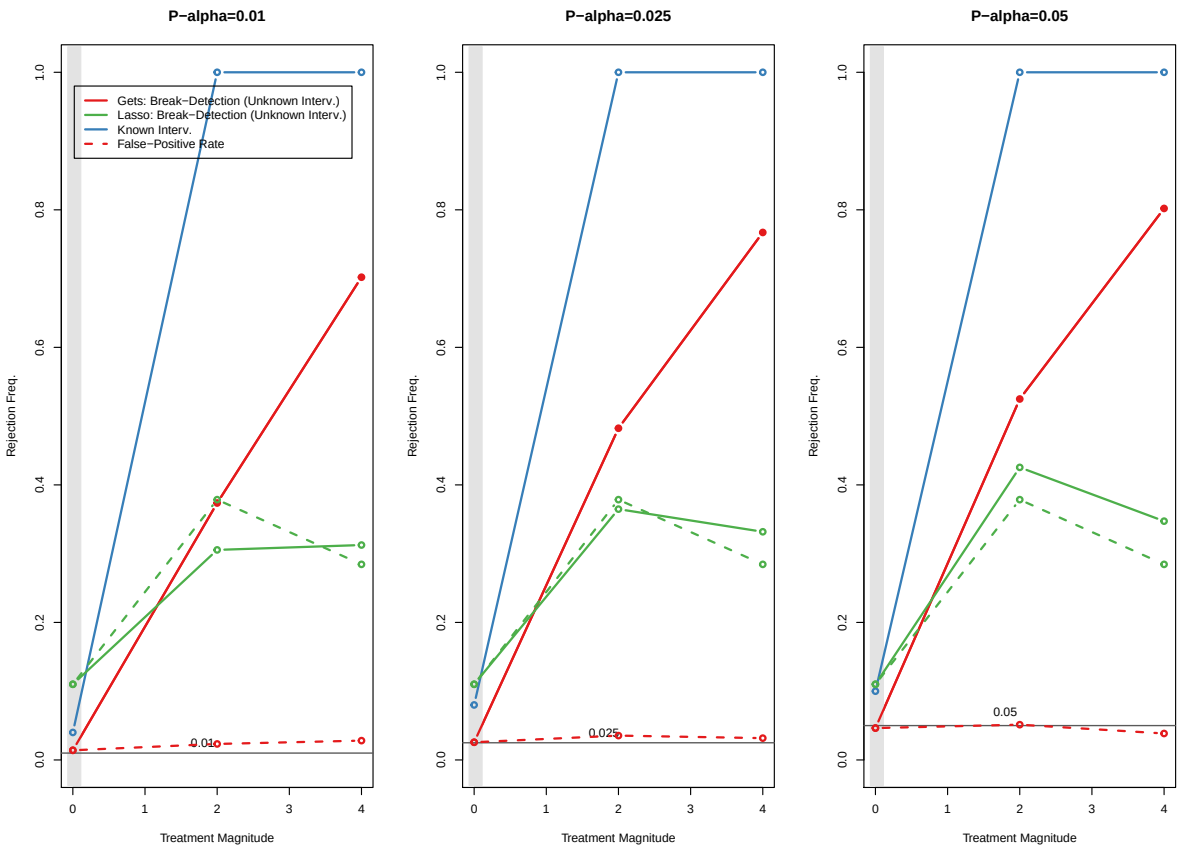

Figure 9: Simulation: Detecting treatment with **five** unknown treated units using 'gets' (SIS) and the adaptive LASSO compared to a 'known' treatment with $N = 10$

(gauge) again is stable around the specified nominal level of significance. Such control is more difficult to achieve when using the LASSO due to not targeting the false-positive rate.

