Assessing the Heterogeneous Impact of Economy-Wide Shocks: A Causal Machine Learning Approach Applied to Colombian Firms

Abstract

Our paper presents a causal Machine Learning (ML) methodology to study the heterogeneous effects of economy-wide shocks and applies it to the impact of the COVID-19 crisis on exports. This method is applicable in scenarios where, due to the pervasive nature of the shock, it is difficult to identify a control group that is not affected by the shock and to determine ex ante differences in shock intensity across units. In particular, our study investigates the effectiveness of different machine learning techniques in predicting firms' trade and, by building on recent developments in causal ML, these predictions are used to reconstruct the counterfactual distribution of firms' trade under different COVID-19 scenarios and investigate the heterogeneity of treatment effects. Specifically, we focus on the probability of Colombian firms surviving in the export market under two different scenarios: a COVID-19 setting and a non-COVID-19 counterfactual situation. On average, we find that the COVID-19 shock decreased a firm's probability of surviving in the export market by about 20 percentage points in April 2020. We study the treatment effect heterogeneity by employing a classification analysis that compares the characteristics of the firms on the tails of the estimated distribution of the individual treatment effects.

Keywords: Economy-wide shocks; Causal Machine Learning; Heterogeneous Treatment Effects; COVID-19; International Trade.
JEL Codes: F14; F17; D22; L25

1 Introduction

This paper presents a methodology to study the heterogeneous effects of economy-wide shocks, applicable in scenarios where neither a control group unaffected by the shock nor an ex-ante definition of the intensity of the shock for each unit is possible. We apply this methodology to the impact of the COVID-19 crisis on exports. In particular, we aim to estimate the causal effect of COVID-19 on a firm's probability of survival in the export markets and to study the heterogeneity of this effect. The main hurdles for this evaluation task are related to the pervasiveness of the COVID-19 shock. On the one hand, the fact that all firms are eventually exposed to the effects of the COVID-19 crisis makes it hardly possible to find a control group of firms to be used to build a counterfactual non-COVID-19 scenario. On the other hand, adopting a continuous treatment variable would imply defining ex-ante the main patterns through which the COVID-19 shock has affected firm-level trade. This task is highly demanding, given that the economy-wide impact of the shock is coupled with complex interdependencies between firms and products across sectors and countries. The paper's main idea is to address these evaluation challenges, which are present when studying the heterogeneous impact of economy-wide shocks, by leveraging the predictive capabilities of Machine Learning (ML) techniques.

To face the COVID-19 crisis, governments implemented social distancing and lockdown policies, exacerbating supply and demand shocks (World Bank, 2020). In a interconnected world, the impact of the pandemic on international trade has gained great attention (Felbermayr and Görg, 2020; Antràs et al., 2023; Bonadio et al., 2020; Evenett, 2020). Global trade, which is typically more volatile than output and tends to fall sharply during a crisis, has shown the biggest fall since the 2009 global financial crisis. From the beginning of the COVID-19 epidemic, scholars underlined that, though its impact on international trade could have been comparable to the Great Trade Collapse of 2008-2009, this time, the demand-side shock is accompanied by a supply-side shock (Baldwin and Tomiura, 2020). Moreover, this supply-side effect could be reinforced by a supply-side contagion via importing/supply chains, which have grown in relevance during the last decade (Antras and Chor, 2022). In other words, supply disruptions in the countries providing intermediate inputs to a given country are likely to hurt also its export performance (Halpern et al., 2015; Navas et al., 2020).

We focus on Colombian exporters because of the vulnerability of the Colombian economy to the COVID-19 shock and the availability of detailed customs data. As in many other developing and industrialized countries, Colombia experienced domestic supply and demand shocks in 2020, with factory closures, the suspension of some public services and supply chain disruptions.

By interpreting exporters' dynamics as a complex learning process,[1] this paper's first

[1]Firms have heterogeneous and incomplete information about the trade opportunities. This is true both on the exporting and the importing side of firm activities. For example, in Albornoz et al. (2012) and Eslava

contribution is exploring and comparing the effectiveness of different ML techniques in predicting firms' trade status in two different scenarios, a COVID-19 and a non-COVID-19 setting. ML techniques have been successfully applied to predict firm performances in high-dimensional contexts (Bargagli-Stoffi et al., 2021) in which the number of potentially relevant explanatory variables is very high. Our paper fits into a nascent literature that is applying ML techniques to study international trade patterns (Breinlich et al., 2022) and, up to what we know, in our study for the first time ML techniques are used to investigate firm-level international trade performance. Estimating more accurately the likelihood of a firm's success in exporting could be useful to increase the effectiveness of export promotion agencies (Van Biesebroeck et al., 2015) by helping them target their activities. However, the effectiveness of ML in improving the prediction of a firm's success cannot be taken for granted, especially for developing countries, as shown by McKenzie and Sansone (2019).

This paper's second and main contribution is to show how to use these predictions to estimate the causal effect of the COVID-19 shock at the firm level and to study its possible heterogeneity. We use the estimated ML model with the best performance in predicting the 2019 export status of firms exporting in 2018 to build a 2020 non-COVID-19 counterfactual outcome for firms exporting in 2019. Then, we compare these counterfactual non-COVID-19 firm-level export probabilities with the predicted probabilities of the best-performing ML model using the characteristics of 2019 exporters to predict their export status in 2020. The latter estimated probabilities summarize the information on the observed COVID-19 scenario and express it in a metric comparable with the estimated counterfactual non-COVID-19 outcomes. In the literature using ML counterfactuals when no control group is available (Cerqua and Letta, 2020; Fabra et al., 2022), it is instead common to estimate causal effects by comparing the counterfactual predictions with the observed outcome in case of treatment, following the so-called "consistency assumption": if the outcome in case of treatment is observed then it also represents the potential outcome under treatment. We follow Chernozhukov et al. (2023)[2] by using ML techniques to reconstruct firm potential outcomes in the case of no treatment and also to predict the outcomes in the treatment scenario. From a methodological standpoint, our study represents the pioneering application and adaptation of the generic machine learning tools proposed by Chernozhukov et al. (2023) in a context where a control group is unavailable.[3] Furthermore, we provide guidance on utilizing in-time placebo tests to assess the credibility of counterfactual estimates. Additionally, we compare the estimation results of the average treatment effect and treatment effect heterogeneity obtained by employing the predicted outcomes in the case of treatment, as proposed by

et al. (2015) exporting firms are uncertain and learn about the appeal of their products and, more in general, about the profitability of exporting their products on the international markets. By searching for clients and observing their realized profitability, firms update their beliefs about their capabilities in international markets.

[2]See formulas 2.6 and 2.7.

[3]For applications of the generic machine learning methodology in economics see Deryugina et al. (2019); Magnan et al. (2021); Baiardi and Naghi (2024); Buhl-Wiggers et al. (2024).

Chernozhukov et al. (2023), with those obtained using the observed outcomes of treated units (i.e., following the Cerqua and Letta (2020) and Fabra et al. (2022) approach). Our findings suggest that while the estimates of the average treatment effect remain robust across methodologies, the former approach should be preferred when the objective is to identify the observations with the highest and lowest treatment effects, and subsequently determine the factors contributing to treatment effect heterogeneity.

Examining the heterogeneous effects of economy-wide shocks is a crucial undertaking as it represents the foundational stage in devising policy interventions intended to mitigate their deleterious outcomes and reactivate economic operations. However, from a methodological point of view, investigating the treatment effect heterogeneity is not a straightforward task when its potential determinants are many. The traditional approach splits the sample into groups to assess the significance of the difference in the treatment effects of the groups. Unfortunately, this approach is prone to overfitting, and finding statistically significant differences out of all possible splits might be entirely due to random noise. Recently, new tools based on ML have been developed to identify subgroups that are particularly responsive to the treatment (Athey et al., 2019; Chernozhukov et al., 2023). Building on the recent progress in causal ML application to the analysis of heterogeneous effects, in this paper we adopt an agnostic ML model to investigate treatment effect heterogeneity. In particular, we interpret the estimated effects stemming from our ML counterfactual empirical model by using the Sorted Effects method (Chernozhukov et al., 2018, 2023). This method focuses on the tails of the estimated distribution of the firm-level treatment effect to identify the units that are most affected and those that are least affected by the treatment (whose characteristics are compared). We provide evidence that contrasting the estimated counterfactual outcomes with the outcomes predicted for the treatment scenario (and not directly with the observed outcomes under treatment) is crucial to correct the estimation error arising from the imperfect reconstruction of the unobservable counterfactual.

Our paper is connected to the literature on the heterogeneous impact of the COVID-19 shock on trade. Using firm-level monthly data on Spanish trade in goods, de Lucio et al. (2020) find that exports decreased more in countries that introduced strict policies to contain COVID-19 and for goods that are consumed outside the household, particularly between March and May, showing how Spain's export performance during the pandemic depends on COVID-19-induced demand shocks in export markets and the characteristics of products. Using monthly bilateral product-level trade flows that cover three-quarters of world trade, Berthou and Stumpner (2024) also find that the impact of the COVID-19 shock on exports was particularly strong in the spring of 2020, and that demand shocks related to COVID-19 impacted exports directly (shocks in importing countries) but also indirectly (shocks in third countries). Using a sector-level gravity model, Espitia et al. (2021) show that, during the COVID-19 crisis, sectors that tend to be relatively less internationally integrated suffered less from foreign shocks but were more vulnerable to domestic shocks. Using data on Chinese

imports at the country-product level, also Liu et al. (2021) show that the COVID-19 effects are heterogeneous, being weaker for medical goods and stronger for durable consumption goods. All these papers base their identification strategy of the average COVID-19 effect on the cross-country differences in the implementation of lockdown measures over time and study treatment effect heterogeneity by focusing on subsamples or interacting the treatment variable with other possible determinants of heterogeneity. We share with these studies the ambition to estimate the causal impact of COVID-19 on trade and its possible heterogeneity. However, we use a different approach based on constructing a counterfactual using the predictive power of ML that, as explained above, recognises that all firms are directly or indirectly affected by this economy-wide shock and that it is very challenging to define ex-ante a variable summarising the (differential) intensity of the shock for each firm. Moreover, we implement the heterogeneity analysis by using a classification analysis that safeguards against the risks of overfitting and multiple testing. Among the possible determinants of heterogeneity, we also consider a firm's diversification on the export and import side. Therefore, our study is also related to the international trade literature on the role of diversification in mediating the impact of adverse shocks (Kramarz et al., 2020; Grossman et al., 2021; Lafrogne-Joussier et al., 2022).

Using our innovative ML approach, we find that the COVID-19 shock reduced the probability of a Colombian firm surviving in the export market in April 2020 by around 20 percentage points. Our analysis of the estimated distribution of treatment effects shows that there is considerable heterogeneity behind these average results. We highlight that more affected firms tend to be small-sized and more exposed to export destinations and import source countries that are more severely hit by the containment policies related to the COVID-19 shock. We identify the firms most and least affected by COVID-19 and compare their characteristics by combining the Sorted Partial Effects methodology with our causal ML approach, which shows that integration into global value chains on the import side is an important determinant of exporters' resilience to the COVID-19 shock. Our findings contribute to the development of targeted recovery policies by identifying the firms most affected by exogenous widespread shocks.

The paper is structured as follows. Section 2 details our empirical strategy. Section 3 presents the firm-level data, variables employed in the analysis, and descriptive statistics. Section 4 reports the estimation results, and Section 5 summarizes our findings and discusses the relevance and limitations of our analysis.

2 Methodological framework

This section lays out our empirical approach to estimating the effect of an economy-wide shock on firms' survival probabilities in export markets and exploring its heterogeneity based on firms' observable attributes.

2.1 Our Causal ML setup

We aim to study the (conditional) average effect of an economy wide shock (e.g., the COVID-19 effect in our specific application) on the probability that the cohort of firms that were exporting in a given month during the pre-shock year t_{s-1} (e.g., year 2019 in our specific application) will export again during the same month of the year of the shock t_s (that is 2020 in our specific application).[4] Therefore, the empirical analysis is carried out separately for each month.[5] This allows the effects of the explanatory variables (e.g., the hypothesized determinants of firm export status) to vary throughout the year.

For economy-wide shocks such as COVID–19, there is no unambiguous definition of an "untreated" group because, plausibly, all firms are subject to the shock. Consequently, if we define the potential outcome for firm i at time t under treatment status $D \in \{0, 1\}$ as Y_{it}^{D}—where D indicates the presence of the shock—the standard Conditional Independence Assumption (CIA), $Y_{i,t_s}^{0} \perp\!\!\!\perp D_{i,t_s} \mid X_{i,t_{s-1}}$, used to identify the Average Treatment Effect on the Treated (ATT) when a contemporaneous control group is available cannot be invoked as the assumption of common support is violated since $P(D_{i,t_s} = 1 \mid X_{i,t_{s-1}}) = 1$. Indeed, in this setting, the ATT coincides with the Average Treatment Effect (ATE) for the cohort of COVID–19-exposed firms, ATE $= \mathbf{E}(Y_{i,t_s}^{1} - Y_{i,t_s}^{0})$.[6] Furthermore, even an identification strategy based on comparing individual firms subject to different treatment intensities does not seem feasible due to the complex and ex-ante unknown paths through which firms are potentially exposed to treatment. Though we study whether treatment effect heterogeneity depends, inter alia, on firm-specific measures of the intensity of the COVID-19 shock,[7] the intensity of treatment effect might also depend on other firms' characteristics, such as the identity of suppliers and clients, the characteristics of the traded final product, among many others, that we cannot know in advance and whose interactions are *a priori* unknown.

Therefore, we refer to all the observations $(Y_{i,t_s}, X_{i,t_{s-1}})$ for the cohort of firms that exported in t_{s-1} as the treated group (i.e., all observations belonging to our sample at t_s). Moreover, invoking the so-called consistency assumption, we assume that for the treated group the observed outcome in the year of the shock Y_{i,t_s} represents the potential outcome

[4] Although the primary analysis focuses on the extensive margin, the proposed methodology is general and can be readily extended to continuous outcomes, allowing an analysis of the intensive margin. Descriptive evidence for the intensive margin is provided in the Appendix.

