Causal inference and policy evaluation without a control group

Causal inference and policy evaluation without a control group*

Augusto Cerqua†, Marco Letta†, Fiammetta Menchetti‡

First version: December 2022
This version: October 2024

Abstract

Without a control group, the most widespread methodologies for estimating causal effects cannot be applied. To fill this gap, we propose the Machine Learning Control Method, a new approach for causal panel analysis that estimates causal parameters without relying on untreated units. We formalize identification within the potential outcomes framework and then provide estimation based on machine learning algorithms. To illustrate the practical relevance of our method, we present simulation evidence, a replication study, and an empirical application on the impact of the COVID-19 crisis on educational inequality. We implement the proposed approach in the companion R package `MachineControl`.

Keywords: potential outcomes framework, counterfactual forecasting, machine learning, short panels, panel cross-validation, educational inequality.

JEL-Codes: C18, C21, C53, I24.

---

† Department of Social Sciences and Economics, Sapienza University of Rome, Rome, Italy (IT). Email at: `augusto.cerqua@uniroma1.it`; `marco.letta@uniroma1.it`

‡ Department of Statistics, Computer Science and Applications, University of Florence, Florence, Italy (IT). Email at: `fiammetta.menchetti@unifi.it`

*We are grateful to Guido Imbens and Fabrizia Mealli for thoughtful conversations and discussions. We also thank Andrea Albanese, Guglielmo Barone, Michele Battisti, Iavor Bojinov, Fabrizio Cipollini, Alessio D'Ignazio, Christina Gatmann, Anna Gottard, Giulio Grossi, Martin Huber, Michael Knaus, Joanna Kopinska, Giuseppe Maggio, Alessandra Mattei, Raffaele Mattera, Giovanni Mellace, Andrea Mercatanti, Samuel Nocito, Gabriele Pinto, Jason Poulos, Donato Romano, Federico Rucci, Jacques-François Thisse, Luca Tiberti, Giuseppe Ragusa, Giuliano Resce, and Bas van der Klaauw for valuable suggestions on earlier versions of this work. The paper has benefited from helpful comments and suggestions by audiences at many conferences and seminars.

"In history there are no control groups. There is no one to tell us what might have been."
Cormac McCarthy, All the Pretty Horses (1992)

1 Introduction

Today, the econometric toolbox for causal panel analysis features many alternative approaches for estimating causal effects, such as difference-in-differences (Card and Krueger, 1994), the synthetic control method (Abadie et al., 2010), two-way (Angrist and Pischke, 2009) and interactive (Bai, 2009) fixed effects models, and matrix completion methods (Athey et al., 2021). However, all of these popular methodologies depend on a critical requirement: the availability of untreated units forming a credible control group, without which they cannot be applied. This poses a significant empirical challenge in observational studies, as there are at least two relevant cases in which a suitable control group does not exist: (i) when the treatment simultaneously affects all units, such as in the context of a large-scale shock or a nationwide program with universal participation (Duflo, 2017); (ii) when only a subgroup of units receives the treatment, but the set of untreated units cannot form a valid control group due to violations of the no-interference assumption (Cox, 1958).[1] Under these challenging but not uncommon circumstances, the use of standard causal panel data methods is precluded, leaving a gap in the econometric toolkit of empirical researchers.

We fill this gap by introducing the Machine Learning Control Method (MLCM), a new estimator based on flexible counterfactual forecasting via machine learning (ML). The MLCM is a versatile technique that can leverage any available supervised ML algorithm and enables the identification and estimation of several policy-relevant causal parameters—including individual, average, and conditional average treatment effects (CATEs)—in practically relevant environments without a control group. It is flexible in terms of data structure, as it can be applied across a wide variety of panel settings, including very short panels and contexts with staggered adoption. The intuition behind the MLCM is straightforward: when you cannot rely on untreated units to build the counterfactual, you can forecast it. Specifically, use supervised ML techniques trained and tuned on the pre-treatment data to forecast the evolution of post-treatment outcomes in the absence of the treatment. Then, estimate treatment effects as the difference between observed and forecasted post-treatment outcomes.

Our approach builds upon and bridges three different methodological currents: causal panel data methods, time series forecasting, and causal ML. From the literature on causal panel analysis, we adopt the conceptual and identification framework based on the potential outcomes model (Rubin, 1974). Since earlier contributions (Ashenfelter and Card, 1985;

[1] Spillovers are common in many empirical settings where the observations are correlated in time, space, or both, and ignoring them can lead to very misleading inferences (Sobel, 2006). In principle, spillovers can be modeled and accounted for at the cost of additional and often restrictive assumptions. However, in practice, most methods just postulate the absence of interference.

Card, 1990) to its more recent developments (Arkhangelsky et al., 2021; Roth et al., 2023), this entire literature revolves around addressing the "fundamental problem of causal inference" (Holland, 1986) in an observational panel data setting where, for each unit, it is impossible to observe both potential outcomes at once. Therefore, causal inference can be viewed as a missing data problem, with the key goal being the imputation of missing potential outcomes for the treated units (Imbens and Rubin, 2015). Most widely used causal panel data methods achieve this imputation by leveraging untreated units to construct a control group. Our key contribution to this literature is to establish identification conditions in the absence of a control group and to introduce a novel estimator of causal effects that imputes missing potential outcomes without relying on untreated units.

The time series literature features some forecasting approaches that researchers have employed for counterfactual building in scenarios without a control group.[2] Some of them are straightforward: the mean method estimates the counterfactual "no-treatment scenario" of the dependent variable Y_i as the average of its pre-treatment values, while the naive method considers the last pre-treatment value of Y_i as the counterfactual (Hyndman and Athanasopoulos, 2021). A more sophisticated forecasting approach is the interrupted time series (ITS) analysis, first introduced by Box and Tiao (1975). ITS analysis generally fits regression models or autoregressive integrated moving average (ARIMA) models to the entire time series of observed data; this is done after postulating a structure on the intervention effect (e.g., constant level shift). By construction, ITS delivers a single average effect for the whole post-intervention period. More recent studies (Brodersen et al., 2015; Chernozhukov et al., 2021; Menchetti et al., 2023; Rambachan and Shephard, 2021) consider estimation and inference with other time series techniques (such as Bayesian Structural Time Series, ARIMA models, or impulse response functions) in time series settings where no direct control group may be available. From a panel data perspective, these methods have significant limitations. First, they are designed for cases with a single treated series. Second, they often impose restrictive assumptions and functional forms. Third, they can be implemented only when many pre-treatment periods are available. However, in the traditional panel data literature, the number of units is much larger than the number of time periods (Arkhangelsky and Imbens, 2024), and the time dimension is typically substantially smaller than in time series settings, rendering these methods mostly inapplicable. We draw from this literature the idea of forecasting counterfactuals by exploiting the temporal information. We go beyond it by proposing a more flexible, ML-powered method explicitly designed for counterfactual forecasting with panel data.

In recent years, a new methodological literature at the intersection between causal inference with panel data and ML has evolved. The matrix completion methods developed

[2] A rich tradition also exists for forecasting with panel data (e.g., Arellano and Bonhomme (2011); Baltagi (2008); Liu et al. (2020)). However, our interest lies in *counterfactual* forecasting. In this regard, the Rubin causal framework is absent from this tradition, which focuses on different estimands and non-causal settings compared to the recent causal panel data literature (Arkhangelsky and Imbens, 2024).

by Athey et al. (2021) use nuclear norm regularization and observed control outcomes to impute missing elements in the counterfactual matrix of untreated unit/period combinations. Semenova et al. (2023) propose an approach for estimating CATEs characterized by high-dimensional parameters in both homogeneous cross-sectional and unit-heterogeneous dynamic panel data settings, leveraging ML techniques for improved inference. Artificial control methods (Carvalho et al., 2018; Masini and Medeiros, 2021) and synthetic learners (Viviano and Bradic, 2023) harness supervised ML algorithms to predict counterfactuals and assess treatment effects in a panel setting with a single treated unit, many untreated units, and a long pre-intervention window. All these methods need a set of units assumed to be completely unaffected by the treatment to build a counterfactual scenario. Therefore, none of these causal ML techniques can be applied in the challenging econometric setting we focus on. We share with this literature the idea of harnessing the power of ML in the service of causal inference rather than for pure prediction, exploiting the fact that counterfactual building is ultimately a predictive task (Varian, 2016). We extend it by developing a new causal ML estimator that estimates individual treatment effects without depending on the no-interference assumption or leveraging unaffected units.

We begin by thoroughly discussing and formalizing identification without a control group in the potential outcomes framework. We then propose an adaptive estimator based on a horse-race competition between several different ML algorithms, a novel cross-validation (CV) procedure for model assessment and selection tailored for panel data, and block-bootstrapping for inference. The MLCM comes with a set of diagnostic and placebo tests and is characterized by a high level of generality: it delivers individual treatment effects that can either be the object of interest or aggregated into several policy-relevant causal estimands. Since our approach allows for unrestricted treatment effect heterogeneity, we also propose an automated search for heterogeneity based on an easy-to-interpret regression tree.

To showcase the potential of the MLCM, we present an extensive simulation study, a replication of the minimum wage application in Callaway and Sant'Anna (2021) without using control units, and an empirical application in which we investigate the effects of the COVID-19 pandemic on educational inequality in Italian Local Labor Markets (LLMs). We find that the pandemic led to a generalized drop in students' performance, which is particularly pronounced in LLMs that before COVID-19 were characterized by higher levels of unemployment and inequality and lower educational attainment.

At the time of writing, we are aware of only one other method for assessing causality in panel settings without using control units.[3] In independent work subsequent to ours, Botosaru et al. (2023) propose estimating average treatment effects in the absence of a control group by regressing pre-treatment outcomes on known basis functions of time. Compared to

[3]There are also some empirical studies in energy economics that, without proposing a new formal method, have applied counterfactual prediction using ML algorithms on very high-frequency (hourly) electricity data (Abrell et al., 2022; Jarvis et al., 2022; Prest et al., 2023).

the MLCM, this method relies on different assumptions, such as absence of interference and the validity of a central limit theorem, and imposes functional form restrictions. Overall, we view the two methods as complementary, depending on the assumptions one is willing to make and on the characteristics of the available data.[4]

The possibility of moving beyond the reliance on untreated units paves the way to evaluating many potential treatments—such as universal policies, large-scale shocks, and programs engendering interactions between units—that, due to the lack of a valid comparison group, have so far been under-explored in empirical studies. To answer these causal questions, interested researchers can implement the MLCM on real-world datasets using the companion R package `MachineControl`.[5]

The rest of this paper is organized as follows. Section 2 outlines the causal framework by defining the identification assumptions, the causal estimands, and the estimators. Section 3 describes the implementation process of the MLCM and presents the simulation study. Section 4 illustrates the empirical application, while Section 5 concludes.

2 The causal framework

In this section, we present the causal framework for an observational short-panel setup where the intervention is a simultaneous policy change or shock that affects directly or indirectly the entire statistical population under scrutiny.[6] We start by defining the assumptions and the corresponding causal estimands within the potential outcomes framework. We then provide the proof that such estimands can be identified under those assumptions. Finally, we introduce novel estimators for scenarios without a suitable control group. To make the notation (and the method) as general as possible, the definitions given in this section assume the presence of some contemporaneous covariates unaffected by the intervention. Including these covariates in the MLCM can potentially mitigate certain assumptions. However, as we will discuss later, caution must be exercised in their use and selection.

[4]For instance, the estimator by Botosaru et al. (2023) can be employed in data-scarce environments where only outcome data are available, or when the researcher is comfortable relying on parametric assumptions. In contrast, the MLCM is more flexible and better suited for data-rich or high-dimensional settings where information on many covariates is available. This context is particularly relevant when the primary goal is to uncover heterogeneity in treatment effects associated with differences in observed characteristics.

[5]The latest version of the package is available on GitHub at this link.

[6]To use language from Arkhangelsky and Imbens (2024), in this benchmark case, we consider a thin panel matrix, with a large number of cross-sectional units and a small number of time periods. However, the framework easily accommodates the use of long panels and fat and square matrices. In addition, the MLCM can be applied to both balanced and unbalanced panels, as well as in block-assignment and staggered adoption settings where one cannot credibly rely on the plausibility of the no-interference assumption (i.e., outcomes of never-treated or not-yet-treated units are contaminated by spillovers).

2.1 Assumptions

Denote with $Y_{i,t}$ the outcome of unit i at time t, and let $W_{i,t} \in \{0, 1\}$ be a random variable describing the treatment assignment of unit $i = 1, \ldots, N$ at time $t = 1, \ldots, t_0, \ldots, T$, where 1 indicates the treatment, 0 indicates control, and t_0 denotes the intervention date.[7] As we are focusing on a single and simultaneous intervention, we can then write $W_{i,t} = 0$ for all i and $t \leq t_0$, $W_{i,t} = 1$ for all i and $t > t_0$, therefore, $t_0 + 1$ is the first post-intervention period. In other words, all units are treated simultaneously and remain treated thereafter. Under the potential outcomes framework, for each unit, we can define a potential outcome under $W_{i,t} = 0$ and a different potential outcome under $W_{i,t} = 1$.

We now introduce the assumptions that are needed to define, identify, and estimate causal effects in an observational panel setting without controls, offering novel theoretical insights such as identification conditions for non-linear, multi-step-ahead causal effects. We emphasize that such assumptions are not particularly restrictive, as the MLCM is designed to be highly flexible. Our method, whose implementation is detailed in Section 3, starts by running several ML algorithms on pre-treatment data and then selects the one producing the most accurate forecasts. Given that it is not possible to know in advance which ML method will be chosen, the notation is kept as general as possible.[8]

The first assumption contributes to define potential outcomes and establishes the crucial link between them and the treatment assignment. We rely on a less stringent version of the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1974; Imbens and Rubin, 2015), retaining only its second part: the treatment is the same for all the units. This is an *a priori* assumption that the potential outcomes do not depend on different treatment intensities or the mechanism used to assign the treatment.

Assumption 1 (Weak SUTVA)**.** *There are no hidden forms of treatment leading to different potential outcomes.*

While we maintain this no-multiple-versions-of-treatment assumption, we drop the first part of SUTVA, namely, the no-interference assumption (Cox, 1958).[9] We consider this as one of the main advantages of our approach because, while it has long been known that violations and failures of the no-interference assumption can lead to misleading inferences in many social science settings (Sobel, 2006), this strong assumption is rarely questioned

[7] The word "treatment" is commonly used in the context of randomized controlled trials. As we are dealing with an observational study, we use the terms "treatment", "intervention", "policy", and "shock" interchangeably.

[8] We believe this is an advantage over most existing approaches in the causal ML literature: while Carvalho et al. (2018) and Masini and Medeiros (2021) focus on LASSO and Wager and Athey (2018) adopt tree-based methods, the MLCM is flexible and can easily adapt to different data-generating processes. A notable exception in this context is the synthetic learner proposed by Viviano and Bradic (2023). In contrast to their technique, our method is based on a horse-race competition between several alternative ML algorithms, whereas the synthetic learner relies on an ensemble procedure that combines various estimators.

[9] Following Imbens and Rubin (2015), we consider the case with general equilibrium effects as a scenario in which there are widespread violations of the no-interference assumption.

or tested in practice (Chiu et al., 2023). This is a key departure from most evaluation methods and it is important to delve into its implications. SUTVA-related interference among units can be of two types: (1) from treated to control units, and (2) among treated units themselves.[10] First, we completely circumvent pitfalls regarding Type-1 interference, since our counterfactual scenario is generated without relying on untreated units. Second, we do not postulate the lack of interference among treated units and allow for any possible Type-2 interference. We avoid relying on the no-interference-among-treated assumption because of the likely presence of interference across both the temporal and cross-sectional dimensions in social science applications (Xu, 2023). Consequently, our focus is on estimating the *total* effect of the treatment, which encompasses both the direct effect on a unit from the treatment it receives and the indirect effects arising from the spillover and general equilibrium effects. We claim that in real-world scenarios with likely interaction between units, focusing on the total effect of the treatment is a sensible choice as it captures the actual impact on each unit under the realized treatment assignment mechanism. Under Assumption 1, we can index the potential outcomes to the common treatment path: $\mathrm{Y}_{i,t}(1)$ indicates the potential outcome under the treatment, and $\mathrm{Y}_{i,t}(0)$ is the counterfactual outcome.

The next assumption contributes to the definition of causal effects.

Assumption 2 (Additivity)**.** *We assume that the intervention produces an additive effect on the potential outcomes, i.e.,*

$$\mathrm{Y}_{i,t}(1) = \mathrm{Y}_{i,t}(0) + \Delta_{i,t}. \tag{1}$$

This is not a restrictive assumption, since it holds after a suitable transformation of the data. For example, we can start from a multiplicative effect and then apply a logarithmic transformation to find an additive structure. Since inference is conducted using block-bootstrap, if it is necessary to express the effect in its original form, the inverse transformation can be applied to the bootstrap distribution, so as to recover the original multiplicative effect and its confidence interval. In conjunction with Assumption 4 (see below), this assumption implies that we effectively adopt a Partially Linear Model, akin to seminal causal machine learning methods (Chernozhukov et al., 2018; Wager and Athey, 2018). The Partially Linear Model strikes a balance between structure and flexibility: the causal-effect component of the model remains simple and interpretable, while the untreated potential outcome can exhibit almost arbitrary complexity (Chernozhukov et al., 2024).

Assumption 3 rules out the possibility of anticipatory effects of the intervention on both the outcome and the covariates (Abadie et al., 2010; Borusyak et al., 2023).

[10]Even if the treatment is the same for all units, residual spillover effects may arise due to interference among the treated units. Consider, for example, a study investigating excess mortality from COVID-19. While the pandemic affects all areas, those with fewer intensive care beds may need to send patients to neighboring hospitals, prompting an additional (indirect) rise in mortality rates there also. Spillovers due to individuals' characteristics within the same treatment group are discussed in Ogburn and VanderWeele (2014).

Assumption 3 (Absence of anticipation)**.** *Let* $\mathbf{X}_{i,t} = (\mathrm{X}_{i,t}^{(1)}, \ldots, \mathrm{X}_{i,t}^{(m)})$ *be an* m*-dimensional vector of covariates that are predictive of the outcome* i *at time* $t \leq t_0$*. We assume: i) absence of anticipatory effects of the intervention on the covariates and the potential outcomes, i.e.,* $\mathrm{Y}_{i,t}(1) = \mathrm{Y}_{i,t}(0)$ *and* $\mathbf{X}_{i,t}(1) = \mathbf{X}_{i,t}(0)$ *for all* $t \leq t_0$*; ii) future covariates do not affect current potential outcomes; iii)* (optional) *some of the covariates remain unaffected by the policy in the post-intervention period (post-treatment exogeneity of the covariates), i.e., for all* $t > t_0$*,* $\mathbf{X}_{i,t}(1) = \mathbf{X}_{i,t}(0)$*.*

This assumption implies that in the pre-intervention period, the observed outcome corresponds to the potential outcome absent the policy, i.e., $\mathrm{Y}_{i,t} \equiv \mathrm{Y}_{i,t}(0)$ and that $\mathbf{X}_{i,t} \equiv \mathbf{X}_{i,t}(0)$ throughout all the analysis period. Additionally, Assumption 3 precludes any anticipatory actions by agents, implying that they cannot alter or manipulate their outcomes prior to receiving the treatment. This assumption can be empirically tested by verifying whether pre-treatment effects are, on average, zero. Finally, post-treatment exogeneity of the covariates is not essential for identification, as in many applications, it is more realistic to use only lagged covariates. Thus, part (iii) of Assumption 3 is relevant only in cases where post-treatment values of some covariates are needed to control for post-treatment confounders and these covariates can be considered as exogenous (Brodersen et al., 2015).[11]

The next identification assumption pertains to the potential outcomes model, similar to Carvalho et al. (2018) and Masini and Medeiros (2021), but instead of using control units, we rely on a dynamic specification based on lagged outcomes.