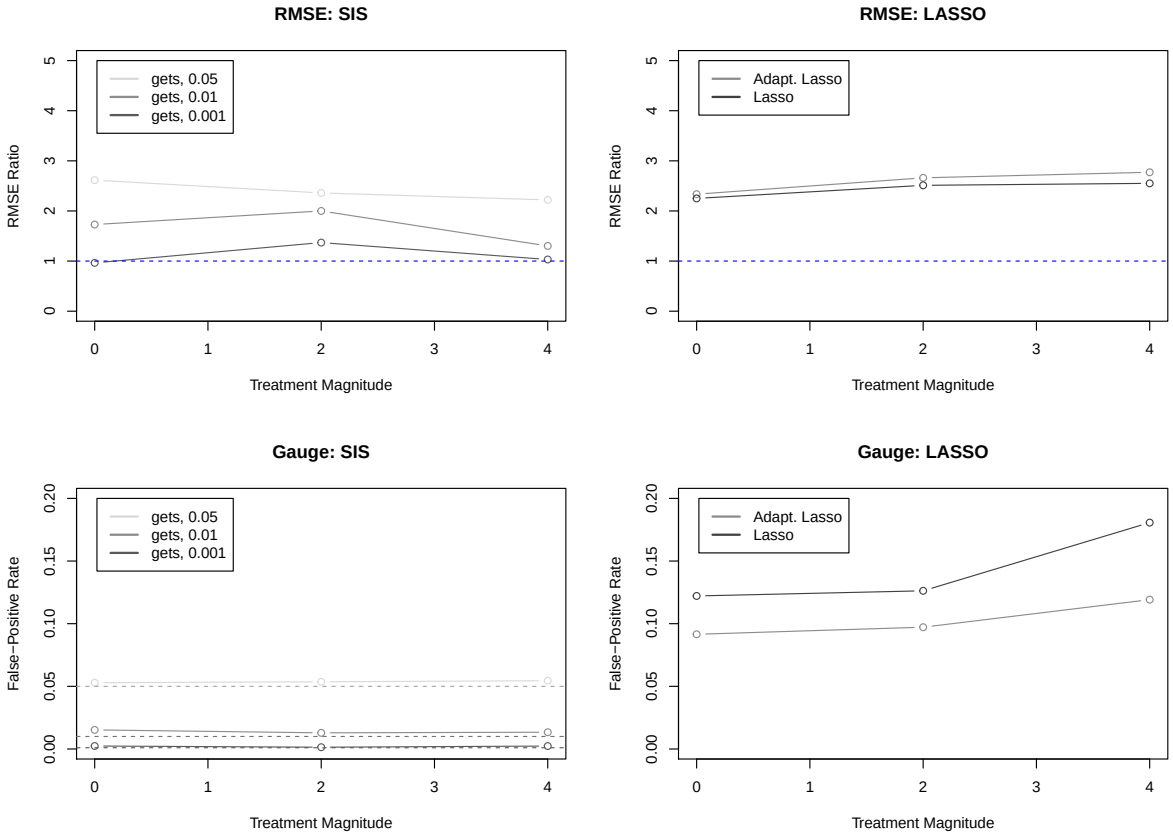

Figure 10: Top: RMSE of estimated 'known' single treatment effect when searching for additional treatment relative to known TWFE estimator. Bottom: False positive rate of detected breaks.

6.4 Simple Time Series Approach

We estimate a simple time series model of Basque GDP per capita in (65) and demonstrate that in absence of control groups, we are unable to detect the impact of ETA Basque terrorism *a-priori*. We model the log of GDP per capita as a function of log investment (one of the original control variables in Abadie and Gardeazabal), while searching for structural breaks in the intercept using step-indicators with 'gets' at a conservative target significance level of $\gamma_c = 0.001$:

SIS – Time Series for Basque Country only:

$$log(GDPpc)_t = \beta_0 + \beta_1 log(Inv)_t + \sum_{s=1966}^{1995} \tau_s 1_{\{t \geq s\}} + \epsilon_t \tag{65}$$

Estimation results of this time series model are shown in Table 4 and Figure 11. While multiple breaks are found, the negative impact of ETA terrorism on GDP per capita in the Basque region is not detectable due to the lack of control regions. There are no detected breaks with negative coefficients during the period that ETA was active.