[5] In line with the literature on this topic (see, e.g., de Lucio et al., 2020; Berthou and Stumpner, 2024; Espitia et al., 2021; Liu et al., 2021), we perform a monthly analysis as the COVID-19 shock has evolved rapidly and unevenly over the months in 2020, as described in the Appendix A.

[6] While we assume ATT = ATE because we think that all firms are subject to the economy-wide COVID-19 shock, this does not imply that all firms experience a non-zero effect. Some treated firms may have a negligible or even positive impact from the shock. In such cases, our ATT estimate is not biased but simply reflects treatment–effect heterogeneity. However, if some firms were in fact not subject to the COVID-19 shock (i.e., untreated), our methodology would underestimate the ATT.

[7] These indexes are described in detail in section 3. They are based on firms' past export and import activities in different countries and on the time-varying strength of the virus and the stringency of the policies aimed at mitigating its spread.

in case of treatment Y^1_{i,t_s}.[8]

As is common when studying the effects of widespread shocks, we must therefore use the information about behavior before the shock to estimate the counterfactual behavior (in the hypothetical situation without the shock) during the actual shock. This process involves forecasting the future conduct of entities based on their historical behavior, an application perfectly suited to ML techniques, which are designed for such out-of-sample prediction tasks. In line with the reasoning of Varian (2016), and drawing parallels with the applications employed by Cerqua and Letta (2020), and Fabra et al. (2022), we harness the predictive strength of ML techniques. This allows us to construct a hypothetical scenario for firm-level outcomes during the shock period, using pre-shock data concerning firms' export behaviors and attributes. We will use information on the export status in a given month of t_{s-1} (that is 2019 in our specific application) for firms that were exporting in the same month of year t_{s-2} (e.g., year 2018 in our specific application) and the observed characteristics of such firms in t_{s-2} to learn the counterfactual function that we apply to the treated group for estimating $Y^0_{it_s}$. We refer to the observations $(Y_{it_{s-1}}, X_{it_{s-2}})$ for the cohort of firms that exported in t_{s-2} as the control group.

The main assumptions used to reconstruct the unobserved counterfactual outcome during the year of the shock using the pre-shock observed firms' behaviour are: (i) absence of anticipatory effects of the shock on covariates measured at t_{s-1} and t_{s-2} and on the outcome at t_{s-1}, (ii) stability of the counterfactual function in time and (iii) a common support assumption. They are explained in detail below.

(i) *No anticipation effects on outcomes and covariates.* Neither observed outcomes at $t_s - 1$ nor covariates at $t_s - 2$ and $t_s - 1$ are affected by the shock happening at t_s:

$$Y_{i,t_{s-1}} = Y^0_{i,t_{s-1}}, \quad X_{i,t} = X^0_{i,t} \quad \text{for} \quad t = (t_s - 1, t_s - 2) \tag{1}$$

Notice that, since the treatment occurs at t_s, for (1) to hold it is sufficient to rule out any effect of the future shock at t_s on the observed $Y_{i,t_{s-1}}$. Moreover, (1) implies that the CIA holds for the control group at t_{s-1}, that is, $Y^0_{i,t_{s-1}} \perp\!\!\!\perp D_{i,t_{s-1}} \mid X_{i,t_{s-2}}$. Consequently,

[8] The consistency assumption corresponds to the first component of the Stable Unit Treatment Value Assumption (SUTVA; Keele, 2015), which requires that there are no hidden forms of treatment. We maintain this assumption by adopting a deliberately broad definition of treatment — being in the sample at time t_s — which, while encompassing a potentially different intensity of the shock for each unit, remains useful for generating policy recommendations concerning the firms relatively more affected by the shock. These recommendations are informative even if treatment-effect heterogeneity is partly confounded by heterogeneity in the treatment itself, since firms that are more affected may be so either because their characteristics are correlated with higher treatment intensity or because the treatment interacts with their characteristics. The second component of SUTVA — the no-interference assumption — is not relevant in our setting because we do not rely on a contemporaneous control group that could be indirectly affected and our focus is on estimating the total effect of the treatment, which by construction includes both the direct effect on a unit from the treatment it receives and any indirect effects arising from spillovers or general-equilibrium adjustments. Disentangling between direct and indirect effects lies beyond the scope of this paper.

at t_{s-1} the conditional expectation of the observed outcome coincides with that of the potential outcome under no treatment: $\mathbf{E}[Y^0_{i,t_{s-1}} \mid X_{i,t_s-2}] = \mathbf{E}[Y_{i,t_{s-1}} \mid X_{i,t_s-2}]$.

(ii) *Stability of the Counterfactual Function.* This assumption is about the stability of the function that expresses the expected value of the conditional potential outcome in time. Define $Y^0_{i,t} = f^0_t(X^0_{i,t-1}) + u^0_{i,t}$, where $f^0_t(\cdot)$ is a generic model or function representing the relationship between explanatory variables and the outcome in the absence of the shock such that $\mathbf{E}[Y^0_{i,t}|X^0_{i,t-1}] = f^0_t(X^0_{i,t-1})$. Under (i), for $t = t_s - 1$ we have that $Y_{i,t_s-1} = f^0_{t_s-1}(X_{i,t_s-2}) + u^0_{i,t_s-1}$ such that $\mathbf{E}[Y_{i,t_s-1}|X_{i,t_s-2}] = f^0_{t_s-1}(X_{i,t_s-2})$. The second assumption states that the function f^0_t does not depend on t, i.e., it is stable over the two considered years:

$$f^0_{t_s-1} = f^0_{t_s} = f^0. \tag{2}$$

Under assumptions (i) and (ii), if $X_{i,t_{s-1}} = X_{i,t_s-2}$ then $\mathbf{E}[Y^0_{i,t_s} \mid X_{t_{i,s-1}}] = \mathbf{E}[Y^0_{i,t_{s-1}} \mid X_{i,t_s-2}]$. In other words, the conditional expectation of the potential outcome under no treatment at t_s coincides with that at t_{s-1}.

(iii) *Common support.* The support of the distribution of the explanatory variables of the firms belonging to the treated group is included in the support of the distribution of the explanatory variables of the firms belonging to the control group:

$$P(D_{i,\{t_s,t_s-1\}} = 1|X_{i,\{t_s-1,t_s-2\}}) = e(X_{i,\{t_s-1,t_s-2\}}) < 1 \tag{3}$$

where $D_{i,\{t_s-1,t_s\}}$ is a dummy variable indicating whether an observation belongs to the treated group or to the control group, and $X_{i,\{t_s-1,t_s-2\}}$ are the corresponding explanatory variables. Therefore, this expression defines a condition on the values of the propensity score, which we denote as $e(X_{i,\{t_s-1,t_s-2\}})$. This assumption allows nonparametric identification of the (conditional) average effects.

Thanks to these assumptions, when $X_{i,t_{s-1}} = X_{i,t_s-2}$, we have that $\mathbf{E}[Y^0_{i,t_s} \mid X_{i,t_s-1}] = \mathbf{E}[Y^0_{i,t_{s-1}} \mid X_{i,t_s-2}] = \mathbf{E}[Y_{i,t_{s-1}} \mid X_{i,t_s-2}]$: the conditional expectation function of the potential outcome in case of no treatment at t_{s-1} can be identified by computing the conditional expectation function of the observed outcome at t_{s-1} and it coincides with the conditional expectation function of the potential outcome in case of no treatment at t_s.

Since f^0 is in practice unknown, we must estimate it. Under the above assumptions, we can write $Y^0_{i,t_s} = f^0(X_{i,t_s-1}) + u^0_{i,t_s}$, such that $\mathbf{E}[Y^0_{i,t_s}|X_{i,t_s-1}] = f^0(X_{i,t_s-1})$, and we can use data on $t_s - 2$ and $t_s - 1$ to estimate $Y^0_{i,t_s-1} = f^0(X_{i,t_s-2}) + u^0_{t_s-1}$ and retrieve $\hat{f}^0$. By applying this invariant estimated function to the covariates of $t_s - 1$, we can obtain the predictions for the counterfactual (without the shock) outcome in t_s:

$$\hat{Y}^0_{i,t_s} = \hat{f}^0(X_{i,t_s-1}) = Y^0_{i,t_s} - \overbrace{\mathcal{E}^0_{i,t_s}}^{\textit{Prediction error}} - \overbrace{u^0_{i,t_s}}^{\textit{Orthogonal error}}. \tag{4}$$

The model we utilize to derive this counterfactual (and the counterfactual itself) is referred to as the "Shock Unaware Machine" (SUM), a term acknowledging the ML techniques employed in constructing the counterfactual and the fact that no information about the shock is used in the analyses. In the application we present in this paper, we rely on the "K-fold" cross-validation method (with $K = 5$) to discriminate between the considered ML techniques. We randomly divide the set of exporters observed in $t_s - 2 = 2018$ (considering the exporting success during the same month in $t_s - 1 = 2019$ as the outcome) into 5 equally sized groups and obtain the predictions for the firms belonging to a group by estimating $Y_{i,2019} = f^0(X_{i,2018}) + u^0_{i,2019}$ with different ML models on the firms belonging to the other groups. Then we compute the accuracy of the different models for each month and choose the model with the best average performance across months. Notice that this comparison is entirely based on the pre-pandemic accuracy of the ML models by comparing the predictions $\hat{Y}_{i,2019}$ with the observed $Y_{i,2019}$, not on its merits in predicting the firms' outcomes in 2020. Finally, we obtain the $\hat{Y}^0_{i,2020}$ by estimating $Y_{i,2019} = f^0(X_{i,2018}) + u^0_{i,2019}$ on the entire set of 2018 units (also in this case month by month) and, as shown in (4), applying the estimated function $\hat{f}^0$ to the set of 2019 units. Given that during the first three months of 2020 Colombia was in practice not exposed to COVID-19 (and therefore $Y_{i,2020} = Y^0_{i,2020}$), if assumption (2) holds, we expect that in those months the accuracy of the predictions $\hat{Y}_{i,2019}$ obtained in the cross-validation step for 2019 will be very similar to the accuracy of $\hat{Y}^0_{i,2020}$ for 2020.

Following Cerqua and Letta (2020) and Fabra et al. (2022), we define the simple comparison of the observed outcome under the shock in t_s with the estimated counterfactual outcome as an estimator of the individual-specific shock effect α_i. This comparison is represented as:

$$\hat{\alpha}_i = Y_{i,t_s} - \hat{Y}^0_{i,t_s}. \tag{5}$$

This provides the full distribution of estimated individual treatment effects, that is, each unit's Conditional Average Treatment Estimate (CATE) estimate (Salditt et al., 2024).

The ATE and the CATE$_z$, that is the Conditional Average Treatment Effect for those units with $Z_{i,t_s-1} = z_{i,t_s-1}$ where Z is a subset of the variables X ($Z \subset X$),[9] are estimated by averaging these estimated individual treatment effects. Therefore, the estimators of ATE and CATE$_z$ based on $\hat{\alpha}_i$ can be defined as:

$$\bar{\hat{\alpha}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\alpha}_i, \qquad \bar{\hat{\alpha}}_Z = \frac{1}{N_z}\sum_{i\in C_z}\hat{\alpha}_i, \tag{6}$$

where N is the number of observations, $C_z = \{i : Z_{i,t_{s-1}} = z_{i,t_{s-1}}\}$ and $N_z = |C_z|$.

As it is shown below, $\bar{\hat{\alpha}}_z$ is an unbiased estimator of CATE$_z$ if in the relevant subsample the mean of the expected prediction error, $(1/N_z)\sum_{i\in C_z} E[\mathcal{E}^0_{i,t_s}]$, is zero.

[9] For example, CATE$_{textile}$ is the average treatment effect for firms belonging to the textile industry.

$$
\begin{aligned}
\mathbf{E}\left[\bar{\hat{\alpha}}_z\right] &= \mathbf{E}\left[\frac{1}{N_z}\sum_{i\in C_z}\hat{\alpha}_i\right] = \frac{1}{N_z}\sum_{i\in C_z}\mathbf{E}[\hat{\alpha}_i] \\
&= \frac{1}{N_z}\sum_{i\in C_z}\mathbf{E}\left[Y_{i,t_s} - Y^0_{i,t_s} - \mathcal{E}^0_{i,t_s} - u^0_{i,t_s}\right] \\
&= \underbrace{\frac{1}{N_z}\sum_{i\in C_z}\alpha_i}_{CATE_z} - \frac{1}{N_z}\sum_{i\in C_z}\mathbf{E}[\mathcal{E}^0_{i,t_s}] - \frac{1}{N_z}\sum_{i\in C_z}\underbrace{\mathbf{E}[u^0_{i,t_s}]}_{=0 \text{ by (i) and (ii)}}
\end{aligned}
\tag{7}
$$

Notice that in the second row, we substitute Eq. (4) plugged in Eq. (5). Similarly, $\bar{\hat{\alpha}}$ is an unbiased estimator of ATE if the mean of the expected prediction error, $(1/N)\sum_{i=1}^{N} E[\mathcal{E}^0_{i,t_s}]$, is zero, and $\hat{\alpha}_i$ is an unbiased estimator of the unit i's CATE if $E[\mathcal{E}^0_{i,t_s}] = 0$.