Assumption 4 (Dynamic potential outcomes model)**.** *Denote with* $\mathbf{X}_{i,t}^{(q)} = (\mathbf{X}_{i,t}, \ldots, \mathbf{X}_{i,t-q})$ *the collection of covariates up to time* $t-q > 0$*, and let* $\mathrm{Y}_{i,t-1}^{(p)}(0) = (\mathrm{Y}_{i,t-1}(0), \ldots, \mathrm{Y}_{i,t-1-p}(0))$ *be the collection of past potential outcomes up to time* $t - p > 0$*, with* $p \in \{0, \ldots, t-2\}$ *and* $q \in \{0, \ldots, t-1\}$*. We assume that, for all* $t = 1, \ldots, T$*, the potential outcome absent the policy is as follows,*

$$\mathrm{Y}_{i,t}(0) = h\left(\mathrm{Y}_{i,t-1}^{(p)}(0), \mathbf{X}_{i,t}^{(q)}\right) + \epsilon_{i,t} \tag{2}$$

where $h(\cdot)$ *is some flexible function of the past lags of the outcome, the contemporaneous covariates, and the past lags of covariates;* $\epsilon_{i,t}$ *is a zero-mean, uncorrelated error term following a generic distribution* $f(\cdot)$

Notice that Equation (2) assumes poolability, i.e., homogeneous intercept and slope for all units, meaning that once we account for a highly predictive set of covariates and include temporal dynamics through lagged outcomes, the residual component is essentially random noise.[12] We posit this assumption can be met in practice by using a subset of very predictive covariates out of a larger initial set selected *ex ante* on the basis of domain knowledge. This

[11] Although motivating covariates' choice is often sufficient, it can be also verified by testing for the presence of treatment effects on each covariate: those that are significantly impacted by the intervention must be removed from the model.

[12] For panel data studies with small T and large N, it is usual to pool the observations (Baltagi, 2008).

approach ensures that major sources of heterogeneity among units are effectively addressed. In addition, we can provide empirical support for this assumption by examining the fit of the pooled model during the pre-intervention period. A good fit, indicated, for instance, by low Mean Squared Error (MSE), would suggest that the pooled model adequately captures the variations in the data. In cases where the researcher also wants to account for time-invariant, group-level unobserved heterogeneity (e.g., units' fixed effects), the covariate set could be augmented by adopting recent approaches based on sufficient representations for categorical variables (Johannemann et al., 2019).

We remark that $h(\cdot)$ can be either a linear or a non-linear function. In the linear case, no additional assumptions are needed for the identification of causal effects. In the non-linear case, no further assumptions are required if the effect is defined only a single step ahead, as in our empirical application. Conversely, identification of non-linear multi-step-ahead causal effects is notoriously difficult and has not yet been fully explored in the literature (for time series in non-causal settings, see Chen et al. (2004)). Therefore, our paper also contributes to the formal identification of causal effects in this challenging context. Specifically, when $h(\cdot)$ is non-linear and the focus is on multi-step-ahead causal effects, we propose modeling the potential outcomes in the post-intervention period as follows.

Assumption 5 (Post-intervention non-linear multi-step-ahead model)**.** *For a positive integer $k \geq 2$, denote by $\mathrm{Y}_{t_0+k|t_0}(0)$ the expected potential outcome absent the policy, given covariates and pre-intervention lags of the outcome, i.e., $\mathrm{Y}_{i,t_0+k|t_0}(0) = \mathrm{E}[\mathrm{Y}_{i,t_0+k}(0)|\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(q)}_{i,t_0+k}]$. The potential outcomes absent the policy at time t_0+k are functions of their past lags up to t_0, their conditional expectations $\mathrm{Y}_{i,t_0+1|t_0}, \ldots, \mathrm{Y}_{i,t_0+k-1|t_0}$, and the covariates $\mathbf{X}^{(Q)}_{i,t_0+k}$, with $Q = \max\{k-1, q\}$,*

$$\mathrm{Y}_{i,t_0+k}(0) = g_k\left(\mathrm{Y}_{i,t_0+k-1|t_0}(0), \ldots, \mathrm{Y}_{i,t_0+1|t_0}(0), \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(Q)}_{i,t_0+k}\right) + \epsilon_{i,t_0+k} \tag{3}$$

where ϵ_{i,t_0+k} is a zero-mean, uncorrelated error term, and $g_k(\cdot)$ is a flexible function which can be different at each horizon k.

Notice that Equation (3) does not impose any particularly restrictive assumption on the potential outcomes. Indeed, it merely formalizes the dependence structure between potential outcomes and their conditional expectations already implied by Equation (2), e.g., for $k = 2$ and $p = q = 0$, we have $\mathrm{Y}_{i,t_0+2}(0) = h(\mathrm{Y}_{i,t_0+1|t_0}(0) + \epsilon_{i,t_0+1}, \mathbf{X}_{i,t_0+2}) + \epsilon_{i,t_0+2} = g_2(\mathrm{Y}_{t_0+1|t_0}(0), \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}) + \epsilon_{i,t_0+2}$.

The panel CV procedure—thoroughly described in Section 3—which we propose to estimate the pre-intervention potential outcome model in Equation (2), will reveal whether the winner of the horse-race is a linear or non-linear model. If a linear model is selected, Equation (2) remains valid for all $t = 1, \ldots, T$. Otherwise, we can use the non-linear specification in Equation (3) to identify and estimate post-intervention counterfactual outcomes after t_0+1.

Finally, the potential outcome model defined by Equation (2) does not assume stationar-

ity of the outcome variable. In similar contexts, studies often adopt a weaker assumption of conditional stationarity (D'Haultfœuille et al., 2023; Hoderlein and White, 2012). Formally, let $f_{\mathrm{Y}_{i,t}(0)|\,\mathrm{Y}^{(p)}_{i,t-1}(0),\mathbf{X}^{(q)}_{i,t}}$ be the conditional density of the potential outcomes under control. Conditional stationarity assumes that this density remains unchanged one step ahead, i.e., $f_{\mathrm{Y}_{i,t}(0)|\,\mathrm{Y}^{(p)}_{i,t-1}(0),\mathbf{X}^{(q)}_{i,t}} = f_{\mathrm{Y}_{i,t+1}(0)|\,\mathrm{Y}^{(p)}_{i,t}(0),\mathbf{X}^{(q)}_{i,t+1}}$. In other words, its functional form, which matches the form of the error term, is time-invariant, although the conditional mean can still be time-varying.[13] This means that the error term would be distributed in the same way during the post-intervention period. In our case, as we explain in Section 2.2, we target the causal effect at the individual level and specifically its expected value conditional on past information. Therefore, for identification, estimation, and inference, we do not require the error distribution to remain unchanged after the intervention. Instead, it is sufficient that the error term remains uncorrelated and has a mean of zero, a condition already embedded in Assumptions 4 and 5. In other words, we only require that, in the absence of the policy, the potential outcomes model would continue to be correctly specified. This implies the absence of other unforecastable shocks or policies affecting the outcome of interest in the post-intervention window. While it is not possible to explicitly test for this, we can assess how well the model fits the pre-intervention data via the rigorous panel CV procedure (see Section 3).

2.2 Causal estimands

We introduce novel estimands that could be of general interest for observational studies based on panel data in the absence of control units. The causal estimands discussed here pertain to Case (i) defined above, namely, when the treatment affects all, or most, of the available units, simultaneously. In Supplemental Appendix A, we also define causal estimands for the Case (ii) scenario with violations of the no-interference assumption.

We begin with the unconditional individual causal effect estimand.

Definition 1. *For any $k \geq 1$, the unconditional causal effect of the policy on unit i at time $t_0 + k$ is,*

$$\Delta_{i,t_0+k} = \mathrm{Y}_{i,t_0+k}(1) - \mathrm{Y}_{i,t_0+k}(0). \tag{4}$$

Note that this estimand arises naturally as a consequence of Assumption (2). However, identification, estimation, and inference of this causal effect would require stringent assumptions (e.g., strong stationarity of the potential outcome distribution). Therefore, our focus is on its conditional mean.

[13] Consider this simple (non-linear) example: $Y_t = \sin Y_{t-1}^2 + \varepsilon_t$ where $\varepsilon_t \sim f(0, \sigma_\epsilon^2)$. Under conditional stationarity, we have that $Y_t|Y_{t-1} \sim f(\sin Y_{t-1}^2, \sigma_\epsilon^2)$, $Y_{t+1}|Y_t \sim f(\sin Y_t^2, \sigma_\epsilon^2)$ and so on. Notice that while the conditional mean is time-varying, the form of the distribution $f(\cdot)$ remains the same. More generally, in our setting of interest, a short panel with small T and large N, non-stationarity is not an issue of particular concern (Baltagi, 2008). Nevertheless, as we will show below, the simulation results reported in Supplemental Appendix D demonstrate that even in the case of an explosive autoregressive coefficient (i.e., when the coefficient of the past lag of the outcome Y_{t-1} is set at $\phi = 1.2$), both the bias and the coverage rates of the MLCM are in line with the results obtained for stationary processes (i.e., when $\phi = 0.8$).

Definition 2. *For any $k \geq 1$ and some $p \in \{0, \ldots, t_0 - 1\}$, $q \in \{0, \ldots, t_0 + k - 1\}$ with $Q = \max\{k-1, q\}$, the expected individual effect of the policy, conditional on past outcomes and covariates, at time $t_0 + k$ is,*

$$\begin{aligned}\tau_{i,t_0+k} &= \mathrm{E}[(\mathrm{Y}_{i,t_0+k}(1) - \mathrm{Y}_{i,t_0+k}(0)) \,|\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(Q)}_{i,t_0+k}] \\ &= \mathrm{Y}_{i,t_0+k}(1) - \mathrm{E}[\mathrm{Y}_{i,t_0+k}(0) |\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(Q)}_{i,t_0+k}] \\ &= \mathrm{Y}_{i,t_0+k}(1) - \mathrm{Y}_{i,t_0+k|t_0}(0). \end{aligned} \tag{5}$$

This estimand emphasizes that, because the potential outcomes are temporally related, the causal effect of the policy should be defined conditionally on past information. Furthermore, in many situations, the full distribution of Δ_{i,t_0+k} is not of interest unless we specifically target causal effects at the median or other quantiles. Typically, the mean effect is the primary concern, making τ_{i,t_0+k} a more suitable summary measure of policy effects. The following theorem establishes the identification conditions for the individual effect τ_{i,t_0+k}.

Theorem 1. *For $k = 1$, τ_{i,t_0+k} is identified under Assumptions 3 and 4. Under a linear potential outcomes model, the identifying conditions for $k \geq 2$, are the same as for $k = 1$. If the potential outcomes model is non-linear, τ_{i,t_0+k} is identified under Assumptions 3 and 5.*

Proof. We illustrate the proof in a simple case where the potential outcome absent the policy depends only on one previous lag and contemporaneous covariates, i.e., Equations (2) and (3) become, respectively,

$$\mathrm{Y}_{i,t}(0) = h(\mathrm{Y}_{i,t-1}(0), \mathbf{X}_{i,t}) + \epsilon_{i,t}$$

$$\mathrm{Y}_{i,t_0+k}(0) = g_k(\mathrm{Y}_{i,t_0+k-1|t_0}(0), \ldots, \mathrm{Y}_{i,t_0+1|t_0}(0), \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+k}) + \epsilon_{i,t}$$

corresponding to the case $p = q = 0$. The general proof is reported in Supplemental Appendix B. The proof proceeds by induction and is organized distinguishing the two cases $k = 1$ and $k = 2$.

- $k = 1$. In the expression $\tau_{i,t_0+1} = \mathrm{Y}_{i,t_0+1}(1) - \mathrm{Y}_{i,t_0+1|t_0}(0)$, the first term $\mathrm{Y}_{i,t_0+1}(1)$ is the observed outcome under the policy and thus is immediately identified. Since we are in the case $p = q = 0$, we have that $Q = \max\{k-1, q\} = 0$, so the second term can be written as

$$\begin{aligned}\mathrm{Y}_{i,t_0+1|t_0}(0) &= \mathrm{E}[\mathrm{Y}_{i,t_0+1}(0) |\, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+1}] && \text{Def. of } \mathrm{Y}_{i,t_0+1|t_0}(0) \text{ for } p = Q = 0 \\ &= \mathrm{E}[h(\mathrm{Y}_{i,t_0}(0), \mathbf{X}_{i,t_0+1}) |\, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+1}] && \text{Assumption 4 - Eq.(2)} \\ &= h(y_{i,t_0}(0), \mathbf{x}_{i,t_0+1}) = y_{i,t_0+1|t_0}(0). \end{aligned}$$

  The last expression follows from the fact that $h(\cdot)$ is deterministic once we observe $\mathrm{Y}_{i,t_0}(0) = y_{i,t_0}(0)$ and $\mathbf{X}_{i,t_0+1} = \mathbf{x}_{i,t_0+1}$. As a result, $\mathrm{Y}_{i,t_0+1|t_0}(0)$ is identified from observed data, and so is τ_{i,t_0+1}. Notice that in this proof, we never used the linearity

of the potential outcome model. Thus, the proof at $k = 1$ is the same even when $h(\cdot)$ is non-linear.

- $k = 2$. In the expression $\tau_{i,t_0+2} = \mathrm{Y}_{i,t_0+2}(1) - \mathrm{Y}_{i,t_0+2|t_0}(0)$, the first term $\mathrm{Y}_{i,t_0+2}(1)$ is the observed outcome under the policy. Thus, the identification of the causal effect depends solely on the second term, $\mathrm{Y}_{i,t_0+2|t_0}(0)$. We now distinguish between a linear and a non-linear model specification:

- Linear case. Assume the following linear model $h(\mathrm{Y}_{i,t-1}(0), \mathbf{X_t}) = b_1\, \mathrm{Y}_{i,t-1}(0) + \mathbf{b_2}\mathbf{X}_{i,t}$, where $\mathbf{b_2}$ is an $m-$dimensional vector of covariates' coefficients. Since we are in the case $p = q = 0$, we have that $Q = \max\{1, 0\} = 1$, so the term $\mathrm{Y}_{i,t_0+2|t_0}(0)$ can be written as,

$$
\begin{aligned}
\mathrm{Y}_{i,t_0+2|t_0}(0) &= \mathrm{E}[\mathrm{Y}_{i,t_0+2}(0)|\, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] && \text{Def. } \mathrm{Y}_{i,t_0+2|t_0}(0) \text{ for } p = 0, Q = 1 \\
&= \mathrm{E}[h(\mathrm{Y}_{i,t_0+1}(0), \mathbf{X}_{i,t_0+2})|\, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] && \text{Assumption 4 - Eq.(2)} \\
&= \mathrm{E}[b_1\, \mathrm{Y}_{i,t_0+1}(0) + \mathbf{b}_2\mathbf{X}_{i,t_0+2}|\, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] && \text{Linear model} \\
&= b_1\, \mathrm{E}[\mathrm{Y}_{i,t_0+1}(0)|\, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+1}] + \mathbf{b}_2\mathbf{x}_{i,t_0+2} && \text{Assumption 3} \\
&= b_1 y_{i,t_0+1|t_0}(0) + \mathbf{b}_2\mathbf{x}_{i,t_0+2}.
\end{aligned}
$$

- Non-linear case. The identification of $\mathrm{Y}_{i,t_0+2|t_0}(0)$ can be based on the formulation for the post-intervention potential outcome model defined by Equation (3),

$$
\begin{aligned}
\mathrm{Y}_{i,t_0+2|t_0}(0) &= \mathrm{E}[\mathrm{Y}_{i,t_0+2}(0)|\, \mathrm{Y}_{i,t_0}(0), \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] && \text{Def. } \mathrm{Y}_{i,t_0+2|t_0}(0) \text{ for } p = 0, Q = 1 \\
&= \mathrm{E}[g_2(\mathrm{Y}_{i,t_0+1|t_0}(0), \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2})|\, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] && \text{Assumption 5 - Eq.(3)} \\
&= \mathrm{E}[g_2(h(\mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+1}), \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2})|\, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] \\
&= g_2(y_{i,t_0+1|t_0}(0), y_{i,t_0}, \mathbf{x}_{i,t_0+2}).
\end{aligned}
$$

The last expression follows from the fact that $g_2(\cdot)$ is deterministic once we observe $\mathrm{Y}_{i,t_0} = y_{i,t_0}$, $\mathbf{X}_{i,t_0+1} = \mathbf{x}_{i,t_0+1}$ and $\mathbf{X}_{i,t_0+2} = \mathbf{x}_{i,t_0+2}$. In addition, even though Assumption 3 was not explicitly mentioned in this part of the proof, notice that it is implicit in Assumption 5, as by Equation (3), the potential outcomes never depend on future covariates.

The proof for a generic time $t_0 + k$ follows analogously by induction. □

The following estimands are unit-averages of the individual effects defined above. For the reasons previously stated, we focus on averaging the conditional individual effects, which can be considered the panel analog of the ATE. We also frame them from a finite-sample perspective, as the MLCM is designed for statistical and econometric settings where the treatment affects, either directly or indirectly, the entire population under study.[14]

[14] For further details on the difference between the finite-sample and superpopulation perspectives in causal

Definition 3. *For any $k \geq 1$, the ATE at time $t_0 + k$ is defined as the average of the individual effects across all i units in the panel, with $i = 1, \ldots, N$,*

$$\tau_{t_0+k} = \frac{1}{N} \sum_{i=1}^{N} \tau_{i,t_0+k} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{Y}_{i,t_0+k}(1) - \mathrm{Y}_{i,t_0+k|t_0}(0). \tag{6}$$

Note that in this benchmark scenario (Case (i) where only treated units are available), the term τ_{t_0+k} corresponds to the Average Treatment effect on the Treated (ATT).

The subsequent estimand measures the average individual effect within a subset of units with the same values of selected covariates. This is commonly referred to as the CATE, and it is of particular interest when there is reason to believe that the intervention has produced heterogeneous effects on different subpopulations of units as defined by their distinct characteristics. Following existing literature on CATE in high-dimensional settings with a mix of discrete and continuous covariates (Fan et al., 2022; Knaus et al., 2021; Chernozhukov et al., 2018), we focus on its low-dimensional summary called "group-average treatment effect".

Definition 4. *Let $G_{i,t}$ denote a set of individual characteristics and indicate with N_g the number of units in the population having $G_{i,t} = g$. For any $k \geq 1$, the CATE at time $t_0 + k$ is defined as,*

$$\tau_{t_0+k}(g) = \frac{1}{N_g} \sum_{i:G_{i,t_0+k}=g} \tau_{i,t_0+k}. \tag{7}$$

Note that the set of candidate conditioning variables $G_{i,t}$ used to determine CATEs does not necessarily have to match those utilized for counterfactual forecasting. This is because the variables that predict outcomes can, and often do, differ at least partially from those that predict treatment effect heterogeneity. We also remark that in a general panel setting with more than one post-treatment periods, the above definitions imply the existence of vectors representing estimated average effects, i.e., $(\tau_{t_0+1}, \ldots, \tau_T)$ and $(\tau_{t_0+1}(g), \ldots, \tau_T(g))$. Furthermore, researchers are sometimes also interested in temporal aggregations of such effects.

Definition 5. *The temporal average ATE and CATE are defined, respectively, as,*

$$\bar{\tau} = \frac{1}{T - t_0} \sum_{k=1}^{T-t_0} \tau_{t_0+k} \qquad\qquad \bar{\tau}(g) = \frac{1}{T - t_0} \sum_{k=1}^{T-t_0} \tau_{t_0+k}(g)$$

We now introduce the following Theorem.

Theorem 2. *For $k = 1$, ATE and CATE are identified under Assumptions 1, 3, and 4. Under a linear potential outcomes model, the identifying conditions for $k \geq 2$, are the same as for $k = 1$; if the potential outcomes model is non-linear, ATE and CATE are identified under Assumptions 1, 3, and 5.*

inference, see Imbens and Rubin (2015). In observational panel settings, only Rambachan and Roth (2023) defined causal estimands when the entire population is observed, but their approach relies on control units.

Proof. The proof follows directly from the previous one. By Definitions 3 and 4, we have that both τ_{t_0+k} and $\tau_{t_0+k}(g)$ are aggregations of τ_{i,t_0+k} across different sets of units (all the N units in the panel for ATE, subgroups sharing the same characteristics for CATE). Thus, their unit-average is also identified under the same set of assumptions. We also add Assumption 1 because in this case we are considering aggregation of units, so the effects must be comparable (if Assumption 1 is violated, we would estimate the effects of different forms of treatment that cannot be aggregated). This logic extends to the temporal average effects in Definition 5. □

2.3 Estimators

We now introduce estimators of the causal quantities defined in Section 2.2. Note that imputing future counterfactual outcomes based on past information and covariates is essentially a forecasting problem. Therefore, in the following definitions, the estimator $\widehat{\mathrm{Y}}_{t_0+k|t_0}$ represents the k-step-ahead forecast based on the model trained in the pre-intervention period. We first define an estimator for the individual effect at time t_0+1.