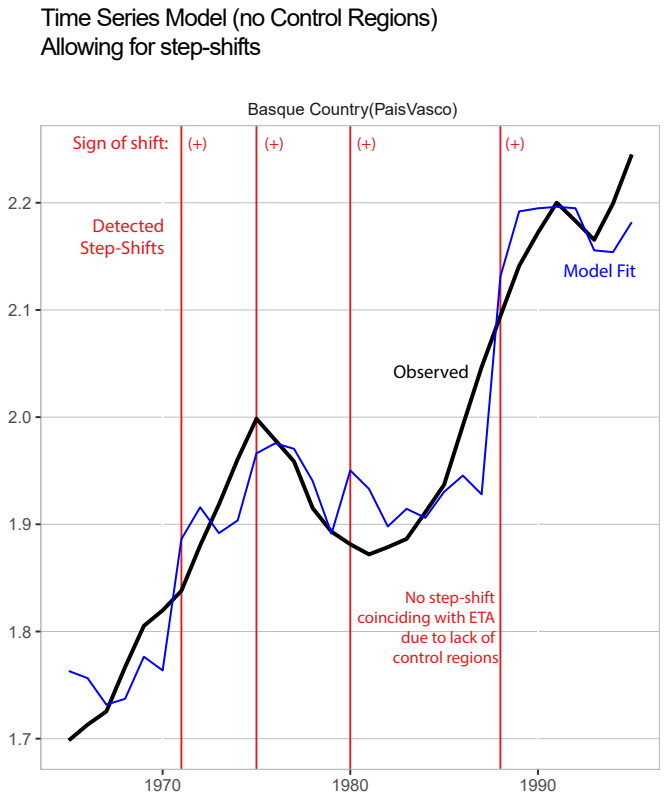

Figure 11: Simple Time Series Model of Basque GDP per Capita – Breaks detected using 'gets' and $\gamma_c = 0.001$.

Table 4: Detecting Breaks in a Simple Time Series Model (Basque Country)

Dependent Variable:	log(GDPpc)
Model:	Time Series
Variables	
Constant	0.5367
	(0.5008)
log(Invest)	0.3788**
	(0.1556)
Break (i=Basq., $t \geq 1971$)	0.1663***
	(0.0321)
Break (i=Basq., $t \geq 1975$)	0.1308***
	(0.0463)
Break (i=Basq., $t \geq 1980$)	0.0536
	(0.0417)
Break (i=Basq., $t \geq 1988$)	0.1737***
	(0.0392)
Fit statistics	
Observations	31
R^2	0.92

Standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 5: Detecting Piece-Wise Constant Treatment: 15-Region Panel Model

Dependent Variable:	log(GDPpc)	
	Unknown Treatment	**Known Treatment**
Model:	gets ($\gamma_c = 0.0001$)	'Known' TWFE
Variables		
log(Invest)	0.1171***	0.1377***
	(0.0121)	(0.0175)
τ: Break (i=Basq, $t \geq$1978)	**-0.1560*****	
	(0.0120)	
τ: Known ETA (i=Basq, $t \geq$1979)		**-0.1553*****
		(0.0182)
τ: Break: (i=Castilla-La Mancha, $t \geq 1972$)	0.1169***	
	(0.0143)	
τ: Break: (i=Extremadura, $t \geq 1987$)	0.1350***	
	(0.0127)	
τ: Break: (i=Galicia, $t \geq 1976$)	0.0980***	
	(0.0121)	
τ: Break: (i=Madrid, $t \geq 1970$)	-0.1256***	
	(0.0176)	
τ: Break: (i=Madrid, $t \geq 1990$)	-0.0903***	
	(0.0150)	
τ: Break: (i=Princip. De Asturias, $t \geq 1986$)	-0.1220***	
	(0.0123)	
τ: Break: (i=La Rioja, $t \geq 1981$)	0.0796***	
	(0.0117)	
τ: Impulse: (i=Madrid, $t = 1965$)	0.0914**	
	(0.0356)	
Fixed-effects		
Region	Yes	Yes
Year	Yes	Yes
Observations	N=15, T=31, NT=465	N=15, T=31, NT=465
Within R^2	0.71	0.29

Standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*