In what follows, we introduce alternative estimators that compare the predicted counterfactual outcomes with predicted outcomes under the treatment scenario, rather than directly with the observed treatment outcomes. The first step is to decompose the observed outcome in t_s under the shock, Y^1_{i,t_s}, into a generic model or function $f^1(X^1_{i,t_s-1})$, capturing the relationship between covariates and the outcome during the shock, and a residual component u^1_{i,t_s} orthogonal to the covariates:

$$
Y^1_{i,t_s} = f^1(X^1_{i,t_s-1}) + u^1_{i,t_s}, \quad s.t.\ \mathbf{E}[Y^1_{i,t_s} \mid X^1_{i,t_s-1}] = f^1(X^1_{i,t_s-1}). \tag{8}
$$

Given that $Y^1_{i,t_s} = Y_{i,t_s}$ and $X^1_{i,t_s-1} = X_{i,t_s-1}$, we can write:

$$
Y_{i,t_s} = f^1(X_{i,t_s-1}) + u^1_{i,t_s}, \quad s.t.\ \mathbf{E}[Y_{i,t_s} \mid X_{t_s-1}] = f^1(X_{i,t_s-1}). \tag{9}
$$

We then define an alternative estimator of the individual-specific shock effect α_i, that is, each unit's Conditional Average Treatment Estimate (CATE) estimate (Salditt et al., 2024), as the difference between the predicted outcome under the shock in t_s and the predicted counterfactual outcome in the absence of the shock for the same firm:

$$
\hat{\alpha}_i = \hat{Y}^1_{i,t_s} - \hat{Y}^0_{i,t_s}, \tag{10}
$$

where $\hat{Y}^1_{i,t_s} = \hat{f}^1(X_{i,t_s-1}) = Y_{i,t_s} - \mathcal{E}^1_{i,t_s} - u^1_{i,t_s}$.

We refer to the model used to predict Y_{i,t_s} (and the predictions $\hat{Y}_{i,t_s}$ themselves) as the *Shock Aware Machine* (SAM). The term "Shock Aware" emphasizes that this model exploits information from the observed shock scenario. Importantly, SAM predictions are expressed in the same metric as the counterfactual predictions, which are produced by the SUM, allowing them to be directly comparable.[10] In our application, the SAM expresses the outcome in 2020 of exporters operating the foreign market in 2019 as a function of their characteristics

[10]Notice that with $\hat{\alpha}_i$, we are comparing a probability (counterfactual) with a binary value (observed outcome), while with $\hat{\alpha}_i$, we are comparing two estimated probabilities.

in 2019 and the information about governments' shock-related stringency measures all over the world coming from Hale et al. (2020).[11] Similarly to the procedure followed to select the best-performing SUM, we rely on a 5-fold cross-validation strategy to obtain a 2020 prediction for each firm that exported in 2019. We randomly group the 2019 exporters into five equally sized subsets and we predict the 2020 outcomes of the firms contained in one subset by using the information of firms contained in the remaining four subsets. In other words, we train the models on a random 80% of the data and test them on the remaining 20% and we repeat the process five times for each different 20% subset, thus obtaining a 2020 prediction for each 2019 exporter.

The ATE estimator and CATE_z estimator based on $\hat{\alpha}_i$ are:

$$\bar{\hat{\alpha}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\alpha}_i, \qquad \bar{\hat{\alpha}}_z = \frac{1}{N_z}\sum_{i\in C_z}\hat{\alpha}_i, \tag{11}$$

As shown below, $\bar{\hat{\alpha}}_z$ is an unbiased estimator of the CATE_z if, in the relevant subsample, the mean of the expected prediction-error difference between SAM and SUM, $\frac{1}{N_z}\sum_{i\in C_z}\mathbf{E}[\mathcal{E}^1_{i,t_s} - \mathcal{E}^0_{i,t_s}]$, is zero:

$$\begin{aligned}
\mathbf{E}\big[\bar{\hat{\alpha}}_z\big] &= \frac{1}{N_z}\sum_{i\in C_z}\mathbf{E}[\hat{\alpha}_i] \\
&= \frac{1}{N_z}\sum_{i\in C_z}\mathbf{E}\big[Y_{i,t_s} - Y^0_{i,t_s} - \big(\mathcal{E}^1_{i,t_s} - \mathcal{E}^0_{i,t_s}\big) - \big(u^1_{i,t_s} + u^0_{i,t_s}\big)\big] \\
&= \underbrace{\frac{1}{N_z}\sum_{i\in C_z}\alpha_i}_{CATE_z} - \frac{1}{N_z}\sum_{i\in C_z}\mathbf{E}[\mathcal{E}^1_{i,t_s} - \mathcal{E}^0_{i,t_s}] - \underbrace{\frac{1}{N_z}\sum_{i\in C_z}\mathbf{E}[u^1_{i,t_s} - u^0_{i,t_s}]}_{=0 \text{ by (i), (ii) and (8)}}.
\end{aligned} \tag{12}$$

Similarly, $\bar{\hat{\alpha}}$ is an unbiased estimator of the ATE if $\frac{1}{N}\sum_{i=1}^{N}\mathbf{E}[\mathcal{E}^1_{i,t_s} - \mathcal{E}^0_{i,t_s}] = 0$, and $\bar{\hat{\alpha}}_i$ is an unbiased estimator of the unit i's CATE if $E[\mathcal{E}^1_{i,t_s} - \mathcal{E}^0_{i,t_s}] = 0$.

Given the definitions of SAM and SUM, to simplify the reasoning in the following, we refer to Eqs. (5) and (10), respectively as

$$\hat{\hat{\alpha}}_i = Y - \hat{Y}_{SUM} = Y - SUM. \tag{13}$$

$$\hat{\alpha}_i = \hat{Y}_{SAM} - \hat{Y}_{SUM} = SAM - SUM. \tag{14}$$

The conditions behind these identification results are not directly testable as they are expressed in terms of the expected values of the prediction error $\mathcal{E}^0_{i,t_s}$ that is a function of the unobservable counterfactual Y^0_{i,t_s}. Table 1 distinguishes the five different scenarios concerning the expected values of $\mathcal{E}^0_{i,t_s}$ and $\mathcal{E}^1_{i,t_s}$ that are relevant in determining whether applying the statistic $\mathbf{T}$ to $Y - SUM$ and $SAM - SUM$ is able to recover the corresponding treatment

[11] See subsection 3.1. We do not introduce these variables explicitly as an argument of $f^1()$ to simplify notation.

effect estimand (e.g., whether averaging the estimated individual treatment effects would recover the average treatment effect).

Table 1: Identification of generic functions of the individual treatment effects, $\mathbf{T}$, according to the corresponding value taken by the prediction errors

	$\mathbf{T}(SAM - SUM)$	$\mathbf{T}(Y - SUM)$
$\mathbf{T}[E[\mathcal{E}^1_{i,t_s}]] \neq 0$ and $\mathbf{T}[E[\mathcal{E}^0_{i,t_s}]] = 0$	×	✓
$\mathbf{T}[E[\mathcal{E}^1_{i,t_s}]] = \mathbf{T}[E[\mathcal{E}^0_{i,t_s}]] = 0$	✓	✓
$\mathbf{T}[E[\mathcal{E}^1_{i,t_s}]] = 0$ and $\mathbf{T}[E[\mathcal{E}^0_{i,t_s}]] \neq 0$	×	×
$\mathbf{T}[E[\mathcal{E}^1_{i,t_s}]] = \mathbf{T}[E[\mathcal{E}^0_{i,t_s}]] \neq 0$	✓	×
$\mathbf{T}[E[\mathcal{E}^1_{i,t_s}]] \neq \mathbf{T}[E[\mathcal{E}^0_{i,t_s}]] \neq 0$	×	×

The estimators based on $Y\text{-}SUM$ identify the population parameters when $\mathbf{T}[E[\mathcal{E}^0_{i,2020}]] = 0$. The estimators based on $SAM\text{-}SUM$ are unbiased whenever $\mathbf{T}[E[\mathcal{E}^1_{i,2020}]] = \mathbf{T}[E[\mathcal{E}^0_{i,2020}]]$. Under the assumption that the strength of the COVID-19 effect on export propensity was at most very limited during the first quarter of 2020, we will use the out-of-sample prediction errors for the first quarter of 2020 as a proxy for the unobservable behavior of $\mathcal{E}^0_{i,2020}$ in the following months. Moreover, as explained in detail in section 4.2, the distribution of the estimated treatment effects during the first quarter will be used to check the credibility of the above assumptions for the set of all 2019 exporters and for different subsets of 2019 exporters defined according to their characteristics $X_{i,2019}$ or to their position in the distribution of such effects.

The inference is performed using bootstrap. Specifically, bootstrap samples are drawn by resampling the training and testing datasets with replacement, preserving their original sizes and proportions, and repeating this process 100 times per month. For each bootstrap iteration, out-of-sample predictions are generated using each ML model trained on the resampled datasets, with hyperparameters fixed at the values previously optimized via cross-validation. Once the predictions are made, the SUM and SAM are calculated as described above for each bootstrap sample within each month. To construct confidence intervals, the predictions across all bootstrap replicates are aggregated, and the empirical distribution of the resulting estimates is used to calculate the percentile-based bounds for the 95% confidence interval, thus capturing the uncertainty in the predicted effects due to sampling variability. However, as a note of caution, we remark that the literature on causal ML has shown that estimators based on ML estimation of the conditional expectation function of potential outcomes, such as $\bar{\hat{\alpha}}$ and $\bar{\alpha}$, inherit the slow convergence rates of the ML method on which they are based and are not asymptotically normal, making inference problematic. The problem with these estimators is that the moment conditions on which they are based are not Neyman-orthogonal. To do a robustness check for the average treatment effects,

we use the AIPW-Double ML estimator (with 5-fold cross-fitting and nuisance parameters estimated with Generalized Random Forest) that is consistent and asymptotically normal for $ATT = ATE = E(Y^1_{t_s} - Y^0_{t_s})$ and $ATE_{\{t_s,t_s-1\}} = E(Y^1_{\{t_s,t_s-1\}} - Y^0_{\{t_s-1,t_s\}})$ (Chernozhukov et al., 2018), in which the average potential outcomes are expressed with a moment condition which is Neyman Orthogonal. $ATT = ATE = E(Y^1_{t_s} - Y^0_{t_s})$ is the parameter we aim to estimate with $\bar{\hat{\alpha}}$ and $\bar{\alpha}$. $ATE_{\{t_s,t_s-1\}} = E(Y^1_{\{t_s,t_s-1\}} - Y^0_{\{t_s,t_s-1\}})$ is the average treatment effect obtained considering the cohorts of treated and control firms together as a unique sample. To identify $ATE_{\{t_s,t_s-1\}}$, we have to make the following additional assumptions: $Y^1_{t_s} \perp\!\!\!\perp D_{t_s} \mid X_{t_{s-1}}$; $\mathbf{E}[Y^1_{t_s} \mid X_{t_{s-1}}] = \mathbf{E}[Y^1_{t_{s-1}} \mid X_{t_{s-2}}]$; $Y_{t_s} = Y^1_{t_s}$; $P(D_{\{t_s,t_s-1,\}} = 1 | X_{\{t_s-1,t_s-2\}}) = e(X_{\{t_s-1,t_s-2\}}) > 0$. Indeed, these assumptions are needed to identify the Average Treatment Effect on the Untreated (ATU), which is defined as $ATU = ATE_{t_{s-1}} = E(Y^1_{t_{s-1}} - Y^0_{t_{s-1}})$, because the $ATE_{\{t_s-1,t_s\}} = ATU * (1 - \sigma(\cdot)) + ATT * (\sigma(\cdot))$ where $\sigma(\cdot)$ represents the share of the treated population. Let's define the following pseudo-outcome[12]

$$\overset{ATE}{\{t_s,t_s-1\}} = \underbrace{\hat{Y}^1_{\{t_s,t_s-1\}} - \hat{Y}^0_{\{t_s,t_s-1\}}}_{\text{outcome predictions}} + \underbrace{\frac{D_{\{t_s,t_s-1\}}(Y_{\{t_s,t_s-1\}} - \hat{Y}^1_{\{t_s,t_s-1\}})}{\hat{e}(X_{\{t_s-1,t_s-2\}})} - \frac{(1 - D_{\{t_s,t_s-1\}})(Y_{\{t_s,t_s-1\}} - \hat{Y}^0_{\{t_s,t_s-1\}})}{1 - \hat{e}(X_{\{t_s-1,t_s-2\}})}}_{\text{weighted residuals}} \quad (15)$$

The AIPW-Double ML estimator (where $\hat{Y}^1$, $\hat{Y}^0$ and $\hat{e}$ are estimates of the nuisance parameters obtained with ML by cross-fitting) is the average of these pseudo-outcomes and identifies the $ATE_{\{t_s,t_s-1\}}$.[13]

Finally, following the nomenclature introduced by Künzel et al. (2019), we note that our estimator $\bar{\hat{\alpha}}_i$ is referred to in the literature as the *T-learner*. In particular, we are applying a T-Learner to the group of treated units. Although the T-learner is a consistent estimator for CATE, it may have shortcomings in finite samples due to a phenomenon known as *regularization bias*. This issue arises because the outcome models for treated and control units, denoted $\hat{f}_1(x)$ and $\hat{f}_0(x)$, are estimated separately and may be subject to different degrees of regularization. Such discrepancies are particularly pronounced when the sample sizes for the two groups differ substantially, and the machine learning methods used employ regularization schemes that are sensitive to sample size. In this case, the model trained on the smaller group may be oversmoothed, potentially introducing artificial variation into the estimated treatment effect.

A second concern relates to *regularization-induced confounding*, which may occur when the covariate distributions of treated and control units are not well aligned. Under such circumstances, the models $\hat{f}^1(x)$ and $\hat{f}^0(x)$ may be effectively trained on disjoint regions of the covariate space. Consequently, their difference could reflect underlying distributional

[12] For notational convenience, the subscript "i" is omitted in the following expression.