Definition 6. *For some $p = 0, \ldots, t_0-1$ and $q = 0, \ldots, t_0$ with $Q = \max\{0, q\}$, an estimator of τ_{i,t_0+1} is the difference between the observed outcome under the policy and the estimated 1-step-ahead forecast, conditional on past information and contemporaneous covariates,*

$$\begin{aligned}
\widehat{\tau}_{i,t_0+1} &= \mathrm{Y}_{i,t_0+1}(1) - \widehat{\mathrm{Y}}_{i,t_0+1|t_0}(0) \\
&= \mathrm{Y}_{i,t_0+1}(1) - \mathrm{E}\left[\widehat{h}(\mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(q)}_{i,t_0+1}) \middle| \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(Q)}_{i,t_0+1}\right] \qquad (8) \\
&= \mathrm{Y}_{i,t_0+1}(1) - \widehat{h}(y^{(p)}_{i,t_0}(0), \mathbf{x}^{(q)}_{i,t_0+1}) \qquad (9)
\end{aligned}$$

where $\widehat{h}(\cdot)$ denotes that the 1-step-ahead forecast is based on the optimized ML parameters.

A detailed description of the estimation algorithm for $h(\cdot)$ will be given in Section 3. Although our empirical application only evaluates causal effects at time t_0+1, one might also be interested in estimating effects over multiple post-intervention periods, as in our simulations and replication study. There are two main strategies for generating multi-step-ahead forecasts (Chevillon, 2007; Petropoulos et al., 2022): the *recursive* approach, where each forecast is defined and estimated using previous forecasts, and the direct approach, where forecasts are produced by estimating separate models for each forecast horizon. In the case of linear models, the recursive approach yields more efficient parameter estimates, especially when the model is correctly specified and for a long forecasting horizon (Pesaran et al., 2011). However, with non-linear models, the recursive approach is known to be asymptotically biased, making the direct approach often preferable (Bontempi and Taieb, 2011). One criticism of the direct forecasting approach is that it assumes independence between the forecasts, as each one is based on a separate model. *Hybrid* forecasts, instead, represent a combination of the recursive and direct approaches, effectively integrating the strengths of

both strategies (Petropoulos et al., 2022). Like the direct approach, hybrid strategies require estimating separate models for each forecasting horizon. However, similar to the recursive approach, each model also includes the direct forecast from the previous step. This method offers full flexibility by allowing different models for each time horizon, while still accounting for the dependency between direct forecasts. This is exactly what Equation (3) captures: each $\mathrm{Y}_{i,t_0+k-1|t_0}(0), \ldots, \mathrm{Y}_{i,t_0+1|t_0}(0)$ is a direct forecast (i.e., the expected potential outcome up to $k-1$ steps ahead from t_0), and $\mathrm{Y}_{i,t_0+k}(0)$ is then modeled as a flexible function of these direct forecasts, along with observed past lags and covariates.

Therefore, depending on whether the potential outcome model is linear or non-linear, the next definition outlines two different estimators for multi-step-ahead evaluations.

Definition 7. *For $k \geq 2$ and for some $p = 0, \ldots, t_0 - 1$ and $q = 0, \ldots, t_0 + k - 1$ with $Q = \max\{k-1, q\}$, an estimator for τ_{i,t_0+k} when $h(\cdot)$ is linear is given by,*

$$\widehat{\tau}^{lin}_{i,t_0+k} = \mathrm{Y}_{i,t_0+k}(1) - \widehat{\mathrm{Y}}_{i,t_0+k|t_0}(0) \tag{10}$$

$$= \mathrm{Y}_{i,t_0+k}(1) - \mathrm{E}\left[\widehat{h}(\mathrm{Y}^{(p)}_{i,t_0+k-1}(0), \mathbf{X}^{(q)}_{i,t_0+k}) \middle| \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(Q)}_{i,t_0+k}\right] \tag{11}$$

whereas, if $h(\cdot)$ is non-linear an estimator for τ_{i,t_0+k} is given by,

$$\widehat{\tau}^{nl}_{i,t_0+k} = \mathrm{Y}_{i,t_0+k}(1) - \mathrm{E}\left[\widehat{g}_k(\mathrm{Y}_{i,t_0+k-1|t_0}(0), \ldots, \mathrm{Y}_{i,t_0+1|t_0}(0), \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(q)}_{i,t_0+k}) \middle| \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(Q)}_{i,t_0+k}\right] \tag{12}$$

where $\widehat{h}(\cdot)$ and $\widehat{g}_k(\cdot)$ denote that the k-step-ahead forecasts are based on the optimized ML parameters.

A detailed description of the estimation algorithm for $g_k(\cdot)$ will be given in Section 3. To clarify how multi-step-ahead forecasts are computed and to highlight the differences between the linear and non-linear case, we provide the following example.

Example 1. *Consider the simple case $p = q = 0$, and let $\widehat{h}(\mathrm{Y}_{i,t_0+k-1}(0), \mathbf{X}_{i,t_0+k}) = \widehat{b}_1\, \mathrm{Y}_{i,t_0+k-1}(0) + \widehat{\mathbf{b}}_2 \mathbf{X}_{i,t_0+k}$. For $k = 2$, the 2-step-ahead forecast $\widehat{\mathrm{Y}}_{i,t_0+2|t_0}(0)$ is given by*

$$\begin{aligned}
\widehat{\mathrm{Y}}_{i,t_0+2|t_0}(0) &= \mathrm{E}[\widehat{\mathrm{Y}}_{i,t_0+2}(0) | \, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] \\
&= \mathrm{E}[\widehat{h}(\widehat{\mathrm{Y}}_{i,t_0+1}(0), \mathbf{X}_{i,t_0+2}) | \, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] \\
&= \widehat{b}_1 \, \mathrm{E}[\widehat{\mathrm{Y}}_{i,t_0+1}(0) | \, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+1}] + \widehat{\mathbf{b}}_2 x_{i,t_0+2} \\
&= \widehat{b}_1 \widehat{\mathrm{Y}}_{i,t_0+1|t_0} + \widehat{\mathbf{b}}_2 x_{i,t_0+2} = \widehat{h}(\widehat{\mathrm{Y}}_{i,t_0+1|t_0}, \mathbf{x}_{i,t_0+2})
\end{aligned}$$

Multi-step-ahead forecasts in the linear case are then computed by recursive substitution, plugging-in the previous forecast (and updated covariates) in the already estimated model $\widehat{h}(\cdot)$. Now, assume that the panel CV procedure selects a non-linear model for the pre-intervention period. The 1-step-ahead forecast $\widehat{\mathrm{Y}}_{i,t_0+1|t_0}(0)$ can be performed easily by following Equation (8). At $k = 2$, we then re-estimate a model $\widehat{g}_2(\widehat{\mathrm{Y}}_{i,t_0+1|t_0}(0), \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1})$ that includes the forecasted counterfactual outcome, as in Equation (3). Thus, the 2-step-ahead forecast

$\widehat{\mathrm{Y}}_{i,t_0+2|t_0}(0)$ *is given by,*

$$\begin{aligned}\widehat{\mathrm{Y}}_{i,t_0+2|t_0}(0) &= \mathrm{E}[\widehat{\mathrm{Y}}_{i,t_0+2}(0)|\,\mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] \\ &= \mathrm{E}[\widehat{g}_2(\widehat{\mathrm{Y}}_{i,t_0+1|t_0}, \mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1})|\,\mathrm{Y}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] \\ &= \widehat{g}_2(\widehat{h}(y_{i,t_0}, \mathbf{x}_{i,t_0+1}), y_{i,t_0}, \mathbf{x}_{i,t_0+2}).\end{aligned}$$

Building on Equations (8-12), the next definition summarizes the ATE and CATE estimators under the MLCM.

Definition 8. *For any positive integer k, finite-sample estimators for the ATE and CATE at time $t_0 + k$ (6) are, respectively,*

$$\widehat{\tau}_{t_0+k} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\tau}_{i,t_0+k} \qquad\qquad \widehat{\tau}_{t_0+k}(g) = \frac{1}{N_g}\sum_{i:G_{i,t_0+k}=g}\widehat{\tau}_{i,t_0+k}. \tag{13}$$

Finally, the estimators for the temporal average ATE and CATE are, respectively,

$$\widehat{\bar{\tau}} = \frac{1}{T-t_0}\sum_{k=1}^{T-t_0}\widehat{\tau}_{t_0+k} \qquad\qquad \widehat{\bar{\tau}}(g) = \frac{1}{T-t_0}\sum_{k=1}^{T-t_0}\widehat{\tau}_{t_0+k}(g). \tag{14}$$

Inference on the estimated causal effects defined above is conducted using block-bootstrap. Refer to Supplemental Appendix C for a detailed description of the block-bootstrap algorithms used to derive confidence intervals for the ATE and CATEs. Finally, we stress that deriving the theoretical properties of the above-defined estimators would deviate from the intended nature of the methodology, since the MLCM can be implemented with any supervised ML algorithm, and it is unknown in advance which one will outperform the others.[15] Instead, in Section 3.3, we perform a simulation study showing that the MLCM can achieve forecast unbiasedness and is able to detect causal effects even in challenging econometric environments with non-linearities and irrelevant covariates included in the model specification.

3 The Machine Learning Control Method

3.1 Departures from the standard machine learning approach

Supervised ML techniques primarily aim to minimize the out-of-sample prediction error, generalizing well on unseen data. The degree of flexibility is the result of a trade-off: increased flexibility can enhance in-sample fit but may diminish out-of-sample fit due to overfitting. ML algorithms tackle this trade-off by relying on empirical tuning to choose the optimal level of complexity.

[15] To use Breiman's words: "Nowhere is it written on a stone tablet what kind of model should be used to solve problems involving data." (Breiman, 2001).

The standard ML approach is to randomly split the sample into two sets, containing, for instance, 2/3 and 1/3 of observations. One then uses the first set to train ML algorithms (training set) and the second to test them (testing set). This introduces a "firewall" principle: none of the data involved in generating the prediction function is used to evaluate it (Mullainathan and Spiess, 2017). The out-of-sample performance of the model on the unseen (held-out) data of the testing set can be considered a reliable measure of the "true" performance on future data. In order to solve the bias-variance trade-off and prevent overfitting, one can rely on automatic tuning using tools such as random k-fold CV on the training sample to select the best-performing values of the tuning parameters in terms of an *a priori* defined metric, such as the MSE.

We depart from this standard ML routine and reorient it towards the counterfactual forecasting goal. First, we do not randomly split the data, but we train, tune, and evaluate the models only on the pre-treatment data (Design Stage); then, we use the final selected model to forecast counterfactual post-treatment outcomes. In this forecasting perspective, the unseen data on which the ML models must generalize well are not the outcomes of different units (as in typical out-of-sample prediction tasks), but future observations of the outcome for the same set of units employed to train the models. Stated differently, the aim is to make ML models learn as best as possible the pre-intervention outcome trend for each treated unit, so as to predict the best possible counterfactual outcome under the no-treatment scenario. The key implication of this unconventional ML setup is the shift in focus: the primary concern becomes ensuring unbiasedness in forecasts, rather than focusing only on forecast accuracy.

Second, and related, we do not carry out hyperparameter tuning and model selection with random k-fold CV. The panel dimension creates an additional challenge regarding how to implement CV, because standard CV does not account for the temporal structure of the data (Arkhangelsky and Imbens, 2024). To address this challenge, we introduce a resampling technique suited for forecasting tasks on panel data—panel CV—which is described below.

Finally, a key concern in ML regards the trade-off between accuracy and interpretability. Such a trade-off is relevant when ML is used for tasks that take into consideration transparency aspects. In principle, our method can be used with any supervised ML routine, including black-box techniques like deep neural networks. However, maintaining some degree of transparency can be important, because higher interpretability of the estimated counterfactuals bolsters the credibility of the proposed approach (Abadie, 2021). In practice, users should decide on the basis of a comparative assessment across a mix of models characterized by different layers of complexity by balancing any improvements in performance from complex models against the loss of interpretability associated with their use.

3.2 Implementation

The MLCM is rooted in the idea of developing a ML forecasting model that can closely reproduce the outcome trajectories of treated units in the pre-intervention window, so that any post-intervention divergence between the observed and forecasted outcomes can be attributed to the treatment under the identification assumptions. To achieve this, the implementation of the MLCM requires ten empirical steps, which are divided into the Design Stage and the Analysis Stage. The full process is summarized in Box 1 and described in detail below.

In the Design Stage, the first step involves the selection of supervised ML algorithms that will play the horse-race of performance testing on the pre-treatment data. This competition among multiple methods is conceptually analogous to the way the best ML learner is selected in the double/debiased ML method (Chernozhukov et al., 2018).[16]

In the second step, we recommend deploying some tweaks involving feature pre-selection and engineering drawing from consolidated practices in applied predictive modeling (Kuhn and Johnson, 2013). A crucial aspect of this pre-processing is the strategy for selecting predictors. There are two main schools of thought: those who advocate for purely data-driven selection argue that one should build a dataset as large as possible, and then let the algorithm autonomously decide which variables matter for the prediction task. Others stress the importance of subject matter knowledge: the researcher should select *ex ante* the relevant predictors, and then feed only those to the algorithm. The rationale is that subject matter knowledge can separate meaningful from irrelevant information, eliminating detrimental noise and enhancing the underlying signal (Kuhn and Johnson, 2013). We propose a hybrid approach: build a large initial dataset on the basis of domain knowledge, then adopt preliminary and data-driven variable selection criteria to drop non-informative predictors.

As outlined in the causal framework, counterfactual forecasting is carried out by using an information set mainly comprising lagged values of outcomes and covariates. However, determining the optimal number of lags to include is an empirical question. We recommend including at least two lagged values of both outcomes and covariates, and then using a data-driven approach to select a subsample of the most relevant features. The latter step allows for a reduction of the risk of overfitting and degradation of performances as well as facilitating interpretability. Finally, feature engineering and data pre-processing matter too because how the predictors enter into the model is also important (Kuhn and Johnson, 2013).

To carry out model selection and validation on the pre-treatment sample, we propose a panel CV approach. Using an alternative CV procedure is necessary since ML methods do not natively handle longitudinal data and are designed for predicting rather than forecasting.

[16]The MLCM can leverage any supervised ML algorithm, including deep learning techniques, some of which, such as RNNs and transformers, are explicitly designed to learn temporal dynamics; however, they cannot be employed in the panel settings we focus on (small T, large N), as they require a large number of time periods to be applicable. It is also possible to stack several different ML algorithms and form complex ensemble learners, but this would come at the cost of a substantial loss in transparency.

Box 1: MLCM implementation

Preliminary

0. **Data splitting.** Split the full sample based on the treatment date: employ only pre-treatment data throughout the Design Stage; use post-treatment data only for estimating causal effects in the Analysis Stage.

Design Stage

1. **Algorithm selection.** Select one or more supervised ML algorithms.
2. **Principled input selection.** Build a large initial dataset on the basis of domain knowledge. To maximize forecasting performances, it is possible to use feature engineering, feature selection, and heuristic rules to pre-process the data and select a subsample of the most relevant predictors.
3. **Panel cross-validation.** For each selected algorithm, tune hyperparameters via panel CV (see Figure 1 and Algorithms 1 and 2).
4. **Performance assessment.** Assess average performance metrics (e.g., MSE) for all the selected algorithms and check what is the best-performing version of the MLCM.
5. **Diagnostic and placebo tests.** Implement a battery of diagnostic and placebo tests to bolster the credibility of the research design.

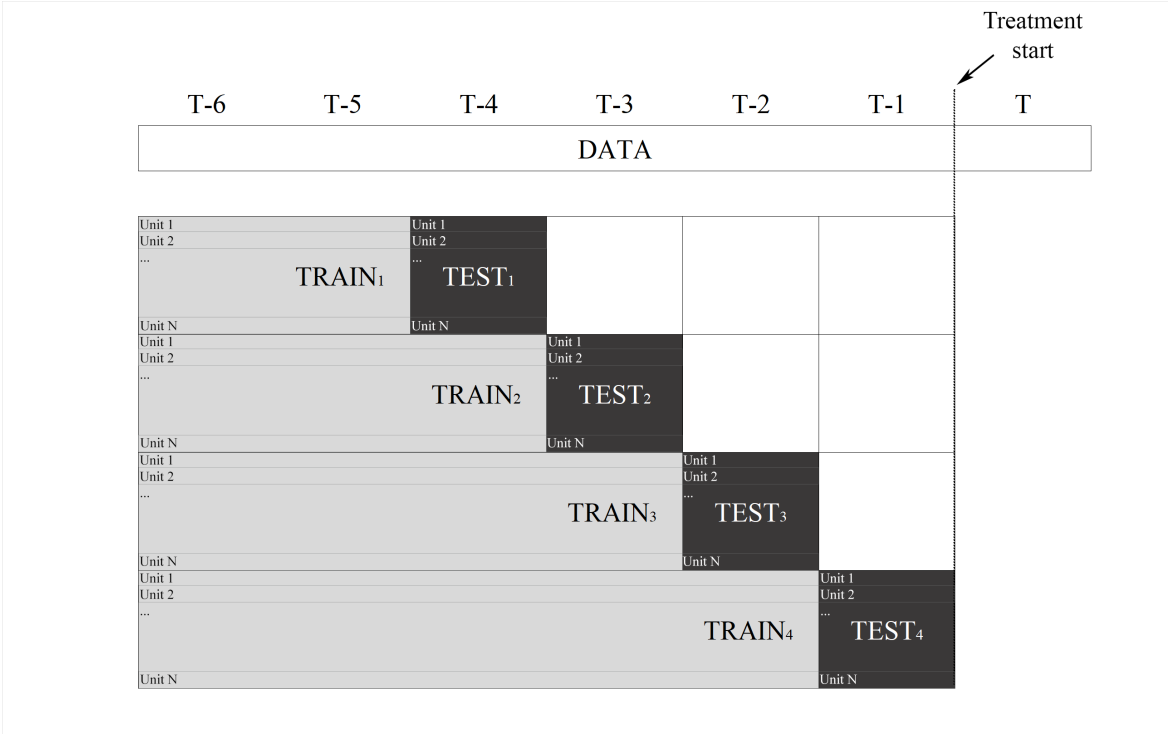
Analysis Stage

6. **Final model selection.** On the basis of the comparative performance assessment in the Design Stage, pick the best-performing model and use that for the Analysis Stage. Start by re-training the model on the full pre-treatment sample using the hyperparameter(s) selected in the Design Stage.
7. **Counterfactual forecasting.** For each unit i, forecast the post-treatment counterfactual outcome $\widehat{\mathrm{Y}}_{i,t_0+k|t_0}(0)$. In case of a large-scale shock affecting important covariates, either rely solely on their past lags or repeat steps 1–6 to forecast $\widehat{\mathbf{X}}_{i,t_0+k|t_0}(0)$ and use these values to improve the forecast of $\widehat{\mathrm{Y}}_{i,t_0+k|t_0}(0)$. Leverage Explainable Artificial Intelligence tools to enhance the model's explainability and transparency of the estimated counterfactual.
8. **Estimation of treatment effects.** For each unit i, estimate the individual treatment effect as in Equation (8) by taking the difference between the observed post-treatment outcome Y_{i,t_0+1} and the ML-generated potential outcome $\widehat{\mathrm{Y}}_{i,t_0+1|t_0}(0)$. For $k \geq 2$, if the selected ML algorithm in Step 4 of the Design Stage is linear, estimate $\hat{\tau}_{t_0+k}$ as in Equation (10), otherwise use Equation (12). Estimate other causal estimands, such as the ATE from Equation (13) or the temporal ATE from Equation (14), by aggregating the individual estimates.
9. **Treatment effect heterogeneity.** To uncover heterogeneity, data-driven CATEs can be estimated as in Equation (13) via a regression tree analysis with the individual treatment effects as the outcome variable and a set of variables potentially associated with treatment effect heterogeneity.
10. **Inference.** Compute standard errors for the ATE and CATEs via block-bootstrap.

Our panel CV approach adapts time series CV based on expanding training windows to a panel setting.[17] The intuition is provided in Figure 1 and constitutes an adaptation from

[17]Here we use an expanding window (as we are in a short-panel setting), but the procedure can also be implemented using a rolling window approach, which might be more appropriate in specific cases (e.g., with long panels or non-stationarity of the outcome).