[13] The procedure consists of the following steps: (a) randomly partition the data into K equally sized folds; (b) for each fold k, leave it out and use the remaining $K-1$ folds to estimate the nuisance functions $\mu(d,x) = \mathbb{E}[Y_i \mid D_i = d, X_i = x]$ and $e_d(x)$; (c) predict the nuisance parameters on the left-out fold k using the estimated models, yielding cross-fitted values $\hat{\mu}^{-k}(d,x)$ and $\hat{e}_d^{-k}(x)$; and (d) repeat steps (a)-(c) so that each fold serves once as the validation set (Knaus, 2022). This cross-fitting approach ensures that no observation is used to predict its own nuisance parameters, thereby reducing overfitting and guaranteeing the validity of inference. These cross-fitted estimates are subsequently used to compute the pseudo-outcomes.

shifts rather than true treatment effect heterogeneity. This issue is often associated with settings where the estimated outcome models correspond to regions with substantially different values of the propensity score—for example, where $\hat{f}_1(x)$ is learned predominantly in areas with high propensity scores and $\hat{f}_0(x)$ in areas with low propensity scores. Since the T-learner does not explicitly adjust for the propensity score or reweight observations to balance treatment groups, it may be more susceptible to this type of bias compared to alternative estimators considering also the distribution of the propensity score, such as the DR-learner introduced in Kennedy et al. (2020) (that approximate the $E[\tilde{Y}^{ATE}_{\{t_s-1,t_s\}}|X]$ as a generic ML problem; the reader is referred to the Appendix for further details).
However, in our application, these concerns are likely to be minimal. First, we leverage a large sample size, which mitigates the risk of over-regularization. Moreover, we employ a range of ML methods with varying levels of regularization intensity and obtain highly consistent results across estimators. Second, the distribution of covariates is balanced between treated and control units as shown by the estimated propensity score which is nearly identical across groups (see Figure Appx.13). This implies that both $\hat{f}_1(x)$ and $\hat{f}_0(x)$ are estimated over comparable regions of the covariate space, thereby limiting the potential for regularization-induced confounding. Third, in practice (see Table 6, Table 7, Table 8 and the monthly comparison of the distribution of estimated CATE in the Online Appendix), the CATE results obtained for the T-Learner are very similar to those obtained with the other meta-learners using the propensity score and with Generalized Random Forest.

2.2 Treatment effect heterogeneity analysis

As a further step, we perform the heterogeneity analysis by adapting the Sorted Partial Effect (SPE) method introduced in Chernozhukov et al. (2018). Formally, the SPEs are defined as percentiles of the α_i, the individual Treatment Effects (TE), and can supply a more detailed summary of the distribution of TE than the Average Treatment Effects (ATE), commonly employed in econometric analysis. The SPEs are defined as

$$\alpha^*(u) = u^{th} - \text{ percentile of } \alpha_i. \tag{16}$$

In our setting, $\alpha^*(u)$ is a function of X_{t_s-1} defined over its distribution in the population of $t_s - 1$ exporters.

The SPEs are used to do a classification analysis (CA) that allocates the $t_s - 1$ exporters into two groups, the most and the least affected by the shock, according to whether their α are lower than $\alpha^*(25)$ or greater than $\alpha^*(75)$, respectively. Notice that, since the shock effect is negative, we have defined as the most (negative) affected units those whose α lie in the left tail of the sorted distribution of treatment effects. Finally, to study the determinants of treatment effect heterogeneity, we focus on the difference in means ($CADiff$) of the X_{t_s-1} across the most and least affected groups. In the estimation, we use sample analogues of

$\alpha^*(u)$ and $CADiff$. We calculate standard errors of $\alpha^*(u)$ and $CADiff$ by bootstrapping the entire estimation process, starting from the initial α_i estimation step.

Starting from B bootstrap replications of all the estimation steps (including the prediction stage), we calculate the $CADiff$ B times. To determine the significance of the $CADiff$ we perform a two-tailed test. The p-values are constructed as follows:

$$2 \cdot min\{Pr(S \geq t|H_0), Pr(S \leq t|H_0)\}$$

being t the observed t test statistic, $t = \frac{\widetilde{CADiff_{original}}}{\tilde{\sigma}}$, drawn from the unknown distribution S. $\tilde{\sigma}$ represents the standard deviation of the bootstrapped $CADiff$. To adjust the p-values and obtain the joint p-values taking into account that we are testing hypotheses jointly on many covariates, we reproduce the "single-step" method employed by Chernozhukov et al. (2018) to control for the family-wise error rate.[14]

The application of the SPE technique presents several advantages in our setting. First, the estimated $\alpha^*(u)s$ provide a summary of the distribution of the estimated treatment effects and, therefore, of treatment effect heterogeneity. Second, the CA identifies the subgroup of the population that is more affected by the treatment and the $CADiff$ studies how the heterogeneity of the treatment effect depends on observables without imposing (additional) functional form assumptions. Third, the $CADiff$ step provides p-value adjustments to account for the joint testing of all the covariates that are considered to detect if observables are associated with treatment effect heterogeneity. In other words, the main idea is to test the null hypothesis that there is no difference between the value of the covariates in the most and least affected group, also taking into account that we perform simultaneous inference on several variables. Simultaneous inference on multiple covariates in the CA and CADiff naturally raises the problem of multiple testing. To address this, we employ

[14] In the following is described the single-step algorithm. We will indicate the bootstrap version of a variable, v, as $\tilde{v}$ and its estimated version (on the original data) as $\hat{v}$. Moreover, $\Lambda(x)^{-u}$ will denote the first moment for the feature x of interest in the least affected group including the observational units i such that $\alpha_i < \alpha^*(u)$. Similarly, $\Lambda(x)^{+u}$ defines the first moment for the variable x of interest in the most affected group including the observational units i such that $\alpha_i > \alpha^*(1-u)$. Since we do not observe α directly, the mentioned quantities are estimated. According to the above convention, the estimated value of $\Lambda(x)^{-u}$ ($\Lambda(x)^{+u}$) will be $\hat{\Lambda}(x)^{-u}$ ($\hat{\Lambda}(x)^{+u}$) indicating the first moment for the variable x of interest for firms i such that $\hat{\alpha}_i > \hat{\alpha}^*(u)$ ($\hat{\alpha}_i < \hat{\alpha}^*(1-u)$). In the present paper $u = 25$, however, we will maintain the more general u notation for the sake of consistency with Section 4.
The single-step algorithm proceeds as follows: 1) for each variable $x \in X_t$, compute $\tilde{\Lambda}(x)^{+u}$ and $\tilde{\Lambda}(x)^{-u}$, bootstrap draws of $\hat{\Lambda}(x)^{+u}$ and $\hat{\Lambda}(x)^{-u}$ respectively. We want to test the null hypothesis, H_0, that $\Lambda^u(x) = 0$, for $\Lambda^u(x) = [\Lambda(x)^{-u}, \Lambda(x)^{+u}]$. 2) Construct a bootstrap draw of the distribution of $(\hat{\Lambda}^{+u}(x) - \hat{\Lambda}^{-u}(x))$, $Z^u_\infty(x)$. The latter is obtained by exploiting the bootstrap version of $\Lambda^{+u}(x)$ and $\Lambda^{-u}(x)$, namely: $\tilde{Z}_\infty(x) = \sqrt{n}(\tilde{\Lambda}^u(x) - \hat{\Lambda}^u(x))$ where $\tilde{\Lambda}^u(x) = [\tilde{\Lambda}(x)^{-u}, \tilde{\Lambda}(x)^{+u}]$ (similarly, $\hat{\Lambda}^u(x) = [\hat{\Lambda}(x)^{-u}, \hat{\Lambda}(x)^{+u}]$). 3) Repeat steps 1) and 2) B times; 4) compute a bootstrap estimator of the variance of Z_∞ as $\hat{\Sigma}^u(x) = \frac{q^u_{0.75}(x) - q^u_{0.25}(x)}{z_{0.75} - z_{0.25}}$ being $q^u_p(x)$ the p^{th} sample quantile of $\tilde{Z}_\infty(x)$ and z_p the p^{th} quantile of a standard normal distribution. 5) Use the latter to construct the test statistic $\tilde{\tau}(X_t) = sup_{x \in X_t} |\tilde{Z}_\infty(x)| \cdot |\hat{\Sigma}^u(x)|^{-1/2}$. A p-value for the null H_0 that $\Lambda^u(x) = 0$ for all $x \in X_t$ of the realization of the estimated statistic, $sup_{x \in X_t} |\hat{\Lambda}^u(x)| \cdot |\hat{\Sigma}(x)|^{-1/2} = s$, is given by the average number of times that $\tilde{\tau}(X_t)$ is greater than s, where $s = \frac{\tilde{\beta}^m_{1,f}}{\hat{\Sigma}^u(x)}$. The $\tilde{}$ indicates simply that the $\beta^m_{1,f}$ has been projected to the bootstrap dimension.

the single-step joint inference procedure developed by Chernozhukov et al. (2018), which controls the family-wise error rate (FWER) via bootstrap-based p-values. This procedure ensures that the joint probability of incorrectly rejecting any true null hypothesis across all covariates remains bounded, thereby maintaining the overall validity of inference across the full set of comparisons. Technically, the validity of the single-step correction relies on the fact that the sorted effects and their induced classification structure are smooth functionals of the data-generating process. In particular, Chernozhukov et al. (2018) show that the sorting and classification operators are Hadamard differentiable, which allows the use of the functional delta method to derive the large-sample distribution of the estimated effects and their differences. This property justifies the application of bootstrap methods for obtaining joint confidence sets and p-values, even in the presence of nonlinear models and complex sorting rules. In our context, this implies that differences between covariates across classification groups are not tested in isolation, but rather as a joint hypothesis — thereby appropriately correcting for the multiplicity of comparisons and preserving valid inference.

Our inference procedure, in particular, is based on *bootstrap resampling combined with Sorted Effects and CA*, as developed by Chernozhukov et al. (2018) and adapted to settings without a contemporaneous control group. Specifically, we implement a *nonparametric empirical bootstrap* in which each sample is drawn with replacement from the original data using multinomial weights $\omega_1, \ldots, \omega_n$ with equal probability $1/n$[15], allows us to capture the sampling variability of the sorted effects and their induced classification structure. We employ $B = 100$ bootstrap replications, selected to balance computational feasibility with sufficient precision, especially given that bootstrapping is nested within a multi-stage machine learning framework involving model fitting and effect estimation across multiple folds and time splits. As anticipated, the procedure serves two core inferential objectives: constructing confidence intervals for CA and CADiff estimates and computing joint p-values to assess the statistical significance of heterogeneous effects across multiple covariates. In our setting, where the counterfactuals are generated from predictive models rather than observed untreated outcomes, a further layer of uncertainty is introduced. To accommodate this, the bootstrap replicates the entire estimation pipeline –model selection, prediction, and classification –treating the *Partial Effects (PE) as the true reference under the null hypothesis*. All classification and treatment effect heterogeneity measures are re-estimated in each bootstrap sample. Finally, the entire inference pipeline is integrated with k-fold cross-validation, preserving stability and validity in the estimation of sorted effects.

2.3 Comparison with Generic ML

Our approach in estimating the individual treatment effect and in performing the heterogeneity analysis is similar to the generic ML technique presented in Chernozhukov et al. (2023),

[15]The reader is referred to Algorithm 2.2 in our supplemental material available upon request.

which is adapted to a situation in which there is no available (contemporaneous) control group (i.e., it is difficult to identify ex-ante firms that are not affected by the shock).

We show that our empirical strategy is built on the same pillars as Chernozhukov et al. (2023), but applies them to a different setting. To simplify the exposition, we refer to Table 2 which provides a simplified representation of our empirical setting.

Chernozhukov et al. (2023) deal with an experimental empirical setting in which one can easily separate a treated group from a control group. In order to study the heterogeneity of the average treatment effect, the first step of Chernozhukov et al. (2023) is to split randomly the sample under analysis in an auxiliary (A) and a main sample (M) of approximately the same size. Then, they employ ML techniques to learn in A the function approximating the potential outcomes in the treatment and non-treatment scenarios, while M is used to make inferences on the key features of treatment effect heterogeneity. In other words, they estimate the function describing the outcome in case of treatment (no treatment) on the subset of treated (non-treated) firms contained in A. These two estimated functions are used to impute the two potential outcomes for each firm contained in the M sample (the difference represents the estimated individual treatment effects) and study the treatment effect heterogeneity estimated for these firms by using, inter alia, the Sorted Effects method (Chernozhukov et al., 2018). This procedure is designed in this way to avoid overfitting (i.e., doing learning and prediction using the same sample), and, starting from the random splitting, it is repeated many times in order to obtain many distributions of estimated treatment effects to which the Sorted Effects method is applied.

Our Setting			Chernozhukov (2020) Setting	
SUM		SAM	$A-M$ splitting	
$(X_{2018}, Y_{2019})_1$	$(X_{2019}, Y_{2020})_{11}$	$(X_{2019}, Y_{2020})_{11}$	$(X_{2019}, Y_{2020})_{11}$	A
$(X_{2018}, Y_{2019})_2$	$(X_{2019}, Y_{2020})_{12}$	$(X_{2019}, Y_{2020})_{12}$	$(X_{2019}, Y_{2020})_{12}$	
$(X_{2018}, Y_{2019})_3$	$(X_{2019}, Y_{2020})_{13}$	$(X_{2019}, Y_{2020})_{13}$	$(X_{2019}, Y_{2020})_{13}$	
$(X_{2018}, Y_{2019})_4$	$(X_{2019}, Y_{2020})_{14}$	$(X_{2019}, Y_{2020})_{14}$	$(X_{2019}, Y_{2020})_{14}$	
$(X_{2018}, Y_{2019})_5$	$(X_{2019}, Y_{2020})_{15}$	$(X_{2019}, Y_{2020})_{15}$	$(X_{2019}, Y_{2020})_{15}$	
$(X_{2018}, Y_{2019})_6$	$(X_{2019}, Y_{2020})_{16}$	$(X_{2019}, Y_{2020})_{16}$	$(X_{2019}, Y_{2020})_{16}$	M
$(X_{2018}, Y_{2019})_7$	$(X_{2019}, Y_{2020})_{17}$	$(X_{2019}, Y_{2020})_{17}$	$(X_{2019}, Y_{2020})_{17}$	
$(X_{2018}, Y_{2019})_8$	$(X_{2019}, Y_{2020})_{18}$	$(X_{2019}, Y_{2020})_{18}$	$(X_{2019}, Y_{2020})_{18}$	
$(X_{2018}, Y_{2019})_9$	$(X_{2019}, Y_{2020})_{19}$	$(X_{2019}, Y_{2020})_{19}$	$(X_{2019}, Y_{2020})_{19}$	
$(X_{2018}, Y_{2019})_{10}$	$(X_{2019}, Y_{2020})_{20}$	$(X_{2019}, Y_{2020})_{20}$	$(X_{2019}, Y_{2020})_{20}$	

Table 2: A simplified representation of our empirical setting in which we compare the methods used in the present paper to those described in Chernozhukov et al. (2023).

As an example, in Table 2, we represent 20 exporting firms observed in 2018 or in 2019. In the context of our setting, the strategy of Chernozhukov et al. (2018) would imply that the 2019 sample, for which we are interested in estimating the average treatment effect, should be divided in two, as shown in the last column of Table 2. However, in the COVID-19 scenario, one cannot easily separate treated and control units because COVID-19 imposes a (at least indirect) treatment over all units, hence preventing the possibility of discerning between treated and controls.[16] Moreover, with respect to Chernozhukov et al. (2023), in our empirical setting, we do not have the necessity to predict the outcome of controls in the case of "no treatment" because we are not interested in estimating the COVID-19 effect on 2018 exporters. Therefore, we do not have to split the controls observed in 2018 in two halves to avoid overfitting and this enables us to reconstruct a counterfactual outcome of no treatment for each 2019 exporter without incurring in overfitting problems. Therefore, in this paper for the SUM we use as an auxiliary sample all the Colombian exporters observed in 2018 (A) and as the main sample (M) all the Colombian exporters observed in 2019. For the SAM, we perform instead a K-Fold splitting in which, iteratively we select 80% of the firms in 2019-2020 as being part of A and the remaining 20% as being part of M. This is shown in the column "Our Setting (SAM)" of Table 2, where different A (and, accordingly

[16] Furthermore, even if we assume that during the first three months of the year there was no COVID-19 effect going on, and therefore we categorize as non-treated (treated) firms operating in those months (in the other remaining months), and we use the non treated firms in the auxiliary sample to learn, it would be problematic to use the learning outcome in case of no treatment during the first three months to predict the outcome in case of no treatment for the treated firms that are those in the last 8 months because of the strong seasonality effects we have. So the outcome during the first three months in case of no treatment would be very different from the outcome of the last months in case of no treatment just because of seasonality effects.

different M) groups are selected according to the different colors of the dashed circles. In this way we avoid overfitting problems and, at the same time, we exploit all the available data by being able to compare the predicted probabilities to export in the COVID-19 with those in the non-COVID-19 scenario for all the observed 2019 exporters.

Lemma D.1 of Chernozhukov et al. (2023) provides a theoretical foundation for conducting valid inference on key features of CATE. This lemma relies on several underlying assumptions. Most notably, Chernozhukov et al. (2023) assume a randomized controlled trial (RCT) setting. In our case, this assumption is not overly restrictive, as the estimated distribution of the propensity score in our sample lies within a narrow range of 0.498 to 0.511, closely approximating random assignment according to the distribution of observables (see Figure Appx.13 in Appendix H). Another possible limitation of our setting is the overlap of firms between $t-1$ and t, given by the panel nature of the dataset, which could introduce overfitting problems in our strategy. The original sample splitting procedure of Chernozhukov et al. (2023) instead is not affected by this problem, using different sets of firms in the A and M samples.

The SPE (Chernozhukov et al., 2018) offers formal inference guarantees. In particular, the bootstrap-based confidence bands are valid under mild regularity conditions. We verify these conditions in our setting following the guidance in Chernozhukov et al. (2018), which centers on the smoothness and non-degeneracy of the estimated treatment effects function, as shown in the Online Appendix. However, the SPE is a method originally designed for doing inference on CATEs obtained with parametric and semiparametric estimators, and not for machine learners. In Chernozhukov et al. (2023), Group Average Treatment Effects (GATES), which have been designed for studying CATE estimates obtained with ML methods, have the same role as SPE, as detailed in section 4.5. The main difference between SPE and GATES is that the former summarises the distribution of CATEs by estimating its percentiles, while the latter divides the support of estimated CATEs in bins by typically using quartiles or quintiles and estimates averages of the CATEs within these bins. Finally, the classification analysis (CA) employed in Chernozhukov et al. (2018) and in this paper is the same as the Classification Analysis (CLAN) presented in Chernozhukov et al. (2023).

Therefore, as a robustness check, in section 4.5 we will follow the procedure outlined by Chernozhukov et al. (2024) to apply the Chernozhukov et al. (2023)'s methodology to a research design characterized by the Conditional Independence Assumption and we compare the results obtained with those conveyed by SPE.

3 Data and Dependent Variable

This study focuses on the social and economic disruption caused by the COVID-19 pandemic and its effect on Colombian exporters. This global health crisis triggered by the COVID-19 pandemic served as a notable example of a large-scale economic shock that profoundly

impacted global trade, with the dynamics of exporters in Colombia being significantly affected. Applying our ML strategy to data collected from Colombian exporting companies during this period can provide key insights into how companies adapt and survive in the face of such widespread disruption. This provides an understanding of market resilience and firm survival dynamics in the context of global trade shocks. Therefore, our dependent variable $Y_{i,t}$ is a dummy variable that takes the value 1 if a firm i is an exporter at time t –given that it was an exporter in $t-1$– and the value 0 if the firm is not exporting at time t.[17] By grounding our research on a specific case study, we maintain its relevance to the specific scenario while preserving its potential for broader applications.

We use monthly export transaction data reported at the Colombian Customs Office (Dirección de Impuestos y Aduanas Nacionales, DIAN) for 2018, 2019, and 2020. For each transaction, we consider the exporter ID as the firm identifier; the date; a 10-digit Harmonized System code (HS) characterizing the product; the product origin within Colombia (department level); the means of transportation of the shipment; the country of destination; and, the free on board value of the export transaction in US dollars. This data set also contains information about the value and origin country from which a given exporter imports. We remove all transactions related to re-exports of products elaborated in other countries. As a result, we ended up with 386,132 customs reports in 2018 (7 741 firms), 402,140 in 2019 (7 831 firms), and 365,626 in 2020 (7 518 firms).

3.1 Control Variables

The selection of control variables is based on the determinants of firm entry and exit in foreign trade (see, e.g., Albornoz et al., 2012; Arkolakis et al., 2021). We classify products at the six-digit level of the HS code. We consider different features of exporters according to their monthly exports: the total export (and import) value, the number of products (NP), the number of export destinations (ND), the number of import origin countries (NO), the Herfindahl-Hirschman indexes at the product level (HH_p) and the destination level (HH_d), and a set of dummies for the destinations and origin countries and continents. We create a set of dummies according to the Colombian department from which the product comes, a set of dummies for the means of transportation used, and a set of dummies classifying the product HS-chapter and HS-section. Moreover, we build two sets of dummy variables indicating whether a firm has experience exporting in specific destinations and product sectors. We also account for the accumulated exporting (importing) experience by summing up the total value exported (imported) during the last twelve months. Furthermore, we create four *size* dummies classifying firms according to the quartiles of the firm-level distribution of the total monthly log-value of exports.

To measure the COVID-19 demand and supply shock, we use the information on government contention measures coming from Hale et al. (2020), which consists of four

[17] More precisely, $Y_{i,t} = 0$ if a firm is no longer active or is active but not exporting at time t.

indexes (ranging from 0 to 100) representing the strength of the measures taken by countries to contain the COVID-19 outbreak. The authors provide an economic index summarizing economic policies (E), a health index summarizing health policies (H), a government index describing the strictness of 'lockdown style' policies (G), and an overall government response index called stringency index (S). The value of these indexes ranges from 0 to 100.[18] We build two variables at the firm level for each of the four indexes, one at the export and one at the import side, by taking a weighted average of the country-level scores according to the proportion of the total monthly value of exports (imports) that a firm ships (source) in each country in 2019. We call these firm-level indexes for a firm i "Containment Index$_{i,j,z}$", with $j = \{E, H, G, S\}$ and $z = \{\text{Imp}, \text{Exp}\}$.[19]

Our final data set is composed of 1,975 covariates. They are presented in detail in Table Appx.1 of Appendix B.

4 Results

4.1 Selection of the machine learning algorithm

We evaluate and contrast the outcomes of several ML techniques against a benchmark logistic regression, aiming to identify the model with superior prediction performance, which is crucial for the consistency of our T-Learner estimator. The out-of-sample predictive efficacy of our empirical models is crucial, given our goal to reconstruct an unobserved counterfactual. The complexity of this task arises from its high dimensionality and complex interdependencies between firms and products from various sectors and export destinations. While an approach focusing on in-sample prediction accuracy might overfit, ML techniques optimally balance the bias-variance trade-off for out-of-sample predictions.[20]

We examine four distinct models: Logit, Logit-LASSO, Logit-Ridge, and Random Forest (RF). The traditional choice for binary dependent variables, Logit, serves as our baseline. Even though literature often shows ML techniques outperforming traditional models with numerous predictors, we have included Logit results for comparison. The main idea of Logit-LASSO is to mitigate overfitting by introducing a penalty term in the Logit log-likelihood function that forces the parameters associated with the less relevant predictors to be exactly zero. On the other hand, Logit-Ridge reduces the coefficients of less significant predictors without eliminating any of them, proving especially useful when many variables play an important role. The main idea behind Random Forest is the wisdom of crowds because it combines the predictions of many uncorrelated models (the trees) obtained by randomly re-sampling

[18] These indexes are released daily. We average this information at the monthly level.

[19] The value of the Containment Stringency Index Import for firms that are not importing corresponds to the value of the Containment Stringency Index for Colombia (as firms are sourcing all their inputs domestically).

[20] Hyperparameter tuning through cross-validation or other theory-driven methods is often critical in order to avoid overfitting.

observations and explanatory variables.[21] For Logit, Logit-Ridge, and Logit-LASSO models we include interactions between the *size* of the company and some of the main product characteristics, *industry*, *sector*, *means of transportation* as well as with *destination country* dummies. Notice that Random Forest uses the variables sequentially and, therefore, with a large enough number of trees, it is not necessary to explicitly introduce interactions as explanatory variables, i.e., the model automatically takes into account the interactions that are useful to accurately predict the outcome.[22] The prediction analysis is repeated for all months between January-December 2020. In Appendix E we perform a series of robustness tests for alternative ML methods (XGBoost and SVM) and panel cross-validation. Although panel cross-validation is more complicated, the results are largely consistent with those in the main text.

Table 3 shows the goodness of fit of the model's predictions using two widely used classification metrics: Root Mean Square Error (RMSE) and the Area Under the Receiver Operating Curve (AUC). The best value for the RMSE is 0, which indicates optimal accuracy with no fixed upper limit. The AUC reaches a value of 0.5 for random predictions and 1 when the outcomes are classified without error. Our preferred metric is the RMSE. The reason is twofold. First, our analysis focuses on estimating *predicted probabilities* of exporting, rather than producing binary classifications. The key object of interest is the probability of continuing to export in the future, conditional on covariates and treatment status, which is central to counterfactual analysis. In this setting, thresholding probabilities to assign binary labels (e.g., classifying 0.51 as 1 and 0.49 as 0) can lead to substantial misrepresentation of the underlying uncertainty, particularly in marginal cases.
Second, model performance in our context is evaluated based on how well the predicted probabilities approximate the true (unobserved) probabilities.
The AUC-ROC is provided as an alternative threshold-independent measure of the quality of the fit. The results are consistent when using different measures of goodness of fit.

The table's upper part displays the accuracy of predictions for the probability of exporting in 2019 based on 2018 exporter data, serving as an out-of-sample performance benchmark in a pre-COVID-19 context using cross-validation. Here, the Logit-LASSO and RF models arise as top performers. The table's middle section also shows the accuracy of models estimated using the exporters' characteristics in 2018 to explain their observed outcomes in 2019; however, these models are now tested using the set of exporters of 2019 and their observed outcomes in 2020. If the functions f_t^0, which represent the relationship between the explanatory variables and the outcome without the pandemic, are sufficiently similar for the

[21] Note that it is important to optimize (tune) the hyperparameters of the models for an accurate prediction. These are the hyperparameters we tuned in our models: Ridge; $[10^{-4}, 10^2]$, best 10^0; Logit-LASSO; $[10^{-4}, 10^2]$ best 10^1; RF, following Probst et al. (2019); n estimators: $[100, 200, 500]$, best 500; max features: ['sqrt', 'log2', None], best sqrt; max depth: $[5, 7, 10]$, best 7; max leaf nodes: $[3, 6]$, best 6; min samples split: $[2, 8]$, best 2.

[22] For more information about all the features included to build the SUM and SAM see Table Appx.1 in Online Appendix B.

year before the pandemic and for the year 2020 (f^0_{2019} and f^0_{2020}; see assumption (ii)), we expect the accuracy of $\hat{f}^0_{2019}$ to be similar in the first three months of 2020 (when there is likely no relevant COVID-19 effect in Colombia) as in the same months of 2019. Indeed, as expected, the accuracy of Logit-LASSO and RF in January, February and March remains unchanged compared to the accuracy found in the upper part of the table. However, in the middle part of the tables, a decrease in accuracy can be observed after April, highlighting the challenges of a model not trained on COVID-19 data when forecasting in an environment affected by COVID-19.

The models in the lower part of Table 3 are trained and tested with the universe of exporters in 2019 and their observed outcomes in 2020. We use these models to create the SAM forecasts. The accuracy of the predictions is very similar to that obtained with the SUM for 2019 and for the first three months of 2020. Our analysis is crucial to achieve accurate predictions because the unbiasedness of our treatment effect estimators depends on the quality of the (counterfactual) prediction accuracy. Both the SUM and the SAM show an acceptable level of accuracy when predictions are made with Logit-LASSO and Random Forest.