Figure 1: Panel cross-validation (one-step-ahead forecasting). *Notes*: this procedure is carried out using exclusively pre-treatment data. Light gray observations form the training sets; dark gray ones constitute the test sets. All the training windows also include past outcomes in the set of predictors.



Hyndman and Athanasopoulos (2021).

We start from a set of candidate algorithms $\mathcal{H} = (h_1, \ldots, h_J)$, each having its own set of parameters $\theta^{(j)}$ and hyperparameters $\gamma^{(j)}$. For example, in the empirical application, we use four learners: LASSO, PLS, random forest and stochastic gradient boosting. For each learner, we carry out panel CV as detailed in Algorithm 1. In short, the ML algorithms are trained repeatedly using an expanding window approach, with hyperparameters tuned at each step to minimize forecast errors. This sequential procedure ensures that, for each unit, there are no "future" observations in the training set and no "past" observations in the validation set. Note that the same procedure can be applied to forecast post-treatment values of important covariates affected by the intervention. We can assume that such covariates follow Equation (2) and use the ML algorithms and the panel CV routine to forecast counterfactual values of $\widehat{\mathbf{X}}_{i,t_0+k|t_0}(0)$, which we can then incorporate into the prediction of $\widehat{\mathrm{Y}}_{i,t_0+k|t_0}$ as in Liu et al. (2020).

To provide evidence about internal validity, we suggest running diagnostic checks (e.g., showing that the distribution of the pre-treatment forecasting errors is approximately Gaussian and centered around zero) as well as placebo tests, which have become a key device for assessing the credibility of research designs in observational settings (Eggers et al., 2024). Following Liu et al. (2022), panel placebo tests can be implemented by hiding one or more periods of observations right before the onset of the treatment and using a model trained on the rest of the pre-treatment periods to predict the untreated outcomes of the held-out

Algorithm 1 Panel CV for one-step-ahead forecasting

Require: $\mathcal{H} = (h_1, \ldots, h_J)$, t_0

1: **for** j in $1 : J$, where J is the total number of candidate algorithms **do**

2: **for** s in $1 : t_0 - 1$ **do**

3: Split the data into a training set $\mathcal{T} = \{\mathrm{Y}_{i,t}, Z_{i,t}\}_{t=1}^{s}$ and a validation set $\mathcal{V} = \{\mathrm{Y}_{i,t}, Z_{i,t}\}_{t=s+1}$, where $Z_{i,t} = \{\mathrm{Y}_{i,t-1}^{(p)}, \mathbf{X}_{i,t}^{(q)}\}$ is a set of features including past lags of the outcome and covariates that should be related to $\mathrm{Y}_{i,t}$.

4: Fit the model in the training set by optimizing a performance measure, e.g., minimizing MSE,

$$\hat{\theta}^{(j)} = \underset{\theta}{\operatorname{argmin}} \sum_{i,t \in \mathcal{T}} (\mathrm{Y}_{i,t} - h_j(Z_{i,t}; \gamma^{(j)}, \theta^{(j)}))^2$$

and compute the 1-step-ahead forecast, i.e., $\mathrm{E}[\mathrm{Y}_{i,s+1} \,|Z_{i,s+1}] = h_j(Z_{i,s+1}; \gamma^{(j)}, \hat{\theta}^{(j)})$.

5: Tune hyperparameters in the validation set by optimizing the one-step-ahead forecast error,

$$\hat{\gamma}^{(j)} = \underset{\gamma}{\operatorname{argmin}} \sum_{i} (\mathrm{Y}_{i,s+1} - h_j(Z_{i,s+1}; \gamma^{(j)}, \hat{\theta}^{(j)}))^2$$

6: Repeat steps 3-5 for all s up to $t_0 - 1$, the optimal values for the hyperparameters will be those that minimize the forecast errors across all horizons.

7: **end for**

8: Repeat steps 1-6 for all j and choose the algorithm in $\mathcal{H}$ which minimizes the one-step-ahead forecast error.

9: **end for**

period(s). If the identifying assumptions are valid, the differences between the observed and forecast outcomes in those periods should be close to zero.[18] Given that we estimate unit-level treatment effects, we are also able to test whether most unit-level placebo differences are close to zero. The Design Stage thus ends with a battery of performance, diagnostic, and placebo tests.

The Analysis Stage starts with final model selection and training: on the basis of the comparative performance assessment in the Design Stage, pick the best-performing model, re-train it on the full pre-treatment sample (using the hyperparameter values obtained in the Design Stage), then use it to forecast counterfactual outcomes in the post-intervention period. Once that is done, treatment effects for each unit are given by the difference between the post-treatment observed data and the corresponding ML-generated counterfactual forecasts. The ATE is the average of the individual effects.

When dealing with a single post-intervention period, step 7 of the Analysis Stage is straightforward. However, when performing a multi-step-ahead forecast using a data-driven ML routine, one must take special care to avoid including post-treatment outcomes in the prediction. Multi-step-ahead forecast becomes more challenging when the selected ML al-

[18] If large discrepancies, i.e., forecasting errors, are found at this stage, this would indicate pre-intervention shocks or policy changes occurring between pre-treatment periods. These factors should be accounted for in the model to bridge the gap between observed and forecasted pre-treatment outcomes.

gorithm is non-linear, as we cannot simply plug-in the forecasted value $\widehat{\mathrm{Y}}_{t_0+1|t_0}(0)$ into the already estimated model. As explained in Section 2.3, one solution is to re-estimate the potential outcome model for values of $k \geq 2$. Importantly, at the end of the panel CV procedure, we know the selected ML algorithm and can adjust our forecasting strategy accordingly. Specifically, we propose using Equation (3) to model post-intervention potential outcomes, as this approach has proven effective in identifying non-linear, multi-step-ahead causal effects, and estimate them with Equation (12). From a practical standpoint, this requires repeating the panel CV for each step of the forecasting horizon. The detailed procedure is described in Algorithm 2.

Algorithm 2 Panel CV for non-linear multi-step-ahead forecasting

Require: h^{win} (selected model after Algorithm 1).

1: Compute $\widehat{\mathrm{Y}}_{i,t_0+1|t_0} = \widehat{h}^{win}(y^{(p)}_{i,t_0}, \mathbf{x}^{(q)}_{i,t_0+1})$

2: **if** h^{win} is a non-linear model and the forecasting horizon $k \geq 2$ **then**

3: **for** $k = 2$ **do**

4: Start from a set of candidate algorithms $\mathcal{G} = (g_2^{(1)}, \ldots, g_2^{(j)})$

5: Split the data into a training set $\mathcal{T} = \{\mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(q)}_{i,t_0}\}$ and a validation set $\mathcal{V} = \{\widehat{\mathrm{Y}}_{i,t_0+1|t_0}, \mathbf{X}^{(q)}_{i,t_0+1}\}$.

6: Fit the model $g_2^{(j)}$ in the training set by optimizing a performance measure, e.g., minimizing MSE,

$$\hat{\theta}^{(j)} = \underset{\theta}{\operatorname{argmin}} \sum_{i,t_0 \in \mathcal{T}} (\mathrm{Y}_{i,t_0} - g_2^{(j)}(\mathrm{Y}^{(p)}_{t_0-1}, \mathbf{X}^{(q)}_{i,t_0}; \gamma^{(j)}, \theta^{(j)}))^2$$

and compute the 1-step-ahead forecast based on g_2, i.e., $g_2^{(j)}(\mathrm{Y}^{(p)}_{t_0}, \mathbf{X}^{(q)}_{i,t_0+1}; \gamma^{(j)}, \hat{\theta}^{(j)})$.

7: Tune hyperparameters in the validation set by optimizing the one-step-ahead forecasting error,

$$\hat{\gamma}^{(j)} = \underset{\gamma}{\operatorname{argmin}} \sum_{i} (\widehat{\mathrm{Y}}_{i,t_0+1|t_0} - g_2^{(j)}(\mathrm{Y}^{(p)}_{t_0}, \mathbf{X}^{(q)}_{i,t_0+1}; \gamma^{(j)}, \hat{\theta}^{(j)}))^2.$$

8: Choose the candidate algorithm in $\mathcal{G}$ which minimizes the one-step-ahead forecast error, retrain the model including the validation set and compute the one-step-ahead forecast $\widehat{\mathrm{Y}}_{t_0+2|t_0} = \widehat{g}_2(\widehat{\mathrm{Y}}_{t_0+1|t_0}, \mathrm{Y}^{(p)}_{t_0}, \mathbf{X}^{(q)}_{t_0+2})$.

9: Repeat steps 4-8 for all $k > 2$

10: **end for**

11: **end if**

During the counterfactual forecasting step, especially when opting for more complex and less transparent techniques, users should consider leveraging innovations in Interpretable Machine Learning and Explainable Artificial Intelligence (Molnar, 2020), such as model-agnostic Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017), to enhance explainability and transparency of the counterfactual building process (Abadie, 2021).

After estimating individual effects, data-driven CATEs can be computed via a regression

tree analysis. Specifically, this approach uses the estimated treatment effects as the outcome variable, regresses them on many potentially associated variables, and lets the algorithm pick the main predictors and their critical thresholds. The resulting tree reports the group-average treatment effects for all units in each terminal node. As for causal trees and forests (Athey and Imbens, 2016; Wager and Athey, 2018), this data-driven search for heterogeneity of causal effects removes a major degree of discretion because the researcher can only select the set of covariates that can be used by the tree to build the subgroups. However, the two approaches differ regarding both purpose and implementation: our data-driven technique is a post-estimation approach aimed at automatically recovering and visualizing the relevant heterogeneity dimensions, while causal trees and forests are counterfactual methods for the direct estimation of heterogeneous treatment effects, which they achieve by leveraging control units. Moreover, we are not interested in the out-of-sample performance of the regression tree, but in retrieving CATEs for the entire population of interest. To this end, there is no need to split the sample into training and testing sets or to prune the tree by adjusting the complexity parameter. However, it may be necessary to set a pre-specified minimum node size to preserve the interpretability of the resulting tree.

Finally, standard errors and confidence intervals for the ATE and CATEs are estimated through the block-bootstrap approach described in Supplemental Appendix C.[19]

3.3 Simulation study

To investigate the performance of the MLCM in detecting average treatment effects in panel datasets, we performed an extensive simulation study using different data-generating processes (linear and non-linear) and various combinations of pre- and post-intervention periods. We generated $1,000$ panel datasets, each consisting of 400 units and $T = 7$ or 12 time periods, according to the following two models,

$$\begin{aligned} \mathrm{Y}_{i,t}(0) &= \phi\, \mathrm{Y}_{i,t-1}(0) + \mathbf{X}_{i,t-1}\boldsymbol{\beta} + \epsilon_{i,t} && \text{(Linear)} \\ \mathrm{Y}_{i,t}(0) &= \sin\left\{\phi\, \mathrm{Y}_{i,t-1}(0) + \mathbf{X}_{i,t-1}\boldsymbol{\beta}\right\} + \epsilon_{i,t} && \text{(Non-Linear)} \end{aligned}$$

where: the error term $\epsilon_{i,t}$ is generated from a Normal distribution with standard deviation $\sigma_\epsilon = 2$, i.e., $\epsilon_{i,t} \sim N(0, 2)$, $\phi = 0.8$ is the autoregressive coefficient, $\boldsymbol{\beta} = (\beta^{(1)}, \ldots, \beta^{(11)})'$ is a $m \times 1$ vector of coefficients and $\mathbf{X}_{i,t-1} = \left(\mathrm{X}^{(1)}_{i,t-1}, \ldots, \mathrm{X}^{(11)}_{i,t-1}\right)$ is a $1 \times m$ vector of 11 predictors, both continuous and categorical, also containing interaction terms and correlated regressors. We allowed the covariates to vary in time and across units in the dataset by adding, respectively, a random term ν_t (which, as will become clear later, varies for different covariates) and a random term $u_i \sim N(1, 1)$. The latter term adds variability

[19] In staggered adoption settings, uncertainty quantification with block-bootstrapping can be extended to other estimands, such as group-time average treatment effects. See the replication exercise in Supplemental Appendix F.

(and thus, heterogeneity) between the units. In particular, the covariates are generated as follows: $X_{i,t}^{(1)} = 0.1t + u_i + \nu_t^{(1)}, \nu_t^{(1)} \sim N(0,1)$, $X_{i,t}^{(2)} = 0.1t + u_i + \nu_t^{(2)}, \nu_t^{(2)} \sim N(0,0.2)$, $X_{i,t}^{(3,4,5)} = u_i + \nu_t^{(3,4,5)}, \nu_t^{(3,4,5)} \sim MVN(0,\Sigma)$, $X_{i,t}^{(6)} = u_i - \nu_t^{(6)}, \nu_t^{(6)} \sim N(0,1)$, $X_{i,t}^{(7)} = \left(0.1t + \nu_t^{(1)}\right)^2 + u_i + \nu_t^{(7)}, \nu_t^{(7)} \sim N(0,0.2)$, $X_{i,t}^{(8)} \in \{0,1\}$, $X_{i,t}^{(9)} \in \{1,2,3\}$, $X_{i,t}^{(10)} = X_{i,t}^{(3)} \cdot X_{i,t}^{(9)}$, $X_{i,t}^{(11)} = X_{i,t}^{(2)} \cdot X_{i,t}^{(8)}$.[20] These choices regarding the data-generating process for the covariates included in our simulation study are driven by the goal of mirroring real-world empirical settings where the methodology might be applied. In addition, to better reflect a typical real-world scenario where only a subset of covariates is relevant, we set certain coefficients to zero: specifically, $\beta^{(1)}$, $\beta^{(8)}$, and $\beta^{(9)}$. The remaining coefficients are generated as follows: $\beta^{(2)} = \beta^{(6)} = \beta^{(10)} = 2$, $\beta^{(3)} = \beta^{(7)} = 1$, $\beta^{(4)} = 2.5$, $\beta^{(5)} = 0.1$ and $\beta^{(11)} = 1.5$. We also remark that, for computational reasons, in this simulation study we focus on a low-dimensional set of covariates. However, the MLCM is particularly well-suited for data-rich environments and sparse settings with many more covariates. Therefore, the simulation results provided here should be interpreted as conservative evidence regarding the performance of the MLCM. Finally, we assume exogeneity of the predictors in both the pre- and post-intervention periods (third part of Assumption 3). Table 1 below provides an overview of the generated datasets.

Table 1: First 9 observations from one of the datasets generated during the simulation study

Time	ID	Y	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$	$X^{(6)}$	$X^{(7)}$	$X^{(8)}$	$X^{(9)}$	$X^{(10)}$	$X^{(11)}$
1	1	16.57	1.71	0.9	1.38	3.61	3.03	0.42	1.09	0	1	1.38	0
2	1	43.62	1.62	1.53	2.4	3.33	3.73	1.14	1.24	0	3	7.19	0
3	1	73.62	2.83	1.71	3.13	3.44	3.99	3.89	3.37	1	1	3.13	1.71
4	1	89.04	1.16	1.91	1.53	3.35	2.25	1.14	1.25	1	3	4.58	1.91
5	1	170.17	2.61	1.62	1.99	1.47	3.46	1.23	3.79	0	2	3.98	0
6	1	153.05	1.43	1.63	1.3	3.3	3.2	1.22	1.65	0	1	1.3	0
7	1	133.43	0.3	1.8	0.46	1.43	3.38	1.32	1.14	1	3	1.38	1.8
1	2	16.88	1.42	1.14	0.23	2.93	3	0.6	2.11	1	3	0.7	1.14
2	2	45.33	1.66	1.75	2.94	3.2	3.57	0.56	1.13	0	2	5.87	0

At $t_0 + 1$ (corresponding to T = 5 in Table 1), we included a fictional intervention that increases the outcome for each unit by 2 standard deviations at time $t_0 + 1$, 1.5 standard deviations at time $t_0 + 2$, and 1 standard deviation at time $t_0 + 3$. In other words, the intervention has a decreasing effect over time. We made this choice because adding a unit-specific component in the covariates generates heterogeneity; therefore, the scale of each Y_i varies across the i's. We measured the performance of the MLCM in terms of both the bias

[20] The variance-covariance matrix of the multivariate normal distribution used to generate covariates 3–5 is set to $\Sigma = \begin{bmatrix} 1 & .5 & .7 \\ .5 & 1 & .3 \\ .7 & .3 & 1 \end{bmatrix}$. Notice that, to put the ML algorithms under further stress, the covariance between $X_{i,t}^{(3)}$ and $X_{i,t}^{(5)}$ is 0.7 so the two variables are highly correlated but $X_{i,t}^{(3)}$ is ten times more important (in terms of coefficient) than $X_{i,t}^{(5)}$.

of the estimated effect from the true impact and the interval coverage. Also note that, under this setup, the number of pre-intervention periods is $t_0 = 4$ when $T = 7$ and $t_0 = 9$ when $T = 12$. As estimators, we employ four different ML algorithms - LASSO, Partial Least Squares, random forest, and stochastic gradient boosting.

The results are summarized in Table 2 below and show that, across all post-intervention horizons, the MLCM achieves a very low bias. This holds true for both linear and non-linear model specifications. The bias tends to decrease when the number of pre-intervention time periods increases, as more information is present in the data. Nevertheless, the bias at the shortest time period is still very low, which reinforces our belief that the MLCM can be effectively used for short panels. For instance, under the linear specification and $t_0 = 4$, the relative bias increases from 0.2% measured at $t_0 + 1$ to 0.9% measured at $t_0 + 3$. However, by increasing the number of pre-intervention time points, the bias drops at 0.4% even at $t_0 + 3$. We also observe that the relative bias under a linear model specification is much lower than that under the non-linear specification, which was expected since non-linearities are typically more challenging to detect. In any case, the bias under a non-linear model specification is still low (the maximum bias is 4.9% and is measured in the most challenging scenario, i.e., at the third time horizon when the pre-intervention series is very short, $t_0 = 4$).

Table 2: Simulation results for the linear and non-linear model specifications

1st Post-intervention period ($t_0 + 1$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	74.24	0.14	0.002	0.94	4.06	0.10	0.024	0.95
9	block	93.56	0.10	0.001	0.95	4.14	0.10	0.023	0.94
2nd Post-intervention period ($t_0 + 2$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	55.68	0.22	0.004	0.92	3.05	0.10	0.031	0.90
9	block	70.17	0.15	0.002	0.91	3.10	0.09	0.031	0.92
3rd Post-intervention period ($t_0 + 3$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	37.12	0.33	0.009	0.93	2.03	0.10	0.049	0.93
9	block	46.78	0.19	0.004	0.92	2.07	0.09	0.046	0.93

Notes. These results are obtained under the following scenario: $N = 400$ units, $\phi = 0.8$, and $\sigma_u = 1$, where σ_u is the standard deviation of the random heterogeneity term u_i. In this table, "Pre-int." denotes the number of pre-intervention periods.

Similar observations can be made for the interval coverage, which was estimated based on $1,000$ block-bootstrap iterations. At the first time horizon after the intervention, the interval coverage is close or equal to the nominal 95% level for both linear and non-linear model specifications. Then, it tends to slightly decrease as we move further away from the intervention, but remains close to the nominal level for both linear and non-linear processes.

Overall, these numerical studies demonstrate that the MLCM can achieve forecast unbiasedness and that the coverage of the estimator is high even when considering short pre-

intervention windows. The simulation results also suggest that when we are interested in the estimation of causal effects k-step ahead from the treatment, having a greater number of pre-intervention periods further mitigates the bias, both under linear and non-linear data-generating processes.

In Supplemental Appendix D, we provide many additional results that show that these key findings remain largely unchanged if we consider alternative data-generating processes for both the outcome variable and the covariates, and if we reduce the number of available units. Notably, when ϕ is set to 1.2 (cf. Table D.1), rather than to 0.8 as in Table 2, i.e., when moving from a stationary to a non-stationary data-generating process with an explosive autoregressive coefficient for the lag of the outcome variable, the MLCM exhibits very similar performance.

4 Empirical application

The main empirical illustration provided below is an original study on the educational effects of the COVID-19 crisis, thus focusing on a simultaneous treatment affecting all available units. In Supplemental Appendix F, to demonstrate how to leverage the MLCM in settings where there are untreated units, but there may be violations of the no-interference assumption, we revisit the application on minimum wage with staggered treatment adoption in Callaway and Sant'Anna (2021) without using their control group. Both empirical illustrations rely on short panel datasets.