Table 3: Goodness of Fit for SUM and SAM in 2018/19 and 2019/20

	RMSE				AUC			
	Logit-LASSO	Logit-Ridge	Random Forest	Logit	Logit-LASSO	Logit-Ridge	Random Forest	Logit
Goodness of Fit for SUM in 2018/19								
Jan	0.40	0.45	0.41	0.64	0.73	0.53	0.73	0.59
Feb	0.41	0.45	0.41	0.64	0.70	0.50	0.71	0.58
Mar	0.41	0.44	0.41	0.65	0.70	0.56	0.71	0.57
Apr	0.40	0.43	0.40	0.63	0.73	0.59	0.73	0.60
May	0.40	0.44	0.41	0.64	0.72	0.52	0.71	0.59
Jun	0.40	0.45	0.41	0.64	0.71	0.50	0.72	0.59
Jul	0.40	0.45	0.40	0.66	0.73	0.50	0.73	0.55
Aug	0.41	0.45	0.40	0.64	0.70	0.51	0.72	0.58
Sep	0.41	0.45	0.40	0.64	0.72	0.50	0.71	0.58
Oct	0.40	0.44	0.41	0.64	0.73	0.58	0.74	0.58
Nov	0.41	0.45	0.41	0.64	0.71	0.51	0.72	0.57
Dec	0.41	0.45	0.41	0.64	0.70	0.50	0.71	0.58
Goodness of Fit for SUM in 2019/20								
Jan	0.41	0.45	0.41	0.75	0.72	0.53	0.72	0.49
Feb	0.41	0.45	0.42	0.64	0.69	0.50	0.69	0.56
Mar	0.40	0.44	0.41	0.63	0.72	0.54	0.73	0.59
Apr	0.48	0.50	0.49	0.70	0.67	0.56	0.66	0.51
May	0.46	0.48	0.46	0.63	0.69	0.51	0.69	0.60
Jun	0.43	0.47	0.44	0.63	0.68	0.50	0.68	0.59
Jul	0.42	0.46	0.43	0.63	0.70	0.50	0.69	0.59
Aug	0.42	0.45	0.43	0.63	0.68	0.51	0.69	0.58
Sep	0.42	0.45	0.42	0.63	0.69	0.50	0.70	0.59
Oct	0.42	0.45	0.43	0.63	0.71	0.59	0.70	0.60
Nov	0.41	0.45	0.41	0.63	0.71	0.51	0.71	0.59
Dec	0.42	0.46	0.42	0.63	0.69	0.50	0.69	0.58
Goodness of Fit for SAM in 2019/20								
Jan	0.41	0.45	0.41	0.71	0.73	0.58	0.74	0.50
Feb	0.41	0.46	0.42	0.70	0.70	0.50	0.70	0.49
Mar	0.40	0.46	0.40	0.71	0.73	0.50	0.73	0.50
Apr	0.42	0.47	0.42	0.69	0.74	0.66	0.73	0.52
May	0.41	0.46	0.41	0.71	0.76	0.74	0.77	0.50
Jun	0.42	0.46	0.42	0.72	0.73	0.69	0.73	0.48
Jul	0.41	0.45	0.42	0.69	0.73	0.63	0.72	0.51
Aug	0.41	0.46	0.42	0.69	0.72	0.50	0.72	0.53
Sep	0.42	0.47	0.42	0.67	0.71	0.50	0.70	0.55
Oct	0.42	0.46	0.42	0.70	0.72	0.50	0.71	0.52
Nov	0.41	0.45	0.41	0.71	0.72	0.52	0.72	0.49
Dec	0.41	0.45	0.42	0.70	0.71	0.51	0.70	0.51

Notes: Results are obtained based on a 5-fold cross-validation strategy. RMSE and AUC are averaged across folds.

4.2 Evaluation of the COVID-19 effect

Both Logit-LASSO and Random Forest reach high accuracy levels in the export status prediction. As explained in section 2, the predicted probabilities are used to estimate the average monthly effect of the COVID-19 shock as the (monthly) average of $\hat{\alpha}_i$ (the difference between the firm-level probabilities of success predicted by the SUM and the SAM.), $\bar{\hat{\alpha}}$. They are presented in Figure 1.

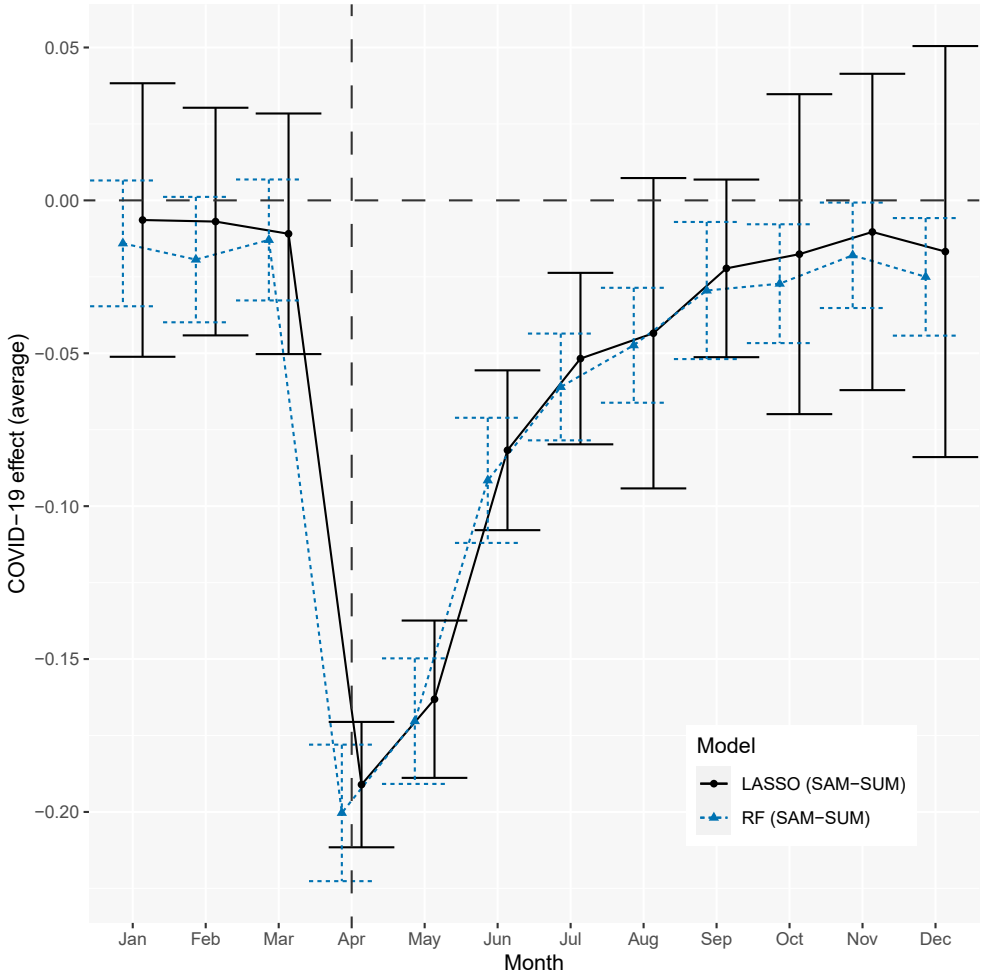


Figure 1: Average Individual Treatment Effect, by months, comparing Logit-LASSO and RF. Standard errors were obtained with 100 bootstrap replications. Confidence intervals for a 5% significance level.

Given the presumption that firms suffered a negligible COVID-19 shock impact during the initial three months of 2020, the treatment effect estimates for this period can be viewed as a placebo test, reminiscent of the in-time placebo test routinely employed in Synthetic Control Methods (Abadie et al., 2015). Detecting a significant COVID-19 effect in the months preceding the actual economic shock would suggest that our model mechanically estimates a COVID-19 effect even in the absence of the stated shock. We conduct these placebo studies also conditioning on exogenous firms' characteristics observed in 2019 by estimating COVID-19 effects for selected subsamples of firms according to such characteristics. We

interpret these additional placebo studies as a robustness check on our results on treatment effect heterogeneity.

As shown in Figure 1, the probabilities obtained from the SUM and the SAM are almost identical on average for January, February, and March. This result is reassuring since only from March 25, 2020, the Colombian government implemented a complete and mandatory lockdown.[23] More in general, we can conclude that our identification strategy is not mechanically recovering COVID-19 effects for a period with low incidence in Colombia and in the rest of the world. We find that the peak of the COVID-19 effect is in April 2020, when we estimate an average difference between the predicted probabilities of exporting of nearly 20 percentage points. In the following months, the estimated average effect declines with time.

The results indicate that both Logit-LASSO and RF models yield comparable performances.[24] Given their good performance and considering that Logit models are frequently used in similar contexts, we opt for Logit-LASSO. It aligns with the conventional approaches and offers greater interpretability as an extension of the traditional model.[25]

The results obtained using the AIPW-Double ML estimator (with 5-fold cross-fitting and nuisance parameters estimated with Generalized Random Forest), which are shown in Figure Appx.12 in the Appendix, are equivalent. It is also interesting to notice that estimated ATT and $ATE_{\{t_s,t_s-1\}}$ are practically the same. This happens because the distribution of the explanatory variables is exactly equal between the treated and the control group, as it is shown in Figure Appx.13 that reports the results of the estimated propensity score for the two groups.

Figure 2 reports the estimated CATE_z by industry, that is the Conditional Average Treatment Effect for those units belonging to different industries. It shows evidence of substantial variations in the quarterly estimated average individual treatment effect by industry. On the one hand, during the first, third, and fourth quarters of 2020, there is no evidence supporting the existence of sectoral heterogeneity in the COVID-19 effect, and the COVID-19 shock is economically and statistically insignificant. Therefore, concentrating on the results for the first quarter, we are able to reject the existence of an effect even within sectors.[26] On the other hand, during the second quarter of 2020, Colombian exporters belonging to almost every industry are found to significantly reduce their probability of surviving in the international markets. This decline is particularly pronounced in industries such as Textiles, Footwear, and Jewelry. However, industries like Food Preparations and

[23] See Appendix A for a detailed description of the measures taken in Colombia in the midst of the COVID-19 crisis.

[24] This is somewhat expected since in Künzel et al. (2019) meta-learners are said to provide better outcomes with generalizable ML-algorithms that perform well for a large variety of data sets.

[25] Non-reported results using RF are equivalent and available upon request.

[26] We have conducted other similar placebo studies conditioning on other variables (e.g., the main destination of exports, the main origin of imports, via (air, land, sea), industry, exported value, imported value) and in all the considered subsamples we do not estimate any significant effect of COVID-19.

Vegetables saw minimal changes in their survival probabilities due to the COVID-19 shock.

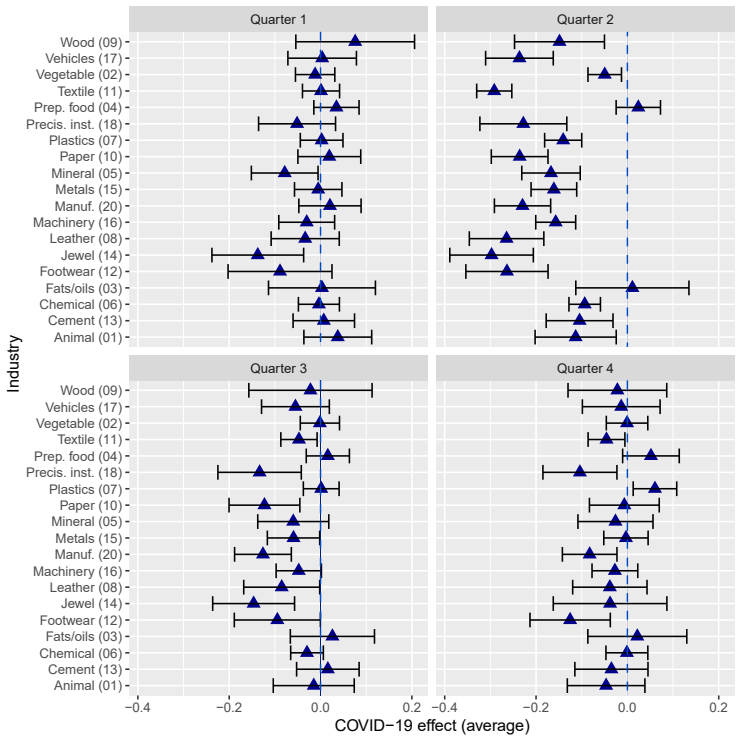


Figure 2: The quarterly mean difference in the predicted probability of success (SAM vs. SUM) by industry, using the Logit-LASSO predictions. Standard errors were obtained with 100 bootstrap replications. Confidence intervals for a 5% significance level.

4.3 Heterogeneity of the COVID-19 effect on Colombian exporters

In this section, we investigate the determinant of possible treatment effect heterogeneity. Figures 3 and 4 show the estimated Sorted Partial Effects (SPE) and Average Partial Effects (APE), which are obtained as explained in section 2 by month and aggregating all the months, respectively. The two figures also report the 95% confidence intervals with blue bands for SPE and black dashed lines for APE.

Significant treatment effect heterogeneity is observed for April and May, with June showing a milder effect. The statistically significant (negative) estimated values of $\alpha^*(u)$ are primarily confined to the distribution's left tail. However, from July onwards, the confidence intervals of the SPEs overlap with those of the APEs, indicating an absence of treatment effect heterogeneity. Interestingly, in the pre-pandemic months, the SPEs closely aligned with the APEs (estimated to be zero). This demonstrates that individual placebo treatment effects are not statistically significant throughout the distribution, not just on average, reinforcing the robustness of our methodology across the entire distribution of treatment effects.

To identify the determinants of treatment effect heterogeneity, we examine the difference in means ($CADiff$) of firm characteristics between the most and least affected groups in

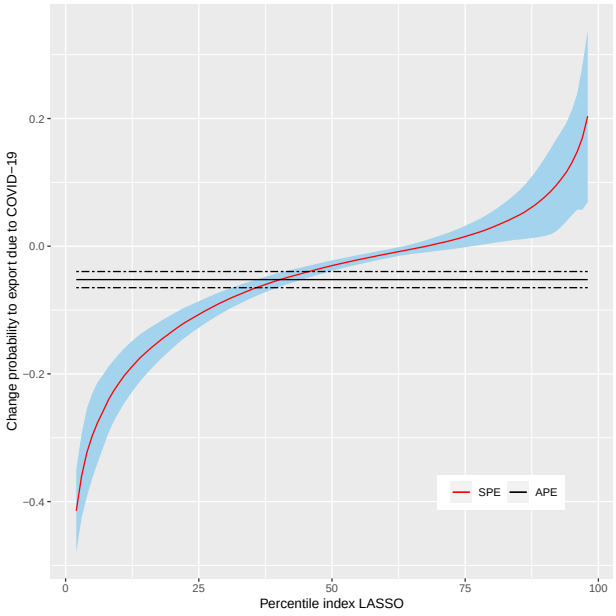


Figure 3: Annual Sorted Partial Effects (SPE) and Average Partial Effects (APE) of COVID-19 on Colombian firm export's status. The treatment effect is calculated as a difference between SAM and SUM predictions. Standard errors were obtained with 100 bootstrap replications. Confidence intervals for a 5% significance level.

Table 4. These groups are defined by whether their estimated α_i is lower than $\alpha^*(25th)$ or greater than $\alpha^*(75th)$, respectively. Therefore, we compute the raw difference in the means of the covariates between the most and the least affected firms by regressing the variables of interest on a constant and a dummy $q = \mathbf{1}_{\{\alpha_i \leq \alpha^*(25th)\}}$ for the observations for which estimated $\alpha_i \leq \alpha^*(25th)$ or $\alpha_i \geq \alpha^*(75th)$. Then, we also provide the difference in adjusted means once we have controlled for the firm sector and month of the year. Controlling for sector and month allows us to perform a *ceteris paribus* analysis, i.e., to dig into the effects of COVID-19 within specific sectors and specific months.

Table 4 is divided into 3 columns according to the control variables included in the regressions: in the first column, we show the unconditional average difference in the firms' characteristics between the most and least affected firms; in the second column, we control for the firm sector; and, in the third column, we control for firm sector and month of observation. The firm characteristics that we consider to explore the sources of COVID-19 treatment effect heterogeneity among Colombian exporters are observed in 2019 (the year before receiving the treatment). First, we check whether the estimated individual treatment effect (TE) differs between the firms contained in the two groups by using the TE as the dependent variable. We then move to firm-sector specific characteristics. In particular, the first set of firm characteristics that we use as dependent variables are dummies indicating the industry where the exporters operate.[27] We also investigate the *CADiff* for the means

[27] We aggregate the 22 industries defined in the main analysis as follows. "Agriculture" contains Animals (01), Vegetables (02), Fats/oils (03), and Prepared Foodstuffs (04). "Chemicals" includes Chemical (06), and Plastics (07). "Manufacturing" contains Machinery (16), Vehicles (17), and Manufactured (20). "Metals"

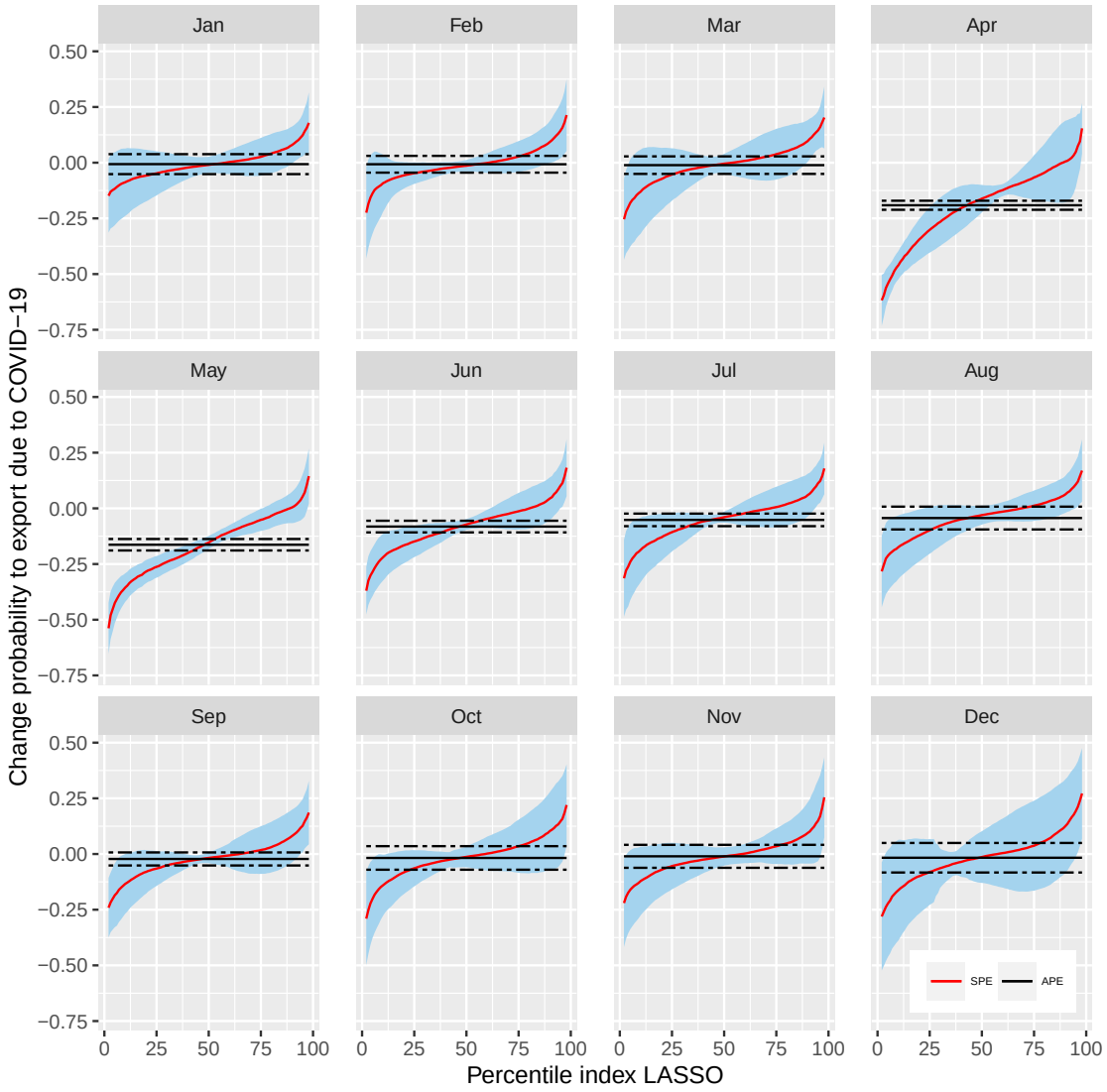


Figure 4: Monthly Sorted Partial Effects (SPE) and Average Partial Effects (APE) of COVID-19 on Colombian firm export status. The treatment effect is calculated as a difference between SAM and SUM predictions. Standard errors were obtained with 100 bootstrap replications. Confidence intervals for a 5% significance level.

of transportation and the months when firms operate. Moreover, to account for the role of diversification patterns, we also consider as dependent variables the number of export destinations (ND), import origins (NO), and products (NP) exported. The weighted Containment Stringency Index that exporters face when exporting (importing) allows us to study to what extent treatment effect heterogeneity depends on these firm-specific measures of exposure to COVID-19 through their activities on international markets. A traditional continuous-DID strategy would have used these intensity variables as treatment variables, assuming that any COVID effect would emanate through them. Finally, including the total

aggregates Mineral (05), Cement (13), Jewelries (14), and Metals (15). "Special" includes Precision Instruments (18), Arms (19), Art (21), and Special (22). "Textile" contains Leather (08), Textile (11), and Footwear (12). Finally, "Wood" aggregates Wood (09), and Paper (10). See Table Appx.2 in the Online Appendix for the complete industry names.

Table 4: Estimated differences in means of the estimated treatment effect and other covariates between the group of more affected and the group of less affected firms ($CADiff$) applying the classification analysis to the $SAM - SUM$ estimates

Outcome variable	(1)	(2)	(3)
TE	-0.3130^{***}	-0.3060^{***}	-0.2790^{**}
Agriculture	-0.1940		
Chemicals	-0.0057		
Manufacturing	-0.0092		
Metals	0.0134		
Special	0.0056^{***}		
Textile	0.1600^{***}		
Wood	0.0292^{***}		
Air	0.2030^{*}	0.1680^{***}	0.2040^{***}
Land	0.0340	0.0249	0.0170
Sea	-0.2360^{***}	-0.1920^{***}	-0.2200^{***}
Jan	-0.0738	-0.0766^{***}	
Feb	-0.0710	-0.0768^{***}	
Mar	-0.0751	-0.0773^{***}	
Apr	0.1860^{***}	0.1950^{***}	
May	0.1770^{***}	0.1820^{***}	
Jun	0.0754	0.0784^{***}	
Jul	0.0132	0.0159	
Aug	0.0021	0.0008	
Sep	-0.0412^{***}	-0.0406^{**}	
Oct	-0.0604^{***}	-0.0609^{**}	
Nov	-0.0723^{***}	-0.0763^{**}	
Dec	-0.0557	-0.0621^{**}	
Number of export destinations (ND)	-0.1990	-0.1640	-0.2480
Number of import origins (NO)	-1.7470	-1.9820^{***}	-2.4440^{**}
Number of exported products (NP)	0.2400	-0.2570	-0.3440
Containment Index Stringency Export	19.3600^{***}	19.5100^{***}	7.1800^{*}
Containment Index Stringency Import	19.1100^{***}	20.8000^{***}	7.2490^{***}
Value Exported (log)	-0.5110^{***}	-0.4490	-0.5700^{*}
Value Imported (log)	-1.8160^{***}	-2.2020^{***}	-2.6860^{***}
Deviation from sectoral mean		✓	✓
Deviation from monthly mean			✓

Notes: column (1) does not include sector or month variables in the regression; column (2) includes sectors in the regression; and, column (3) includes both the sector and month variables. *** means significant at 1%, ** at 5%, and * at 10%. Standard errors are obtained by bootstrapping the whole estimation process, and joint p-values are adjusted to consider the simultaneous testing of all variables.

value exported (imported) by firms –expressed in logarithm– among the variables for which the *CADiff* is computed highlights the difference in the quantities sold (purchased) by most and least affected companies. A discussion of the main findings follows.

Considering the estimated individual treatment effects (TE) as a dependent variable, we find a negative and significant difference between most and least affected firms independently of the set of controls employed. These results show that the most affected exporters–i.e., those located in the first SPE quartile distribution–experienced a decrease in the probabilities of exporting between 27.9 and 31.3 percentage points lower than the one experienced by the least affected firms–i.e., those located in the last SPE quartile.

We found significant differences among firms when examining how different aggregate

sectors are affected. For instance, we detect that the share of textile firms among the most affected 2019 exporters is 16 percentage points higher with respect to the one estimated for the group of the least affected firms. Likewise, there is a difference of 2.9 percentage points in the presence of wood exporters between the most and least affected groups.

We also detect the existence of treatment effect heterogeneity associated with the means of transportation used by exporters in 2019. On the one hand, there are 16.8 to 20.4 percentage points more exporters using air transportation among the most affected than among the least affected firms. However, there are 19.2 to 23.6 percentage points fewer Colombian exporters using the sea for shipping among the most impacted firms compared to the least affected ones (Nitsch, 2022).

Looking at the treatment effect heterogeneity associated with months, the first pattern we notice is that only the months from April to August have a positive estimated parameter. However, only April and May estimated differences are statistically significant. There are 18.6 to 19.5 percentage points (17.7 to 18.2) more firms in April (May) among the most affected than among the least affected firms. From September to November, the coefficients become negative and significant, indicating the beginning stages of recovery.

To evaluate how ex-ante exporter diversification affects the COVID-19 effect, we explore the estimated parameters associated with ND, NO, and NP. We want to investigate whether Colombian exporters' supply chain diversification and export destination diversification help mitigate the COVID-shock. We do not find compelling evidence that ex-ante diversification helps to face a shock of this kind, as we can evince from the estimated parameters associated with ND, NP, and, in the first column, to NO. Following the reasoning of Lafrogne-Joussier et al. (2022), which exploits the COVID-19 crisis to study the export consequences of a country-specific supply-side shock by concentrating on the differential import exposure of French firms to the Chinese early lockdown, one possible explanation is that firms cannot substitute away the partner (or the product) under COVID-19. Another possible explanation, which they offer, is that exporters that do not diversify ex-ante can benefit from some form of ex-post diversification. However, when they restrict the analysis to homogeneous inputs, Lafrogne-Joussier et al. (2022) find weak evidence of a larger COVID-19 effect for firms with non-diversified inputs. They restrict the sample to homogeneous inputs because they want to analyze the COVID-19 effect among inputs expected to be substituted. Similarly, once we control for the sector and, therefore, inter alia, for the fact that some sector has relatively more diversification potential, the negative estimated difference turns statistically significant. Indeed, within sectors, the most affected Colombian exporters tend to import from 1.98 fewer countries in 2019 than the least affected firms. The economic size of this estimate is large as approximately 60 per cent of Colombian exporters are not integrated into global value chains (they do not import), and the mean of NO is approximately 4.16 origins.

The $CADiff$ estimated when using the Export (Import) Containment Stringency Index as dependent variables provides insightful hints on the difficulties of Colombian firms

in exporting (importing) to (from) countries adopting severe stringency measures. In particular, the most affected Colombian exporters face, on average, a higher Export (Import) Containment Stringency Index than those faced by least affected firms by 7.18 to 19.51 (7.25 to 20.80) points, depending on the column in the table.[28]

Finally, the least affected firms exported (imported) 156.7% to 176.83% (614.7% to 1467.3%) more value in 2019 than the most affected firms. As expected, Colombian exporters trading in larger volumes (in value) are more resilient under a COVID-19 scenario. As with diversification, the comparison of the export and import side reinforces the idea that having more experience in sourcing inputs from abroad decreases the strength of the shock.

Our results not only show the uneven impact of the COVID-19 crisis across different covariates, but also highlight the potential of our methodology as a diagnostic tool for targeted policy interventions. By identifying the most affected firms and sectors, our framework can support the allocation of scarce public resources and the design of sector and firm-specific support programs aimed at improving the resilience of the most vulnerable exporters.