4.1 Background

In the aftermath of the COVID-19 pandemic, the inequality legacy of this unprecedented crisis has emerged as an issue of great policy relevance. The available evidence on income inequality documents a decrease driven by short-run and temporary government compensation policies, which were mostly targeted at the poorest segments of the population (Stantcheva, 2022). Concerning education, instead, recent micro-level evidence (Agostinelli et al., 2022; Carlana et al., 2023) shows that school closures had a large, persistent, and unequal effect on learning. The educational gaps caused by the school closures may, in turn, permanently affect the lifetime income possibilities of the current generation of students, with vast repercussions on future inequalities (Werner and Woessmann, 2023). Therefore, it is through the human capital channel that the inequality effects of the pandemic may eventually appear in the medium and long run. It follows that to anticipate longer-term consequences on income inequality and territorial disparities, it is necessary to gauge the magnitude of educational losses and their distribution across areas.

To our knowledge, there is no granular evidence on the geography of the education effects of the COVID-19 crisis. This is mainly due to identification challenges caused by the sudden

spread of the pandemic across the world, which resulted in the absence of a suitable untreated group. Italy is an important case study, as it ranks among the hardest-hit countries and was the first Western country to impose a strict nationwide lockdown. During the lockdown that began on March 9, 2020, schools across the entire national territory were closed and remained so until the end of the school year. This resulted in Italy ranking among the OECD countries with the highest number of weeks of school closures and distance learning (Battisti and Maggio, 2023). In this setting, we leverage the MLCM with non-staggered adoption, given that the treatment—the COVID-19 shock—simultaneously affected all units.[21]

4.2 Data and implementation

We employ LLM yearly data covering the period from 2013 to 2020. We cover all Italian LLMs except some of the smallest ones for which the education data are unavailable due to privacy protection, for a total of 579 LLMs (95% of Italian LLMs and over 99% of the Italian population). The dependent variable is the standardized math test score of fifth-grade students referred to the school year started in 2020. Following the approach by Carlana et al. (2023), the math score is standardized with respect to its pre-pandemic mean. This way, the detected treatment effects can be interpreted as deviations from the pre-COVID baseline. We focus on younger students for two reasons: i) learning disruptions earlier in life typically have longer-lasting and more severe effects, so younger children may have been more heavily impacted (Stantcheva, 2022); ii) primary schools were excluded from the heterogeneous school closures involved in the tier system of regional restrictions implemented by the Italian government since the onset of the second wave (November 2020), so the treatment is the same for all units (Assumption 1). The policy relevance of the math outcome is illustrated by the fact that the most recent report by the OECD Program for International Student Assessment (PISA), published in December 2023, reported that math scores dropped globally in the last few years, making headlines around the world.[22]

The initial pre-treatment information set includes over 150 variables (see Table E.1 in Supplemental Appendix E for a detailed description of the variables). In this set of covariates, we included the first three lags of all the predictors as covariates and three lags of the outcome variable. This implies that we collapse the original 2013–2020 dataset into a dataset covering the period 2017–2020.

The implementation process for this application is reported in Box 2, while the outcomes

[21] We refer to the treatment variable as COVID-19 "shock", which encompasses both the pandemic and the containment policies, as we aim to capture the total effect of the pandemic, i.e., its direct and indirect effects on educational outcomes. In line with previous literature, we consider the restrictive measures, including the lockdown and the school closures, as a manifestation of the pandemic, not as a separate treatment. COVID-related learning losses can originate from many pandemic-induced channels, such as school closures and distance learning, prolonged absence from school due to one or more infections, absence due to parents' concerns about their children's health, and many other transmission mechanisms. At the aggregate level, only the overall impact of the pandemic remains discernible.

[22] See here for coverage of the news from the New York Times.

of the empirical analysis are reported in Section 4.3.

Box 2: Estimating the local impact of the COVID-19 shock on education

Preliminary

0. **Data splitting.** We split the full 2017-2020 dataset into two subsets according to the treatment date: 2017-2019 to be used in the Design Stage; 2020 to be used in the Analysis Stage.

Design Stage

1. **Algorithm selection.** We select four supervised ML algorithms (the same set used for the simulation of Section 3.3): 1) LASSO; 2) Partial Least Squares; 3) stochastic gradient boosting; 4) random forest. As we are agnostic about the functional form of the underlying data-generating process, we opt for a mix of non-linear and linear models.
2. **Principled input selection.** We build an initial LLM dataset with over 150 predictors on the basis of literature insights. From this dataset, we then keep only the most important predictors selected by a preliminary random forest run on the pre-treatment data (see step below).
3. **Panel cross-validation.** For each algorithm, we tune hyperparameters via panel CV, involving iterative estimation on two different training-testing pairs of pre-COVID datasets: i) 2017 training; 2018 testing; ii) 2017-2018 training, 2019 testing.
4. **Performance assessment.** We assess average performance metrics for the four algorithms by comparing average forecasted vs. actual outcomes on the 2018–2019 held-out test data. We then compare the performance of the different MLCM versions.
5. **Diagnostic and placebo tests.** We first check the average distribution of errors with the best-performing model for the 2018-2019 testing sets and then show the map of the unit-level placebo temporal average treatment effects in the pre-COVID period.

Analysis Stage

6. **Final model selection.** On the basis of the comparative performance assessment, we pick the best-performing algorithm (random forest), with its best configuration, and re-train it on the full 2017–2019 sample using the hyperparameters cross-validated in the Design stage.
7. **Counterfactual forecasting.** We apply the final model estimated in Step 6 and forecast, for each LLM i, the counterfactual outcome $\widehat{\mathrm{Y}}_{i,t_0+1|t_0}(0)$.
8. **Estimation of treatment effects.** For each LLM i, we estimate the individual treatment effect by taking the difference between the observed post-COVID outcome Y_{i,t_0+1} and the ML-generated potential outcome $\widehat{\mathrm{Y}}_{i,t_0+1|t_0}(0)$.
9. **Treatment effect heterogeneity.** We estimate data-driven CATEs via a regression tree analysis with the individual treatment effects as the outcome and a host of potentially relevant predictors associated with the heterogeneity of the education effects.
10. **Inference.** We estimate standard errors for the ATE and CATEs via block-bootstrapping by performing $1,000$ bootstrap replications of Steps 6 to 9.

We use a data-driven approach to restrict the information set. More specifically, we follow Athey and Wager (2019) and Basu et al. (2018) and apply a pilot random forest on the pre-treatment data to pick up a subset of the most important predictors according to the importance ranking produced by the forest.[23] In order to select the precise number of relevant

[23]For this preliminary operation, we use default hyperparameter settings that typically perform well with

predictors in a data-driven manner, we include this number as an additional parameter in the subsequent panel CV routine which we employ to tune the hyperparameters of all the selected ML algorithms (LASSO, Partial Least Squares, stochastic gradient boosting, and random forest).[24]

Note that in this application, due to the ubiquitous nature of this shock, we do not invoke the third part of Assumption 3 (exogeneity of post-treatment covariates) and only include pre-treatment values of time-varying variables and time-invariant predictors in the information set. Assumption 3 is satisfied because the pandemic was completely unanticipated. Regarding Assumption 4, given the unprecedented disruptions brought about by the pandemic and based on institutional and subject matter knowledge, we deem it plausible that any divergence from the forecasted counterfactual in educational outcomes can be attributed exclusively to the COVID-19 shock, and not to other concomitant shocks or policies.[25] Finally, Assumption 5 is not needed for identification, as we only estimate one-step-ahead causal effects here.

4.3 Results

4.3.1 Design Stage

Table 3 below reports the average performance—in terms of panel CV MSE—of the four selected ML algorithms in forecasting the standardized math test score of fifth-grade students across the pre-treatment test sets (2018 and 2019).[26]

The best-performing MLCM version is the one using the random forest with twenty variables and 1/3 of them, i.e., six, as candidates at each split. Overall, the fully non-linear random forest and boosting fare better than the linear ML methods, suggesting the presence of non-linearities in the data-generating process. Moreover, the non-linear methods outperform LASSO and Partial Least Squares in all cases, and their performance is stronger in the presence of very few variables and less deteriorated by the inclusion of variables with poor predictive power (cf. Table E.2 in Supplemental Appendix E). See Table E.3 in Supplemental Appendix E for a detailed description of the 20 variables with the highest importance score attributed by the preliminary forest).

random forests (Athey and Wager, 2019). See Figure E.1 for the variable importance ranking of the preliminary forest.

[24]For all algorithms, we tune the subset of most predictive variables (10, 20 or 30) to be included in the analysis according to the preliminary forest. For LASSO, we tune the penalty parameter (the values from 0.1 to 0.9 with an increase of 0.1 in each run); for PLS, we tune the number of components (all possible values from 1 to the number of variables). For boosting, we tune the number of trees (1,000 or 2,000), the maximum depth of each tree (1 or 2), the minimum number of observations in terminal nodes (5 or 10) and the learning rate (0.001, 0.002 or 0.005); for random forest, we tune the number of variables randomly sampled as candidates at each split (1/2, 1/3 or 1/4 of the number of variables), and use a fixed number of 1,000 trees. All candidate hyperparameter values are tested on each testing set.

[25]There were no educational reforms at the primary school level that could have influenced the standardized math test scores of fifth-grade students in the period under scrutiny.

[26]A replication notebook of the main results reported in this section is publicly available here.

Table 3: Performance and placebo estimates

Panel A – Performance	
MLCM using:	**Panel CV-MSE**
LASSO	0.4169
Partial Least Squares	0.4044
Boosting	0.3909
Random Forest	0.3902
Panel B – Placebo estimates with Random Forest	
Time period	**Placebo ATE**
2018	0.0554
	(-0.1182 ; 0.2765)
2019	0.0254
	(-0.0273 ; 0.0728)
Temporal average	0.0404

Notes: The panel CV procedure has selected the 20 most predictive variables from the dataset for all methods. In addition, for boosting, it has selected 2,000 trees, 2 as the maximum depth of each tree, 5 as the minimum number of observations in terminal nodes and 0.005 for the learning rate. The other selected hyperparameters are: 0.3 as the penalty parameter (LASSO), 9 as the number of components (PLS), and 6 for the number of variables randomly sampled as candidates at each split (random forest). Lower and upper bounds of placebo ATE estimates refer to the 95% confidence intervals and are reported within brackets.

Panel CV on pre-treatment periods implicitly enables in-time placebo analysis along the lines of Bertrand et al. (2004) and Liu et al. (2022). In-time placebos are performed on the same pool of treated units where we "fake" that the treatment occurred at time $t_0 - 1$ and use only information up to $t_0 - 1$ to forecast the counterfactuals at time t_0. Since we know what the real values at time t_0 are, we can shift the treatment date artificially backward in time to assess the difference with the main estimates and evaluate the forecasting accuracy and unbiasedness of our estimator. As reported in the table, the placebo ATEs are indistinguishable from zero and statistically insignificant both in 2018 (ATE: 0.0554, with 95% confidence intervals (–0.1182; 0.2765) and 2019 (ATE: 0.0254, with 95% confidence intervals (–0.0273; 0.0728).[27]

In addition, the unit-level placebo map reported in Figure 2 depicts the temporal average (for the years 2018 and 2019) individual 'treatment' effects estimated with the best-performing ML algorithm, random forest, and shows that almost all LLMs exhibit no trace of significant differences between the forecasted and observed math scores in the years before the COVID-19 pandemic. In particular, only 2% of the LLMs report a difference lower than

[27]The counterfactual estimates exhibit substantial similarity regardless of whether a preliminary random forest or an alternative preliminary procedure is employed, such as boosting or selecting covariates based solely on the highest absolute correlation coefficients (e.g., the correlation between the counterfactual estimates and those obtained using default hyperparameter values is 0.9985). Detailed results are available upon request.

–1 standard deviation.[28]

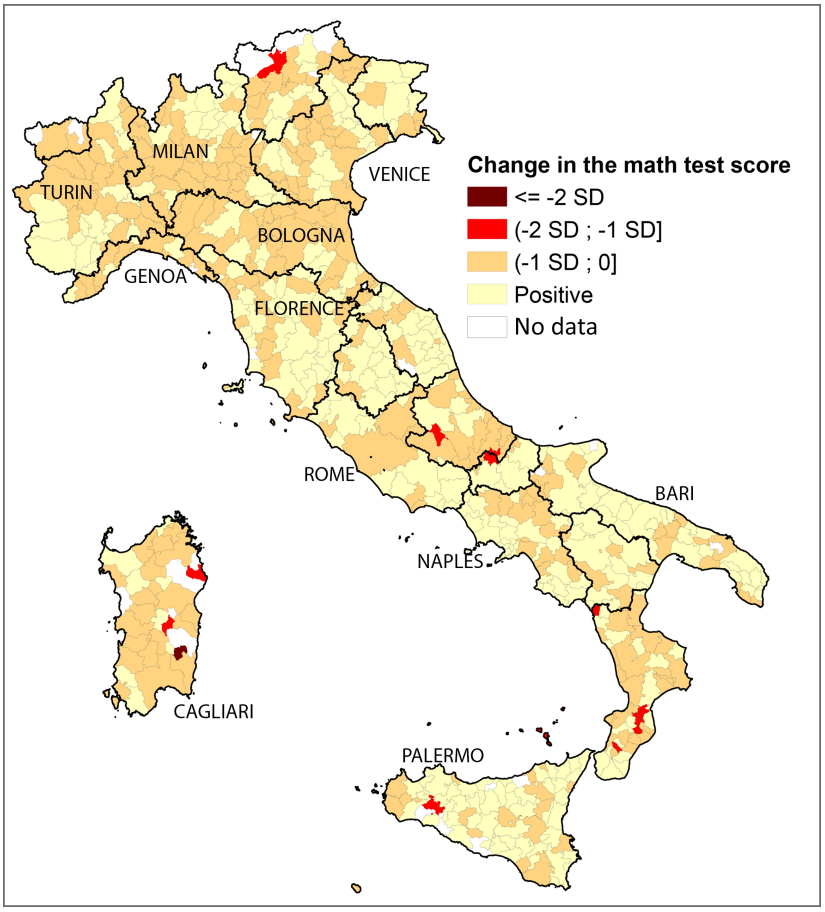


Figure 2: Unit-level panel placebo test—Temporal average 'treatment' effects. *Notes:* the standard deviation (SD) refers to the standardized pre-COVID mean. Estimation via the MLCM with random forest.

Finally, Figure 3 reports a diagnostic test for the performance of the MLCM with random forest. The figure illustrates that the distribution of the placebo forecasting errors is approximately normal and centered around zero, not only for the temporal average but also for the single pre-treatment years. This demonstrates that the ATE estimator is nearly unbiased even when applied on very short panels and without relying on the use of post-treatment information.

[28]Figure E.2 in Supplemental Appendix E presents the placebo maps for the single pre-treatment years (2018 and 2019). In these instances as well, few LLMs exhibit a difference lower than –1 standard deviation between the forecasted and observed math scores prior to the pandemic (4.3% of LLMs in 2018 and 5.7% of LLMs in 2019). Furthermore, the LLMs with extreme values tend to be the smallest ones, which are also the most volatile.

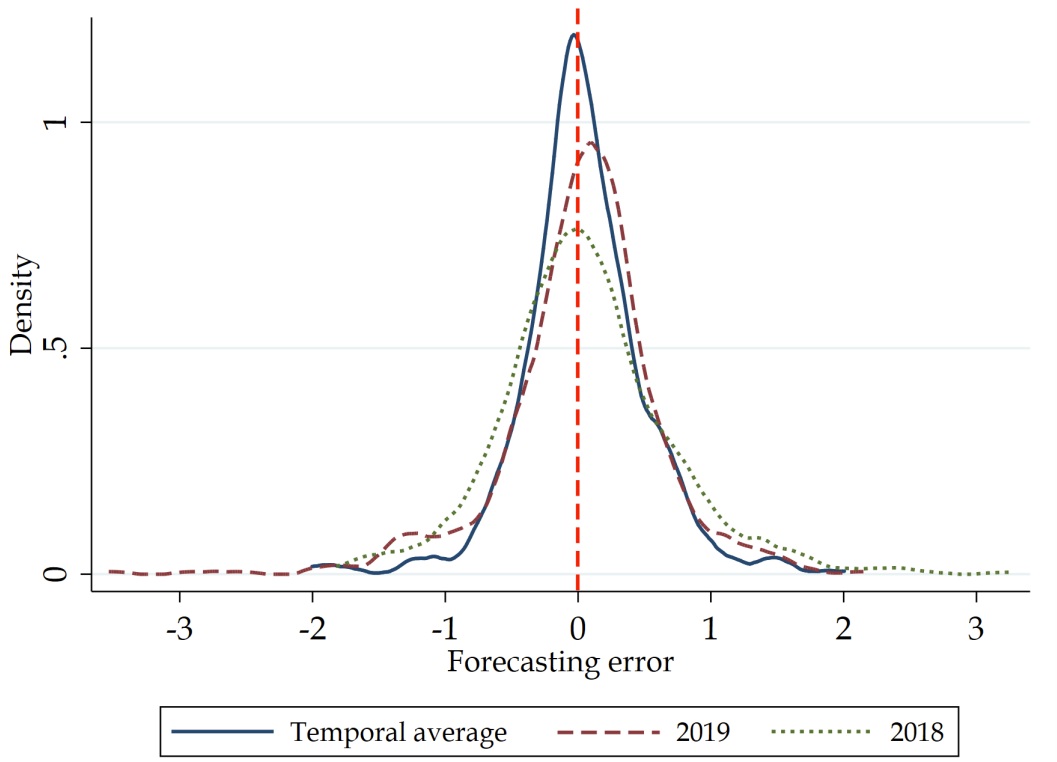


Figure 3: Distribution of the forecasting error. *Note*: Forecasting errors of the MLCM estimated with random forest.

4.3.2 Analysis Stage

Figure 4 shows the local effects of the pandemic on math scores of fifth-grade students across Italian LLMs and reports the ATE. These estimates come from the best-performing technique of the Design Stage, the MLCM with random forest.[29]

The first insight is that the COVID-19 shock engendered major learning losses in Italy. The ATE points to a decrease in the math score of –0.7608 standard deviation (SD) with respect to the pre-pandemic average, and this effect is strongly statistically significant (95% confidence intervals: –0.7976; –0.6884). This result is in line with micro-level studies documenting large negative impacts on educational outcomes in Italy (Battisti and Maggio, 2023; Carlana et al., 2023).

As shown above, some of the smallest LLMs (the most volatile) were imprecisely estimated in the placebo exercise. Since equal weight is attributed to all treated units, the presence of these LLMs might affect the accuracy of the ATE estimate. To address this potential concern, we recommend a sensitivity analysis that re-estimates the ATE following the exclusion of units with the poorest fit in the pre-treatment period. Here, we removed the top 1%, 2%, and 5% of the most imprecisely estimated LLMs in the placebo tests. The resultant ATE estimates were -0.7500, -0.7568, and -0.7341, respectively—each closely aligning with

[29]Figure E.3 in Supplemental Appendix E provides SHAP values for this model for the counterfactual 2020 forecasts.

the ATE estimate of -0.7608.

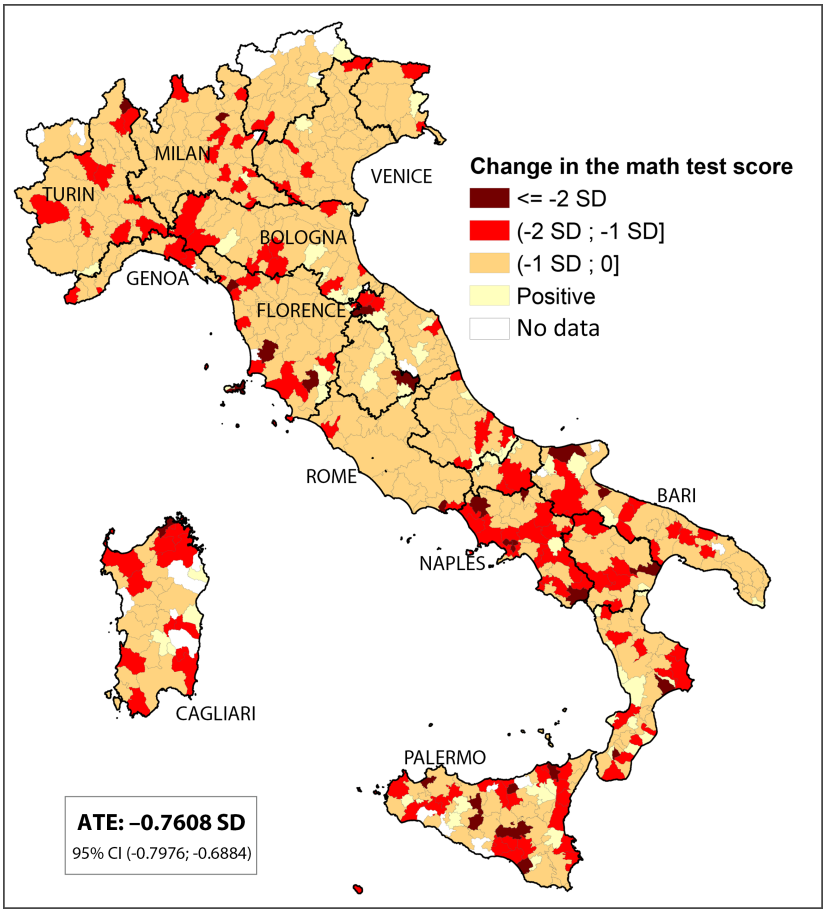


Figure 4: The local impact of the COVID-19 pandemic on education.

The second key finding is that this average impact masks considerable heterogeneity across the Italian territory. Some LLMs experienced much larger drops in students' performances. In particular, the MLCM detects learning losses of more than one standard deviation in clusters of local economies mostly located in the South of Italy (Campania, Apulia, Molise, Calabria, Sicily, and Sardinia), and drops in math scores greater than two SD in scattered areas predominantly concentrated in Central and Southern regions.

Figure 5 presents data-driven CATEs estimated with the regression tree analysis.[30] The algorithm selects only three predictors to construct the tree. The analysis reveals that the most substantial learning losses (1.22 SD below the pre-COVID baseline) occurred in LLMs characterized by an unemployment rate equal to or above 10.19%, a percentage of university graduates below 26%, and a Gini index equal to or above 0.42. Conversely, the territories with an above-average treatment effect (–0.56 SD) exhibit an unemployment rate below

[30]For this analysis, we selected a set of LLM-level variables including, among others, pre-pandemic variables such as the per capita income, the unemployment rate, the Gini index, the share of university graduates, two proxies of access to distance learning that capture the digital divide across areas, and excess deaths registered during the first wave of the pandemic (Cerqua et al., 2021). The full list can be found in Table E.4 in Supplemental Appendix E.

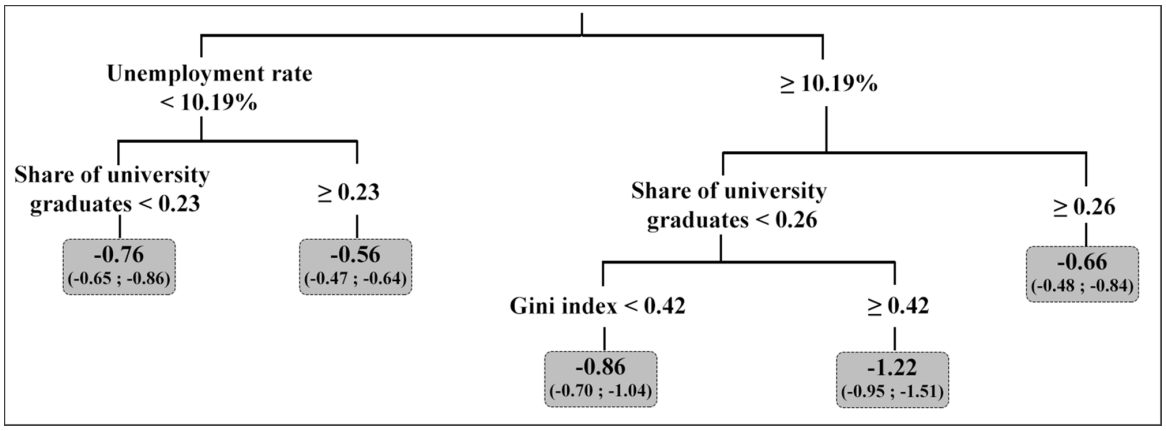


Figure 5: Data-driven CATEs. *Notes*: The values within the terminal nodes report the CATEs (measured in SD) for all LLMs grouped within that node. Lower and upper bounds refer to the 95% confidence interval and are reported in parentheses.

10.19% and a percentage of university graduates equal to or above 23%. Finally, all the CATEs reported in the tree are statistically significant.

Since areas with higher unemployment, greater inequality, and lower education rates at baseline have experienced disproportionate impacts, the pre-existing gaps across the country have been amplified. If such gaps persist, it is likely that they will act as a catalyst for future economic inequality, leading to widened territorial disparities in the medium and long term. In summary, we can anticipate that, in the absence of counterbalancing policy efforts, the pandemic will exacerbate long-standing inequalities that predate COVID-19.

5 Conclusions

Identifying causal effects is challenging without a credible control group. We overcome this challenge by proposing a new method based on counterfactual forecasting with machine learning. The MLCM can employ any off-the-shelf ML algorithm to estimate policy-relevant causal parameters in a wide variety of panel settings, including very short panels, without relying on untreated units. To illustrate its potential, we presented numerical studies, a replication of the minimum wage application in Callaway and Sant'Anna (2021) without using untreated units, and an empirical analysis on the effects of the COVID-19 crisis on education in Italy. The companion R package `MachineControl` provides an easy-to-use implementation of the proposed approach. Large shocks, international economic policies, nationwide policy changes, and regional programs engendering substantial interaction between units, are all real-world scenarios where a control group may not exist. In such cases, researchers can harness the MLCM, which complements the existing econometric toolbox for causal inference and policy evaluation.

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature 59*(2), 391–425. 17, 22

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association 105*(490), 493–505. 2, 7

Abrell, J., M. Kosch, and S. Rausch (2022). How effective is carbon pricing?—a machine learning approach to policy evaluation. *Journal of Environmental Economics and Management 112*, 102589. 4

Agostinelli, F., M. Doepke, G. Sorrenti, and F. Zilibotti (2022). When the great equalizer shuts down: Schools, peers, and parents in pandemic times. *Journal of Public Economics 206*, 104574. 26

Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. 2

Arellano, M. and S. Bonhomme (2011). Nonlinear panel data analysis. *Annual Review of Economics 3*(1), 395–424. 3

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021). Synthetic difference-in-differences. *American Economic Review 111*(12), 4088–4118. 3

Arkhangelsky, D. and G. Imbens (2024). Causal models for longitudinal and panel data: A survey. *The Econometrics Journal*, utae014. 3, 5, 17

Ashenfelter, O. and D. Card (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics 67*(4), 648–660. 2

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association 116*(536), 1716–1730. 2, 4

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360. 23

Athey, S. and S. Wager (2019). Estimating treatment effects with causal forests: An application. *Observational Studies 5*(2), 37–51. 28, 29

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279. 2

Baltagi, B. H. (2008). *Econometric analysis of panel data*, Volume 4. Springer. 3, 8, 10

Basu, S., K. Kumbier, J. B. Brown, and B. Yu (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences 115*(8), 1943–1948. 28

Battisti, M. and G. Maggio (2023). Will the last be the first? School closures and educational outcomes. *European Economic Review 154*, 104405. 27, 32

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275. 30

Bontempi, G. and S. B. Taieb (2011). Conditionally dependent strategies for multiple-step-ahead prediction in local learning. *International journal of forecasting 27*(3), 689–699. 14

Borusyak, K., X. Jaravel, and J. Spiess (2023). Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies, forthcoming*. 7

Botosaru, I., R. Giacomini, and M. Weidner (2023). Forecasted treatment effects. *arXiv preprint*. Available at https://arxiv.org/abs/2309.05639. 4, 5

Box, G. E. and G. C. Tiao (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association 70*(349), 70–79. 3

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science 16*(3), 199–231. 16

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, and S. L. Scott (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics 9*(1), 247. 3, 8

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230. 4, 26, 34, 60, 61, 62, 63

Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review 43*(2), 245–257. 3

Card, D. and A. B. Krueger (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review 84*(4), 772–793. 2

Carlana, M., E. La Ferrara, and C. Lopez (2023). Exacerbated inequalities: The learning loss from covid-19 in italy. *AEA Papers and Proceedings 113*, 489–93. 26, 27, 32, 52

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics 14*(3), 1171–1179. 44

Carvalho, C., R. Masini, and M. C. Medeiros (2018). ArCo: An artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics 207*(2), 352–380. 4, 6, 8

Cerqua, A., R. Di Stefano, M. Letta, and S. Miccoli (2021). Local mortality estimates during the covid-19 pandemic in italy. *Journal of Population Economics 34*(4), 1189–1217. 33, 55

Chen, R., L. Yang, and C. Hafner (2004). Nonparametric multistep-ahead prediction in time series analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology 66*(3), 669–686. 9

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21(1)*, C1–C68. 7, 18

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. *National Bureau of Economic Research Working Paper w24678*. 13

Chernozhukov, V., C. Hansen, N. Kallus, M. Spindler, and V. Syrgkanis (2024). Applied causal inference powered by ML and AI. *arXiv preprint arXiv:2403.02467*. 7

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association 116*(536), 1849–1864. 3

Chevillon, G. (2007). Direct multi-step estimation and forecasting. *Journal of Economic Surveys 21*(4), 746–785. 14

Chiu, A., X. Lan, Z. Liu, and Y. Xu (2023). What to do (and not to do) with causal panel analysis under parallel trends: Lessons from a large reanalysis study. arXiv preprint. Available at `arXiv:2309.15983`. 7

Cox, D. R. (1958). *Planning of experiments*. New York, NY: Wiley. 2, 6

Duflo, E. (2017). The economist as plumber. *American Economic Review 107*(5), 1–26. 2

D'Haultfœuille, X., S. Hoderlein, and Y. Sasaki (2023). Nonparametric difference-in-differences in repeated cross-sections with continuous treatments. *Journal of Econometrics 234*(2), 664–690. 10

Eggers, A. C., G. Tuñón, and A. Dafoe (2024). Placebo tests for causal inference. *American Journal of Political Science 68*(3), 1106–1121. 20

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics 40*(1), 313–327. 13

Hoderlein, S. and H. White (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics 168*(2), 300–314. 10

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960. 3

Hyndman, R. J. and G. Athanasopoulos (2021). *Forecasting: principles and practice*. Melbourne, Australia: OTexts. 3, 20

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge University Press. 3, 6, 13

Jarvis, S., O. Deschenes, and A. Jha (2022). The private and external costs of germany's nuclear phase-out. *Journal of the European Economic Association 20*(3), 1311–1346. 4

Johannemann, J., V. Hadad, S. Athey, and S. Wager (2019). Sufficient representations for categorical variables. arXiv preprint. Available at `arXiv:1908.09874`. 9, 60

Knaus, M. C., M. Lechner, and A. Strittmatter (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal 24*(1), 134–161. 13

Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling*. New York, NY: Springer. 18

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics 17*(3), 1217–1241. 44

Liu, L., H. R. Moon, and F. Schorfheide (2020). Forecasting with dynamic panel data models. *Econometrica 88*(1), 171–201. 3, 20

Liu, L., Y. Wang, and Y. Xu (2022). A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science*. 20, 30

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems 30*. 22

Masini, R. and M. C. Medeiros (2021). Counterfactual analysis with artificial controls: Inference, high dimensions, and nonstationarity. *Journal of the American Statistical Association 116*(536), 1773–1788. 4, 6, 8

Menchetti, F., F. Cipollini, and F. Mealli (2023). Combining counterfactual outcomes and arima models for policy evaluation. *The Econometrics Journal 26*(1), 1–24. 3

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu. com. 22

Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106. 17

Ogburn, E. L. and T. J. VanderWeele (2014). Causal diagrams for interference. *Statistical Science 29*(4), 559–578. 7

Pesaran, M. H., A. Pick, and A. Timmermann (2011). Variable selection, estimation and inference for multi-period forecasting problems. *Journal of Econometrics 164*(1), 173–187. 14

Petropoulos, F., D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. Ben Taieb, and C. B. et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting 38*(3), 705–871. 14, 15

Prest, B. C., C. J. Wichman, and K. Palmer (2023). Rcts against the machine: Can machine learning prediction methods recover experimental treatment effects? *Journal of the Association of Environmental and Resource Economists 10*(5), 1231–1264. 4

Rambachan, A. and J. Roth (2023). A more credible approach to parallel trends. *Review of Economic Studies 90*, 2555–2591. 13

Rambachan, A. and N. Shephard (2021). When do common time series estimands have nonparametric causal meaning? *URL: https://scholar. harvard. edu/shephard/publications/nonparametric-dynamic-causal-modelmacroeconometrics*. 3

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics 235*, 2218–2244. 3

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701. 2, 6

Semenova, V., M. Goldman, V. Chernozhukov, and M. Taddy (2023). Estimation and inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics 14*, 471–510. 4

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association 101*(476), 1398–1407. 2, 6

Stantcheva, S. (2022). Inequalities in the times of a pandemic. *Economic Policy 37*(109), 5–41. 26, 27

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences 113*(27), 7310–7315. 4

Viviano, D. and J. Bradic (2023). Synthetic learner: model-free inference on treatments over time. *Journal of Econometrics 234*(2), 691–713. 4, 6, 44

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242. 6, 7, 23

Werner, K. and L. Woessmann (2023). The legacy of COVID-19 in education. *Economic Policy*, eiad016. 26

Xu, Y. (2023). Causal inference with time-series cross-sectional data: a reflection. *The Oxford Handbook for Methodological Pluralism*. Forthcoming. 7

Supplemental Appendix

A Causal estimands in settings with interference between units

Compared to Case (i), in the scenario with violations of the no-interference assumption, the causal estimands are different. In this section, we define the main causal estimands for the setup in which only a subgroup of units is treated, but there are spillover and general equilibrium effects that prevent using the set of untreated units to build a valid control group (Case (ii) defined above). We can first define the ATT across the treated units at a given point in time.

Definition 9. *For any strictly positive integer k, let $\tau_{i,t_0+k} = \mathrm{Y}_{i,t_0+k}(1) - \mathrm{Y}_{i,t_0+k|t_0}(0)$ be the unit-level treatment effect of the policy at time $t_0 + k$ and denote with N_T the number of treated units. The ATT at time $t_0 + k$ is defined as,*

$$\tau^T_{t_0+k} = \frac{1}{N_T} \sum_{i=1}^{N_T} \tau_{i,t_0+k} = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathrm{Y}_{i,t_0+k}(1) - \mathrm{Y}_{i,t_0+k|t_0}(0).$$

Note that, in this framework, the ATE does not coincide with the ATT because not all units are directly exposed to the treatment.

As a subgroup of the units is not treated but potentially subject to spillovers, we can define the Average Spillover effect on the Affected (ASA) across the untreated units at a given point in time.

Definition 10. *For any strictly positive integer k, let $\tau_{j,t_0+k} = \mathrm{Y}_{j,t_0+k}(1) - \mathrm{Y}_{j,t_0+k|t_0}(0)$ denote the unit-level spillover effect of the policy at time $t_0 + k$ for all untreated units and denote with N_C the number of untreated units. The ASA at time $t_0 + k$ is defined as,*

$$\tau^C_{t_0+k} = \frac{1}{N_C} \sum_{i=1}^{N_C} \tau_{j,t_0+k} = \frac{1}{N_C} \sum_{i=1}^{N_C} \mathrm{Y}_{j,t_0+k}(1) - \mathrm{Y}_{j,t_0+k|t_0}(0).$$

Both causal estimands, $\tau^T_{t_0+k}$ and $\tau^C_{t_0+k}$, can be estimated with the MLCM by applying the same implementation routine defined in the paper separately for the group of treated and untreated units.

B Identification proof

We show here the general identification proof for the causal estimands defined in Section 2.2.

Proof. As in the main text, the proof is organized by distinguishing between $k = 1$ and $k = 2$.

- $k = 1$. Let $\mathrm{Y}_{i,t}(0)$ follow Equation (2). In the expression $\tau_{i,t_0+1} = \mathrm{Y}_{i,t_0+1}(1) - \mathrm{Y}_{i,t_0+1|t_0}(0)$, the first term $\mathrm{Y}_{i,t_0+1}(1)$ is the observed outcome under the policy and thus is immediately identified. Since $Q = \max\{0, q\} = q$, the second term can be written as

$$\begin{aligned}
\mathrm{Y}_{i,t_0+1|t_0}(0) &= \mathrm{E}[\mathrm{Y}_{i,t_0+1}(0) |\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(q)}_{i,t_0+1}] && \text{Def. of } \mathrm{Y}_{i,t_0+1|t_0}(0) \\
&= \mathrm{E}[h(\mathrm{Y}^{(p)}_{i,t_0}(0), \mathbf{X}^{(q)}_{i,t_0+1}) |\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}^{(q)}_{i,t_0+1}] && \text{A.4 - Eq.(2)} \\
&= h(y^{(p)}_{i,t_0}(0), \mathbf{x}^{(q)}_{i,t_0+1}) = y_{i,t_0+1|t_0}(0).
\end{aligned}$$

  The last expression follows from the fact that $h(\cdot)$ is deterministic once we observe $\mathrm{Y}^{(p)}_{i,t_0}(0) = y^{(p)}_{i,t_0}(0)$ and $\mathbf{X}^{(q)}_{i,t_0+1} = \mathbf{x}^{(q)}_{i,t_0+1}$. As a result, $\mathrm{Y}_{i,t_0+1|t_0}(0)$ is identified from observed data, and so is τ_{i,t_0+1}. Notice that in this proof, we never used the linearity of the potential outcome model. Thus, the proof at $k = 1$ is the same even when $h(\cdot)$ is non-linear.

- $k = 2$. In the expression $\tau_{i,t_0+2} = \mathrm{Y}_{i,t_0+2}(1) - \mathrm{Y}_{i,t_0+2|t_0}(0)$, the first term $\mathrm{Y}_{i,t_0+2}(1)$ is the observed outcome under the policy. Thus, the identification of the causal effect depends solely on the second term, $\mathrm{Y}_{i,t_0+2|t_0}(0)$. Since $Q = \max\{1, q\}$, we now have two possible cases: either $q \geq 1$, simplifying to $Q = q$ or $q = 0$, in which case $Q = 1 > q$. Assuming the latter, we now distinguish between a linear and a non-linear model specification:

- Linear case. Assume the following linear model $h(\mathrm{Y}^{(p)}_{i,t-1}(0), \mathbf{X}^{(q)}_t) = \sum_{j=0}^{p} b_1^{(j)} \mathrm{Y}_{i,t-1-j}(0) + \sum_{j=0}^{q} \mathbf{b_2} \mathbf{X}_{i,t-q}$. The term $\mathrm{Y}_{i,t_0+2|t_0}(0)$ can then be written as,

$$\begin{aligned}
\mathrm{Y}_{i,t_0+2|t_0}(0) &= \mathrm{E}[\mathrm{Y}_{i,t_0+2}(0) |\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] && \text{Def. } \mathrm{Y}_{i,t_0+2|t_0}(0) \\
&= \mathrm{E}[h(\mathrm{Y}^{(p)}_{i,t_0+1}(0), \mathbf{X}_{i,t_0+2}) |\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] && \text{A.4 - Eq.(2)} \\
&= \mathrm{E}[b_1^{(0)} \mathrm{Y}_{i,t_0+1}(0) + \sum_{j=1}^{p} b_1^{(j)} \mathrm{Y}_{i,t_0+1-j} + \mathbf{b_2} \mathbf{X}_{i,t_0+2} |\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] && \text{Linearity} \\
&= b_1^{(0)} \mathrm{E}[\mathrm{Y}_{i,t_0+1}(0) |\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}_{i,t_0+1}] + \sum_{j=1}^{p} b_1^{(j)} y_{i,t_0+1-j} + \mathbf{b_2} \mathbf{x}_{i,t_0+2} && \text{A.3} \\
&= b_1^{(0)} y_{i,t_0+1|t_0}(0) + \sum_{j=1}^{p} b_1^{(j)} y_{i,t_0+1-j} + \mathbf{b_2} \mathbf{x}_{i,t_0+2}.
\end{aligned}$$

- Non-linear case. The identification of $\mathrm{Y}_{i,t_0+2|t_0}(0)$ can be based on the formulation for

the post-intervention potential outcome model defined by Equation (3),

$$
\begin{aligned}
\mathrm{Y}_{i,t_0+2|t_0}(0) &= \mathrm{E}[\mathrm{Y}_{i,t_0+2}(0)|\, \mathrm{Y}_{i,t_0}(0), \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] && \text{Def. } \mathrm{Y}_{i,t_0+2|t_0}(0) \\
&= \mathrm{E}[g_2(\mathrm{Y}_{i,t_0+1|t_0}(0), \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}_{i,t_0+2})|\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] && \text{A.5 - Eq.(3)} \\
&= \mathrm{E}[g_2(h(\mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}_{i,t_0+1}), \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}_{i,t_0+2})|\, \mathrm{Y}^{(p)}_{i,t_0}, \mathbf{X}_{i,t_0+2}, \mathbf{X}_{i,t_0+1}] \\
&= g_2(y_{i,t_0+1|t_0}(0), y^{(p)}_{i,t_0}, \mathbf{x}_{i,t_0+2}).
\end{aligned}
$$

The last expression follows from the fact that $g_2(\cdot)$ is deterministic once we observe $\mathrm{Y}_{i,t_0} = y_{i,t_0}$, $\mathbf{X}_{i,t_0+1} = \mathbf{x}_{i,t_0+1}$ and $\mathbf{X}_{i,t_0+2} = \mathbf{x}_{i,t_0+2}$. In addition, even though Assumption 3 was not explicitly mentioned in this part of the proof, notice that it is implicit in Assumption 5, as by Equation (3), the potential outcomes never depend on future covariates.

The proof for a generic time $t_0 + k$ follows analogously by induction. □

C Bootstrap inference for ATE and CATE confidence intervals

Following recent work on causal machine learning estimators for panel data (Viviano and Bradic, 2023), inference on causal effects estimated with the MLCM is performed by block-bootstrap. Since we are dealing with a panel dataset that has an inherent temporal structure, resampling the unit-time pairs independently in a context with dependent data would lead to data leakage. Therefore, we sample the units instead of the unit-time pairs so that if a unit is sampled, its entire evolution over time is sampled as well. This way, the inference is more robust to misspecification issues. This idea is in the spirit of the original block-bootstrap introduced by Carlstein (1986) and Kunsch (1989) that consists in dividing a time series in blocks (overlapping or non-overlapping) and then sampling the blocks with replacement. Since we have a panel dataset, we can consider each unit to form one block. The full algorithm to derive block-bootstrap confidence intervals in our setting is detailed in Algorithm 3 below.

Algorithm 3 Block-bootstrap algorithm for the ATE

Require: N, B,

1: **for** b in 1 : B where B is the total number of bootstrap iterations **do**
2: Sample N units with replacement
3: Only for the resampled units, retain the tuple $(\mathrm{Y}_{i,t}^{(p)}(0), \mathbf{X}_{i,t}^{(q)})^{(b)}$ for all $t = 1, \ldots, T$ (pre- and post-intervention)
4: Compute the bootstrap replication of the ATE, i.e., $\hat{\tau}_t^{(b)}$ by repeating the Design stage (steps 3 to 5) and the Analysis stage (steps 6 to 9) of the MLCM outlined in Section 3
5: **end for**
6: Let $\widehat{G}(c) = \#\{\hat{\tau}_t^{(b)} \leq c\}/B$ be the empirical cumulative distribution function of the B bootstrap replications. Compute a $(1-\alpha)$ percentile interval as $[\widehat{G}^{-1}(\alpha/2), \widehat{G}^{-1}(1-\alpha/2)]$.

Deriving confidence intervals for the CATEs is more challenging than for the ATE. The reason is that CATEs are estimated with a regression tree where the outcome is a collection of individual causal effects. Therefore, if we resample the units at the very beginning and then we estimate B different CATEs, we will end up with very different trees: both the splits and the observations within each terminal node will likely differ. For this reason, we propose to resample directly the observations within each terminal node of the tree. Note that the observations are the estimated individual causal effects (computed by comparing the observed data with the ML forecasts). Our estimand of interest is the group-average of the individual effects, so at each bootstrap iteration, we resample the units in each terminal node, averaging the individual effects and obtaining a bootstrap distribution for the CATE. The full algorithm to derive confidence intervals is described in Algorithm 4 below.

Algorithm 4 Block-bootstrap algorithm for CATEs

Require: $\hat{\tau}_t$, $\mathbf{X}_{i,t}$, D, B,

1: **for** b in $1:B$ where B is the total number of bootstrap iterations and for $d = 1, \ldots, D$ where D is the total number of terminal nodes of the tree **do**
2: Sample with replacement the N_d units in terminal node d
3: Only for the resampled units, retain the tuple $(\hat{\tau}_{i,t}, \mathbf{X}_{i,t})_d^{(b)}$
4: Compute the bootstrap replication of CATE in the terminal node d by averaging the ATEs of the resampled units, i.e., $\hat{\tau}_t(d)$.
5: By definition (13), $\hat{\tau}_t(d)$ is an estimator of CATE, since we are averaging the individual effects among the units in group (terminal node) d, i.e., we are conditioning on covariates
6: **end for**
7: Let $\widehat{G}(c) = \#\{\hat{\tau}_t(d) \leq c\}/B$ be the empirical cumulative distribution function of the d-group CATE in the B bootstrap replication. Compute a $(1-\alpha)$ percentile interval as $[\widehat{G}^{-1}(\alpha/2), \widehat{G}^{-1}(1-\alpha/2)]$.

D Additional simulation results

In this Section of the Supplemental Appendix, we report the results of all the simulation scenarios that we generated by combining the following parameters: i) number of units in the panel ($N = 400$, $N = 200$); ii) autoregressive coefficient ($\phi = 0.8$, $\phi = 1.2$); iii) standard deviation of u_i regulating the amount of heterogeneity between units ($\sigma_u = 1$, $\sigma_u = 0.1$).

Table D.1 shows the results of an "explosive" simulation scenario corresponding to the one reported in the main text ($N = 400$, $\sigma_u = 1$), but with $\phi = 1.2$. Tables D.2 and D.3 report the results for the remaining combinations with $N = 400$ and $\sigma_u = 0.1$.

Tables D.4–D.7 show the results of the simulations when the number of units in the panel reduces to $N = 200$.

Table D.1: Simulation results for $N = 400$, $\phi = 1.2$ and $\sigma_u = 1$

1st Post-intervention period ($t_0 + 1$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	234.28	0.17	0.001	0.94	4.07	0.10	0.024	0.95
9	block	641.62	0.12	0.000	0.94	4.14	0.09	0.023	0.95
2nd Post-intervention period ($t_0 + 2$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	175.71	0.39	0.002	0.93	3.05	0.10	0.032	0.91
9	block	481.21	0.25	0.001	0.92	3.10	0.10	0.031	0.92
3rd Post-intervention period ($t_0 + 3$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	117.14	0.75	0.006	0.93	2.03	0.10	0.050	0.92
9	block	320.81	0.46	0.001	0.94	2.07	0.09	0.045	0.92

Table D.2: Simulation results for $N = 400$, $\phi = 0.8$ and $\sigma_u = 0.1$

1st Post-intervention period ($t_0 + 1$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	73.38	0.27	0.004	0.97	4.07	0.14	0.033	0.97
9	block	92.58	0.15	0.002	0.95	4.14	0.11	0.027	0.95
2nd Post-intervention period ($t_0 + 2$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	55.03	0.37	0.007	0.94	3.05	0.10	0.034	0.91
9	block	69.44	0.21	0.003	0.93	3.10	0.11	0.034	0.93
3rd Post-intervention period ($t_0 + 3$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	36.69	0.72	0.020	0.94	2.04	0.14	0.071	0.95
9	block	46.29	0.33	0.007	0.94	2.07	0.11	0.051	0.93

Table D.3: Simulation results for $N = 400$, $\phi = 1.2$ and $\sigma_u = 0.1$

1st Post-intervention period ($t_0 + 1$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	233.97	0.31	0.001	0.96	4.07	0.14	0.034	0.97
9	block	640.02	0.18	0.000	0.95	4.14	0.11	0.027	0.95
2nd Post-intervention period ($t_0 + 2$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	175.47	0.65	0.004	0.95	3.05	0.10	0.034	0.92
9	block	480.01	0.36	0.001	0.94	3.11	0.11	0.034	0.94
3rd Post-intervention period ($t_0 + 3$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	116.98	1.34	0.011	0.95	2.03	0.15	0.073	0.95
9	block	320.01	0.72	0.002	0.95	2.07	0.10	0.051	0.94

Table D.4: Simulation results for $N = 200$ units, $\phi = 0.8$ and $\sigma_u = 1$

1st Post-intervention period ($t_0 + 1$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	74.23	0.22	0.003	0.93	4.07	0.13	0.033	0.97
9	block	92.07	0.15	0.002	0.94	4.14	0.13	0.032	0.94
2nd Post-intervention period ($t_0 + 2$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	55.67	0.36	0.006	0.90	3.05	0.13	0.044	0.92
9	block	69.06	0.22	0.003	0.90	3.11	0.13	0.042	0.93
3rd Post-intervention period ($t_0 + 3$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	37.11	0.48	0.013	0.93	2.04	0.14	0.068	0.93
9	block	46.04	0.29	0.006	0.92	2.07	0.13	0.064	0.92

Table D.5: Simulation results for $N = 200$ units, $\phi = 0.8$ and $\sigma_u = 0.1$

1st Post-intervention period ($t_0 + 1$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	73.34	0.41	0.006	0.96	4.07	0.19	0.046	0.96
9	block	91.86	0.21	0.002	0.94	4.14	0.15	0.037	0.95
2nd Post-intervention period ($t_0 + 2$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	55.01	0.56	0.010	0.93	3.05	0.15	0.048	0.91
9	block	68.90	0.32	0.005	0.92	3.11	0.15	0.048	0.94
3rd Post-intervention period ($t_0 + 3$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	36.67	1.07	0.029	0.92	2.03	0.22	0.110	0.95
9	block	45.93	0.50	0.011	0.94	2.07	0.15	0.074	0.93

Table D.6: Simulation results for $N = 200$ units, $\phi = 1.2$ and $\sigma_u = 1$

1st Post-intervention period ($t_0 + 1$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	232.58	0.26	0.001	0.94	4.07	0.14	0.033	0.96
9	block	629.30	0.18	0.000	0.94	4.14	0.14	0.033	0.94
2nd Post-intervention period ($t_0 + 2$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	174.44	0.62	0.004	0.93	3.05	0.14	0.045	0.93
9	block	471.97	0.37	0.001	0.92	3.11	0.13	0.042	0.93
3rd Post-intervention period ($t_0 + 3$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	116.29	1.18	0.010	0.93	2.03	0.14	0.069	0.93
9	block	314.65	0.67	0.002	0.93	2.07	0.13	0.064	0.91

Table D.7: Simulation results for $N = 200$ units, $\phi = 1.2$ and $\sigma_u = 0.1$

1st Post-intervention period ($t_0 + 1$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	74.24	0.14	0.002	0.94	4.07	0.19	0.046	0.96
9	block	632.26	0.27	0.000	0.94	4.14	0.15	0.037	0.96
2nd Post-intervention period ($t_0 + 2$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	55.68	0.22	0.004	0.92	3.050	0.14	0.047	0.91
9	block	474.20	0.55	0.001	0.93	3.100	0.14	0.046	0.95
3rd Post-intervention period ($t_0 + 3$)									
		Linear				Non-Linear			
Pre-int.	Bootstrap	True ATE	Bias	Rel. Bias	Coverage	True ATE	Bias	Rel. Bias	Coverage
4	block	37.12	0.33	0.009	0.93	2.03	0.22	0.107	0.95
9	block	316.13	1.07	0.003	0.94	2.07	0.15	0.074	0.94

E Additional application material

Table E.1:
Definition of the variables included in the initial dataset

Dependent variable			
Variable name	**Definition**	**Time period**	**Source**
Standardized math test score	Standardized math test score of fifth-grade students.	School year 2012/2013-school year 2020/2021	National Institute for the Evaluation of Educational Instruction and Training (INVALSI)
Time-invariant variables			
Variable name	**Definition**	**Time period**	**Source**
Economic classification dummies	Without specialization, non-manufacturing (touristic), non-manufacturing (non-touristic), made in Italy, other manufacturing	2011	Italian National Institute of Statistics (Istat)
Dummy district	LLM with at least one industrial district	2011	Istat
Regional dummies	A dummy for each of the 20 Italian regions		Istat
Population dummies	$\leq$ 10,000; (10,000; 50,000]; (50,000; 100,000]; (100,000; 500,000]; >500,000	2011	Istat
Share of individuals with a university degree	Share of individuals aged 30-34 with a university degree	2014	Istat
Share of individuals with a high-school degree	Share of individuals aged 30-34 with a high-school degree	2014	Istat
Share of urban area	Urban surface / Total surface	2012	Istat
Share of population in the periphery	Share of population located in municipalities considered as peripheral or ultra-peripheral according to the SNAI classification	2011	Istat
Number of public libraries	Number of public libraries	2016	Istat
Number of visitors to museums	Number of visitors to museums, galleries, archaeological sites, and monuments per 100,000 inhabitants	2015	Istat
Average maximum temperature	30-year average (1980-2010) of maximum air temperature (°C)	1980-2010	CAMS European Air Quality Reanalysis Data
Average minimum temperature	30-year average (1980-2010) of minimum air temperature (°C)	1980-2010	CAMS European Air Quality Reanalysis Data

Average total yearly precipitation	30-year average (1980-2010) of total precipitation (mm)	1980-2010	CAMS European Air Quality Reanalysis Data
Turnout referendum	Turnout at the constitutional referendum that was held in Italy on 4 December 2016	2016	Ministry of the Interior
Share of Yes at the referendum	Share of Yes at the constitutional referendum that was held in Italy on 4 December 2016	2016	Ministry of the Interior
Percentage of children aged 0-2 years who have utilized childcare services	Children aged 0-2 years who have utilized childcare services provided by Municipalities (nurseries, micro-nurseries, or integrated and innovative services) / Resident children aged 0-2 years * 100.	2016	Istat

Time-varying variables

Variable name	**Definition**	**Time period**	**Source**
Standardized Italian test score	Standardized Italian test score of fifth-grade students.	School year 2012/2013-school year 2018/2019	INVALSI
Percentage of participants to the math test	Percentage of participation in the math test (pupils who took the test out of expected pupils)	School year 2012/2013-school year 2018/2019	INVALSI
Percentage of participants to the Italian test	Percentage of participation in the Italian test (pupils who took the test out of expected pupils)	School year 2012/2013-school year 2018/2019	INVALSI
Average percentage score for math	Average percentage score for math - adjusted for cheating	School year 2012/2013-school year 2018/2019	INVALSI
Average percentage score for Italian	Average percentage score for Italian - adjusted for cheating	School year 2012/2013-school year 2018/2019	INVALSI
Standard deviation for math	Standard deviation of the percentage score for math	School year 2012/2013-school year 2018/2019	INVALSI
Standard deviation for Italian	Standard deviation of the percentage score for Italian	School year 2012/2013-school year 2018/2019	INVALSI
Log of the Gini index	Log of the Gini index	2013-2019	Ministry of Economy and Finance (MEF)

Share of foreign population	Foreigners / population	2013-2019	Istat
Unemployment rate	Resident population aged 15+ not in employment but currently available for work / Labor force	2013-2019	Istat
Share of graduate mayors	Share of municipalities with a mayor with a university degree	2013-2019	Ministry of the Interior
Share of recycled waste	Share of recycled waste	2013-2019	Italian National Institute for Environmental Protection and Research (Ispra)
Share of workers in manufacturing	Share of workers in manufacturing	2013-2019	Istat
Share of old population	Share of population aged >=65	2013-2019	Istat
Declared income per capita	Declared income per capita	2013-2019	MEF
Log of declared income (total)	Log of declared income (total)	2013-2019	MEF
Share of income for pensions	Share of overall declared income for pensions	2013-2019	MEF
Share of income for employment	Share of overall declared income for employment	2013-2019	MEF
Population	Resident population	2013-2019	Istat
Average price per square meter	Average price per square meter - house	2013-2019	Osservatorio del Mercato Immobiliare – Agenzia delle Entrate

Notes. In our application, time-invariant covariates are either features of the LLM not subject to time changes (e.g. area of the LLM) or features taken at a time before 2017. Data from the 2020/2021 school year represent the first available post-pandemic data point for the standardized math test score, as there was no test for the school year 2019/2020 due to the national lockdown. Following Carlana et al. (2023), math test scores have been standardized for the entire dataset using the 2013-2019 values as a benchmark, with a mean of 0 and a standard deviation of 1.

Table E.2: Performances with different numbers of variables

Method	**MSE**		
	10 variables	**20 variables**	**30 variables**
LASSO	0.4316	0.4169	0.6559
Partial Least Squares	0.4172	0.4044	0.4156
Boosting	0.3995	0.3909	0.4037
Random Forest	0.4011	0.3902	0.3935

Table E.3: Definition of the 20 variables included in the estimation process (random forest)

Predictors	**Lag**
Standardized math test score	First, second and third
Standardized Italian test score	First, second and third
Average percentage score for math	First, second and third
Change in the average percentage score for Italian	First
Average percentage score for Italian	Second
Change in the average percentage score for Italian	First
Change in the standard deviation for math	First and second
Change in the unemployment rate	Third
Change in the log of declared income (total)	Third
Log of declared income (total)	Third
Share of individuals with a high-school degree	Fixed (2014)
Dummy Calabria	Fixed
Turnout referendum	Fixed (2016)

Note: These are the 20 variables selected via the panel CV procedure by the preliminary random forest.

Table E.4:
Definition of the variables included in the data-driven CATEs analysis

Variable name	Definition	Time period	Source
COVID-19 impact on the standardized math test score (dependent variable)	Treatment effect of the COVID-19 crisis on the standardized math test score (SD)	School year 2020/2021	Estimated via the MLCM with random forest
Share of income accruing to low-income earners	Share of income accruing to individuals with an overall income $\leq$ €26,000	2019	MEF
Average download speed of the internet	Average download speed of the internet	2018	AGCOM
Share of households without internet access	Share of households without internet access	2018	AGCOM
Tourism LLM	Dummy variable equal to 1 for LLMs specialized in tourism	2011	Istat
Share of individuals with a university degree	Share of individuals aged 30-34 with a university degree	2015	Istat
Population	Resident population	2019	Istat
Per capita expenditure on gambling	Per capita expenditure on gambling	2019	
Unemployment rate	Resident population aged 15+ not in employment but currently available for work	2019	Istat
Excess mortality estimates	Municipality-level excess mortality estimated by applying ML techniques to all-cause deaths data, aggregated at the LLM level	From Feb 21, 2020 to Sep 30, 2020	Cerqua et al. (2021)
Share of temporary contracts	Number of employees with temporary contracts in October divided by the number of employees in October	2015	Istat
Share of jobs in suspended economic activities	Share of jobs in activities suspended in March 2020 by the Italian Government due to the spread of the pandemic	2017	Istat
Per capita income	The amount of money earned per person	2019	MEF
Share of firms having employees in short-time work subsidies	The number of firms with employees receiving short-time work subsidies divided by the universe of firms registered in the Business Register	Average (2015-2018)	Ministry of Labor and Social Policies

Share of population living in peripheral areas	Share of population living in areas defined by Istat as peripheral or ultra-peripheral	Jan 1, 2020	Istat
Dependency ratio	The ratio of those typically not in the labor force (the dependent part, ages 0 to 14 and 65+) and those typically in the labor force (the productive part, ages 15 to 64)	Jan 1, 2020	Istat
Gini index	Gini index	2019	MEF
Average price per square meter	Average price per square meter - house	2019	Osservatorio del Mercato Immobiliare – Agenzia delle Entrate

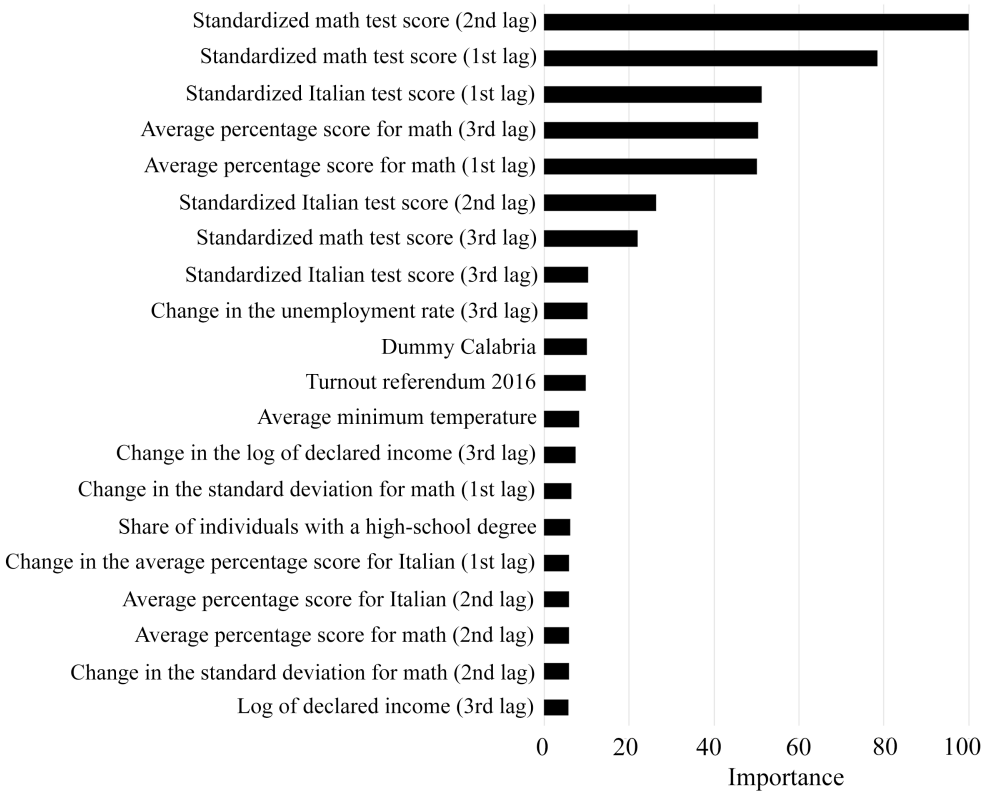


Figure E.1: Variable importance ranking – Preliminary random forest.

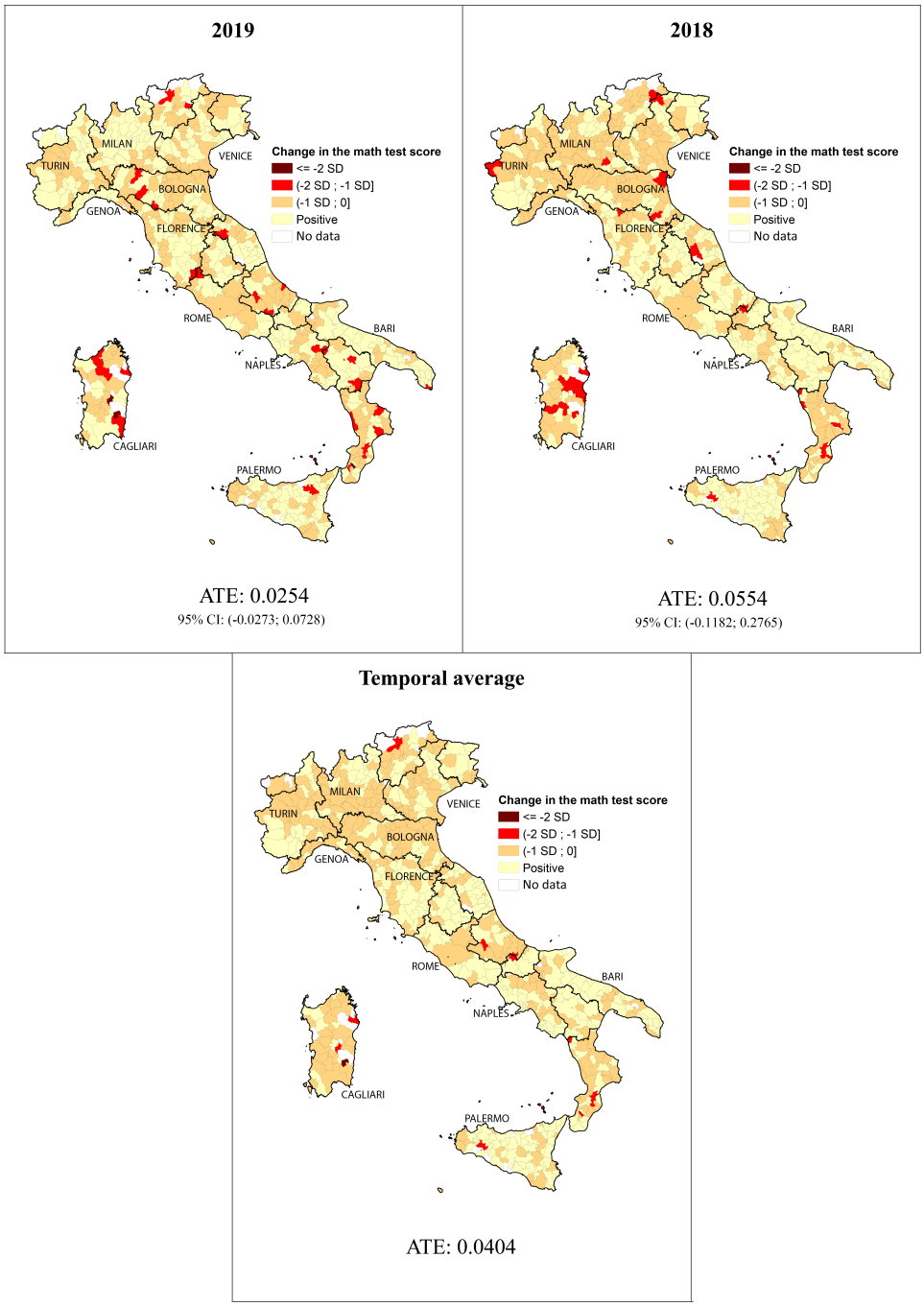


Figure E.2: Unit-level panel placebo test by year.

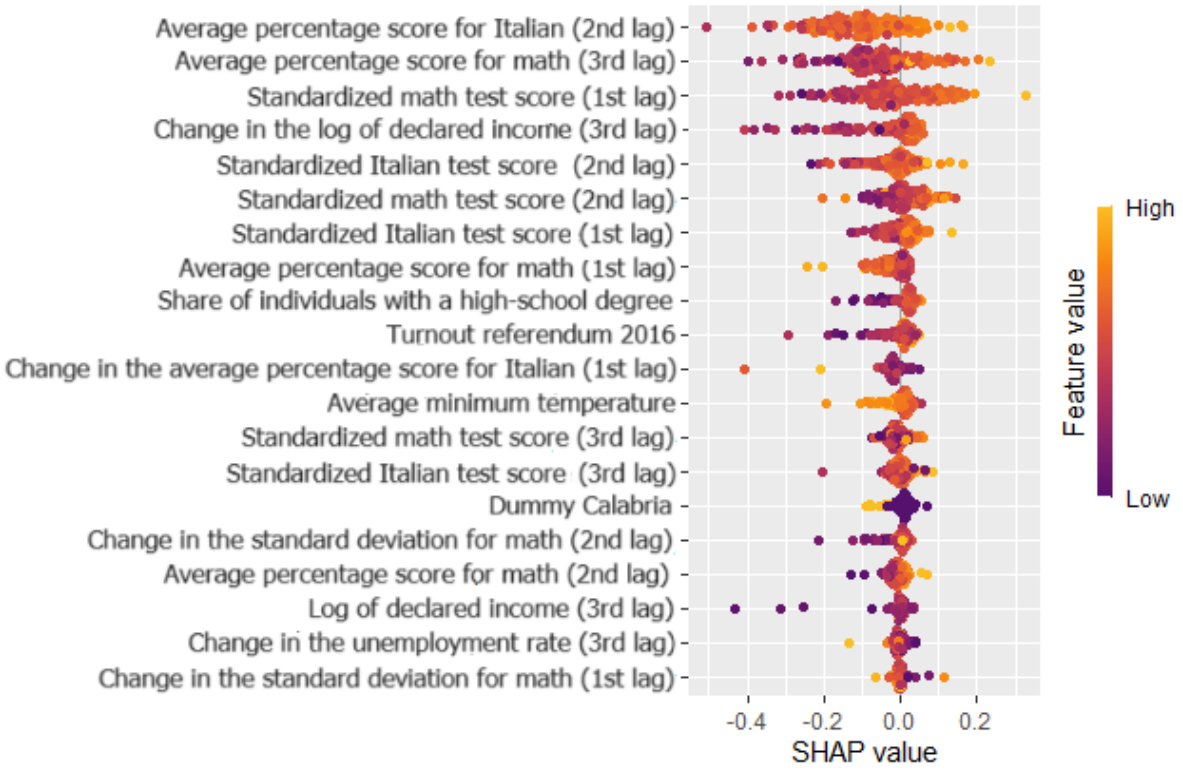


Figure E.3: SHAP values of the counterfactual 2020 forecasts produced by the final selected model (random forest) for a random 20% subsample of LLMs.

F Replication study

In this section, we revisit the empirical application of Callaway and Sant'Anna (2021) on the impact of minimum wage policy on teen employment in the United States (US). The main goal of this exercise is to assess whether the MLCM can replicate findings from a study that used a popular method based on the canonical use of control units. Additionally, we demonstrate how the MLCM can be easily adapted to a setting with staggered adoption of treatment. For the purpose of this replication, we simulate an environment in which the researcher has access to but cannot employ untreated units due to violations of the no-interference assumption. We check the reliability of these estimates by using the results of Callaway and Sant'Anna (2021), obtained via their difference-in-differences estimator for staggered adoption contexts with multiple time periods, as the ground truth.

We use county-level data for the period 2001-2007 from Callaway and Sant'Anna (2021) and apply only minimal pre-processing before proceeding with estimation. First, we discard never-treated units. Second, we employ the same set of variables included by Callaway and Sant'Anna (2021) in their specification with covariates: region dummies, county population, county median income, the fraction of white population that is white, the fraction of the population with a high school education, the county's poverty rate. Third, we also include, as additional predictors, a lag of the outcome variable (the log of teen employment), county fixed effects—using the approach recommended by Johannemann et al. (2019)—and region-year dummies.

We then proceed to estimate the effects of staggered implementation of minimum wage policy on teen employment by forecasting counterfactuals with the MLCM. The key point to keep in mind is that not only we discard never-treated units, but we also do not rely on the available not-yet-treated units to build counterfactuals. Regarding the identification assumptions, we assume additivity and no anticipation, as in Callaway and Sant'Anna (2021), but we replace their parallel trends assumption with our Assumption 4 that implicitly rules out concomitant shocks and policies. This assumption appears reasonable given the context: the years under analysis are those just before the Great Recession, during which no other major employment shock occurred.[31] We also replace their standard SUTVA assumption, which postulates no interference among units, with our weaker version of SUTVA (Assumption 1), which only requires no hidden versions of the treatment.

We employ two versions of the MLCM: one with a fully non-linear model, random forest, and the other with a linear model, LASSO. LASSO outperformed random forest on this specific dataset across all treatment cohorts, making it the chosen method for forecasting counterfactual teen employment levels in the post-treatment periods.[32]

[31] Youth unemployment trends remained relatively stable in the US during these years, with a consistent value of 10.5% in both 2001 and 2007 (the initial and final years of the dataset); see data from the World Bank here.

[32] For this reason, we do not need to rely on Assumption 5, which is only required for identification in

The results of this no-control replication exercise are reported in the event-study graph of Figure F.1 below. This graph illustrates group-time average treatment effects (i.e., the estimand proposed by Callaway and Sant'Anna (2021)) and should be compared with the two sets of estimates reported in Figure 1 of the original study.

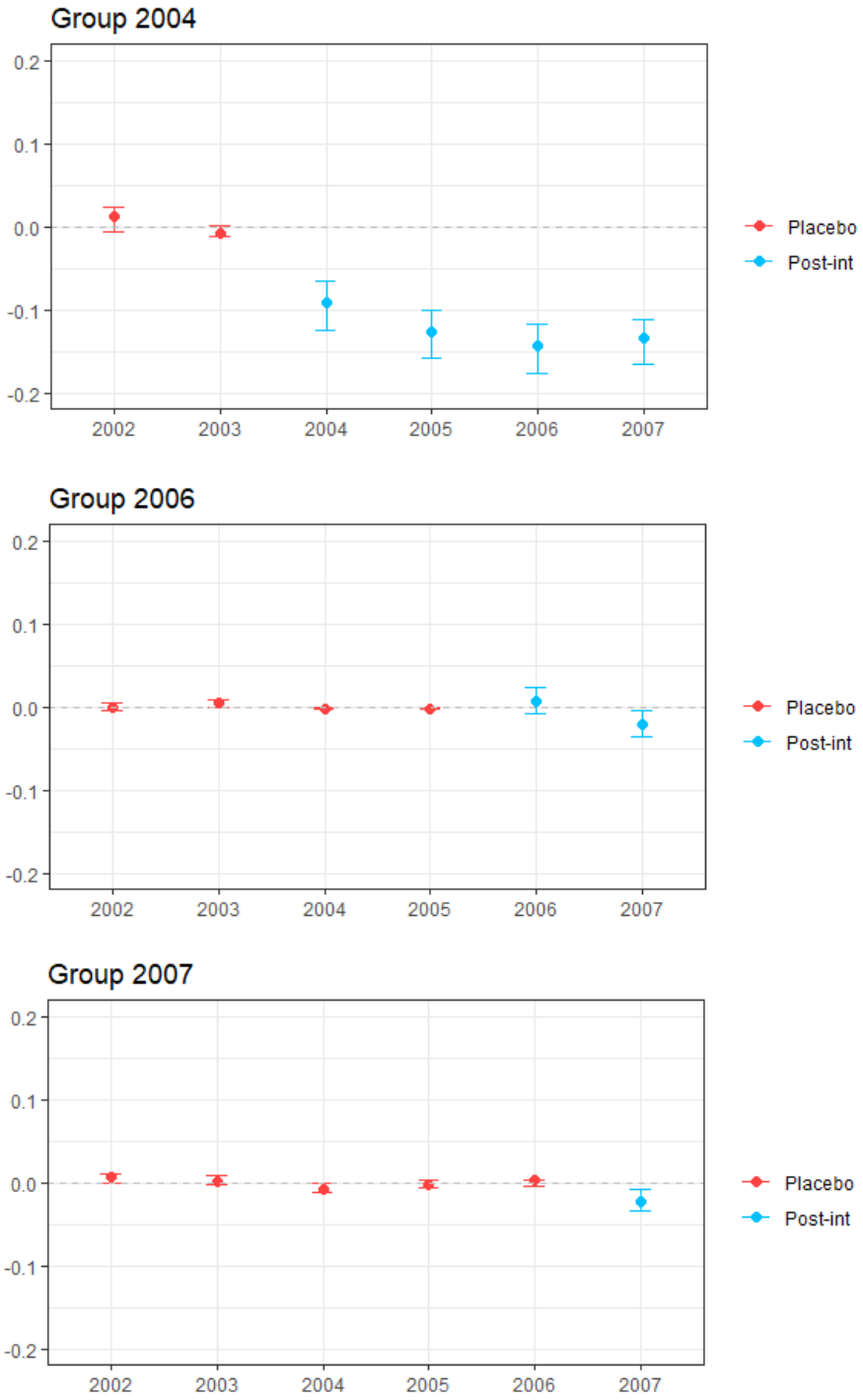


Figure F.1: Group-time average treatment effects – Replication without controls of Callaway and Sant'Anna (2021). *Note*: Confidence intervals estimated using 1000 block-bootstrap replications. To ease comparability, the range of the Y axis is the same used by Callaway and Sant'Anna (2021).

non-linear multi-step ahead forecasting.

Despite completely discarding information from untreated units, we closely replicate their findings: the sign of the post-treatment estimates is always consistent (with the sole exception of the 2006 effect for the cohort first treated in 2006), while the magnitude differs to some extent but consistently falls within the range of confidence intervals from the two specifications of Callaway and Sant'Anna (2021) with conditional and unconditional parallel trends. The qualitative story remains unchanged, pointing to negative impacts of the minimum wage increase on county-level teen employment: our estimate of the 'global' average treatment effect is –3% (95% confidence intervals: –3.8; –2), which aligns with the estimate reported by Callaway and Sant'Anna (2021) of –3.2% for their specification with conditional parallel trends.[33] Lastly, pre-treatment estimates are all indistinguishable from zero. This is an improvement over the original study, where some of the placebo estimates were relatively large and statistically significant, casting doubts on the validity of the parallel trends assumption (Callaway and Sant'Anna, 2021).

Moreover, our county-specific treatment effects can be exploited to investigate patterns of treatment effect heterogeneity. One approach is to leverage these unit-level estimates by mapping them to explore potential spatial patterns. Figure F.2 shows county-specific effects, revealing highly heterogeneous estimates across states. Notably, the minimum wage had a particularly strong negative impact in Michigan, which is also the state that increased the minimum wage the most during the period under analysis (Callaway and Sant'Anna, 2021).

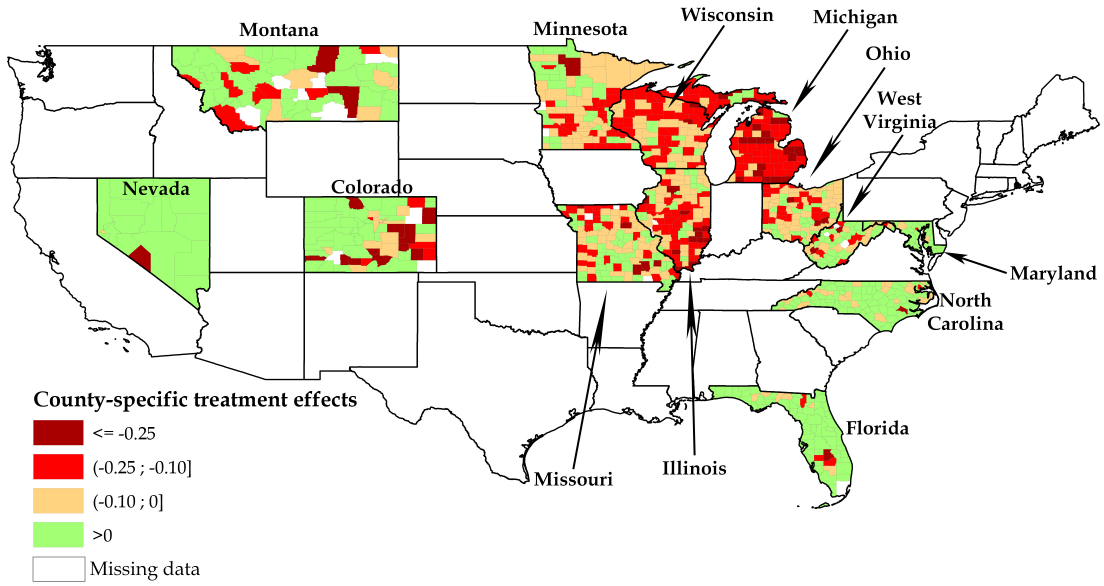


Figure F.2: County-specific treatment effects of the minimum wage on teen employment. *Notes*: These estimates are individual treatment effects averaged over the post-treatment period. The final sample used by Callaway and Sant'Anna (2021) includes county-level teen employment data for 29 states, matched with county characteristics. Of the 29 states considered in the Callaway and Sant'Anna (2021) analysis, 13 are treated. These 13 states comprise 930 counties. However, due to missing data in the original dataset, the number of treated counties is 907.

[33] Callaway and Sant'Anna (2021) refer to this estimate as the 'simple weighted average'. It is the weighted average (by group size) of all available group-time average treatment effects.

This replication study demonstrates the flexibility and applicability of the MLCM in staggered adoption settings or in cases with potential spillovers and general equilibrium effects contaminating the outcomes of untreated units. It also highlights that the MLCM can be used as a robustness check for traditional identification strategies, offering an alternative estimation method to assess the plausibility of the key identification assumptions, validate the main estimates, and rule out spillover effects. Finally, the MLCM generates individual treatment effects, providing a more granular building block compared to the group-time average effects of Callaway and Sant'Anna (2021) and most other traditional estimators leveraging control units, thus allowing for a more comprehensive investigation into the heterogeneity of impacts.