4.4 Estimations based on $Y - SUM$

In this paragraph, following Fabra et al. (2022) and Cerqua and Letta (2020), we use the estimators based on Eq. (13). These estimators capture the differences between the observed outcome, Y (binary variable accounting for the success of a Colombian exporter in 2020), and its counterfactual predictions (SUM). As shown in Figure 5, when the interest lies in estimating the average treatment effects (by months in this case), the results based on $Y - SUM$ do not differ from those obtained by using $SUM - SAM$. We obtain similar results for the two methodologies also in terms of conditional treatment effects based on subgroups defined on firm characteristics (e.g., by industry or main export destination country).

[28] Remember that the Index ranges from 0 to 100.

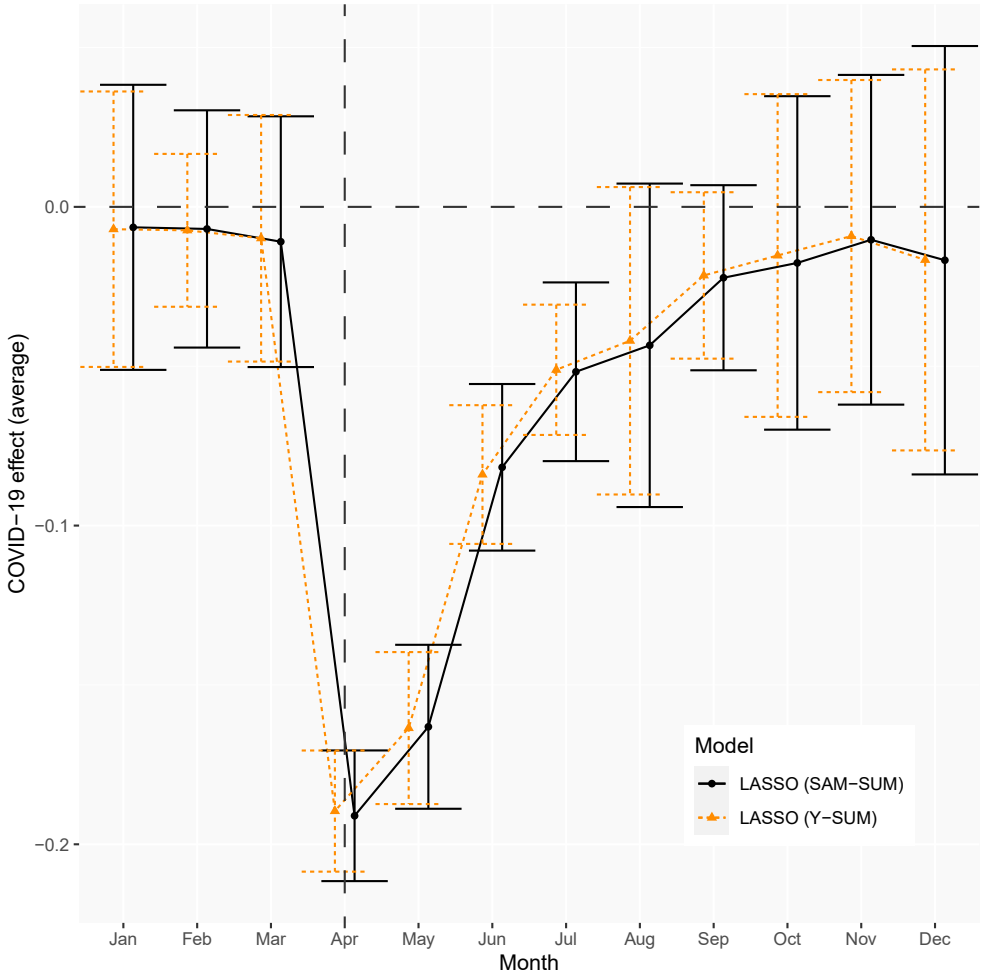


Figure 5: Mean difference in the predicted probability of success (SAM vs. SUM / Y vs. SUM) by month, using Logit-LASSO predictions and (SAM vs. SUM). Standard errors were obtained with 100 bootstrap replications. Confidence intervals for a 5% significance level.

The fact that the two estimators consistently find zero estimated effects for all 2019 exporters (and for subgroups based on the values of individual observables) during the first quarter suggests that the estimation error of both SUM and SAM, $\mathcal{E}^0$ and $\mathcal{E}^1$ respectively, goes to zero when we average the individual treatment effects across the whole distribution of 2019 exporters (or in subgroups defined by one of the possible dimension of treatment effect heterogeneity defined by observables; e.g., by industry or main export destination country).

However, since our goal is to identify the main dimensions of treatment effect heterogeneity by classifying units with the highest and lowest estimated treatment effects, we need also to evaluate how well these alternative estimation strategies perform in identifying treatment effects at the extremes of the distribution of treatment effects. Figure 6 shows the average of the estimated treatment effects obtained with the two estimators for the observations whose estimated treatment effects (by using $Y - SUM$) are contained in intervals defined by two consecutive values of the estimated percentiles of $Y - SUM$. On the one hand, the estimator based on $Y - SUM$ is also identifying significant treatment effect heterogeneity in the first quarter, suggesting that the distribution's estimation error, $\mathcal{E}^0$, is not zero on

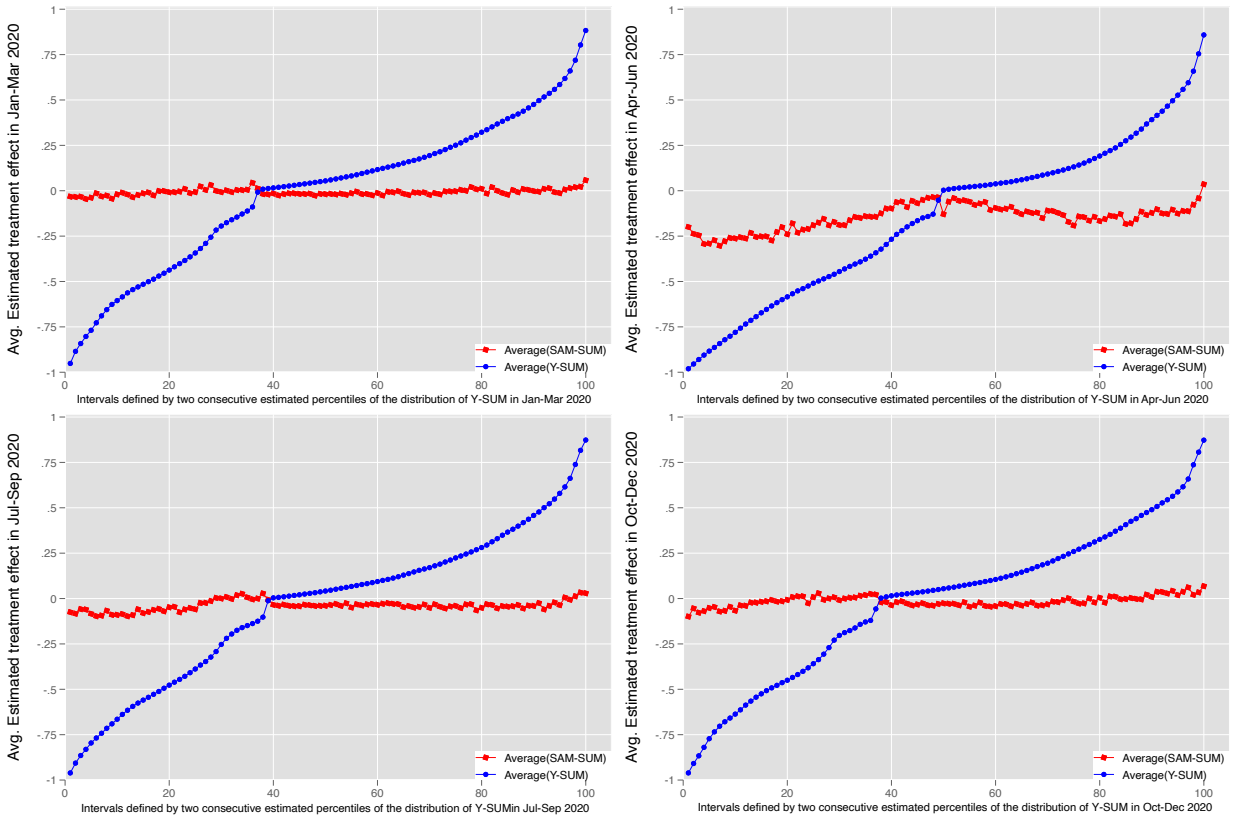


Figure 6: Estimated average treatment effects ($SAM - SUM$, red line, and $Y - SUM$, blue line) by quarter for observations contained in intervals defined by the estimated percentiles of $Y - SUM$.

average in the tails. Moreover, the shape of the $Y - SUM$ curve is similar across quarters, suggesting that this estimation method will be prone to misclassifying units when using the Sorted Effects strategy suggested above. On the other hand, in the first quarter, the shape of the $SAM - SUM$ curve is flat, showing a constant average estimated effect that is zero along the whole distribution of the $Y - SUM$ estimated effects, suggesting that by using the SAM we are able to wash out the estimation error of the SUM because $\mathcal{E}^1 = \mathcal{E}^0$.

This behavior of the estimators based on $SAM - SUM$ is consistent with the results shown in Figure 4 for the Sorted Effects analysis. Figure 7 shows that the intuition on the inadequacy of the $Y - SUM$-based estimators to identify treatment effects on the tail of the distribution is also confirmed by the Sorted Effects analysis based on this estimation strategy. When using the $Y - SUM$ individual level estimates to feed the SPE methodology, we find economically and statistically significant effects of the COVID-19 shock all along the percentile distribution in the first quarter. While it is true that, on average, $\mathcal{E}^0$ tends to be zero across all observations, these findings suggest that this is not true when we focus on specific segments of the treatment effect distribution, particularly in the tails.

Table 5 presents the classification analysis results on the sources of treatment effect heterogeneity when the $CADiff$ is estimated using the $(Y - SUM)$ approach. For all the firm characteristics we examined, we found no statistically significant difference between the most and least affected groups. This is consistent with the inability of the $Y - SUM$

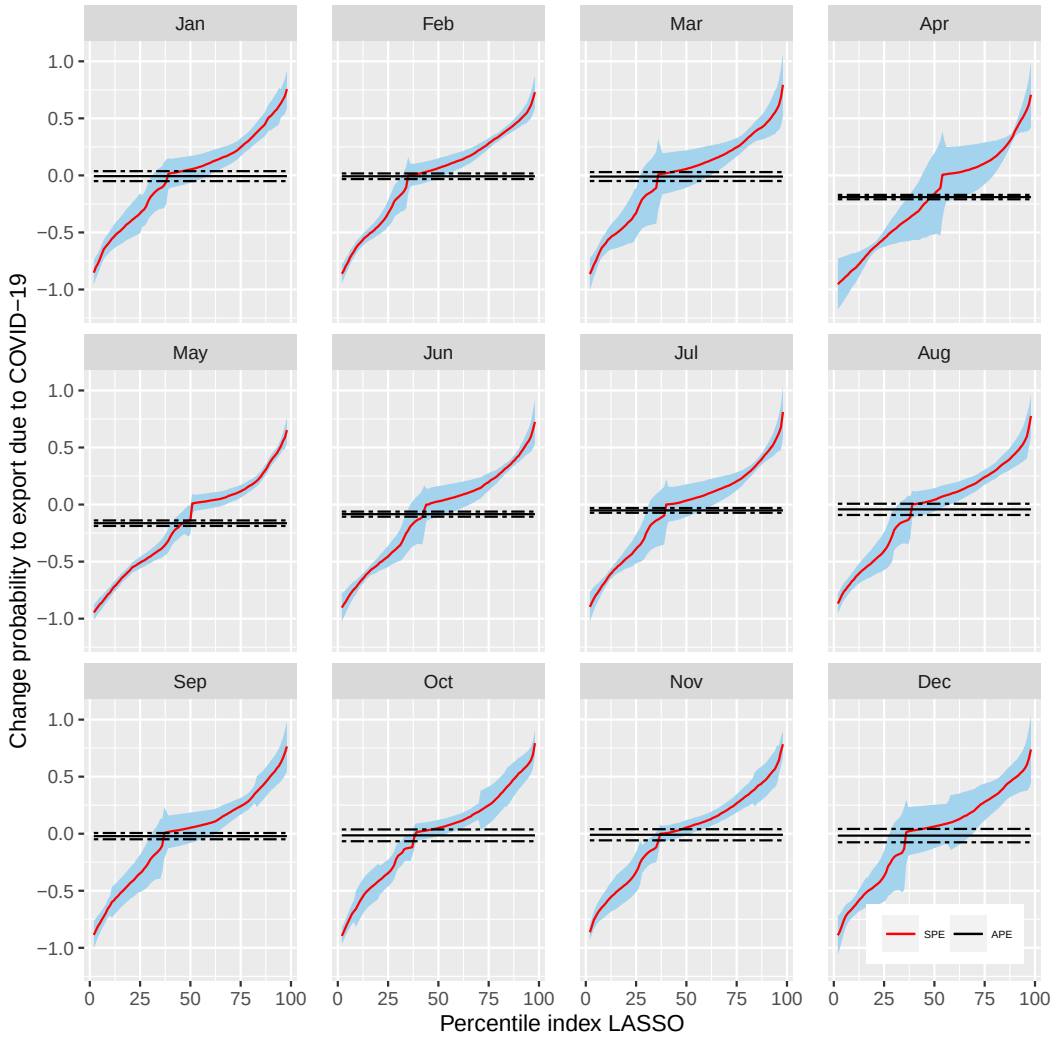


Figure 7: Monthly Sorted Partial Effects (SPE) and Average Partial Effects (APE) of COVID-19 on Colombian firm export's status. TE is calculated as a difference between the observed outcome (Y) and SUM predictions. Standard errors were obtained with 100 bootstrap replications. Confidence intervals for a 5% significance level.

approach to consistently estimate treatment effects in the tails of the α's distribution and, consequently, to identify the groups of the most affected and the least affected firms. In other words, such groups will be contaminated by the inclusion of firms wrongly classified due to the estimation error $\mathcal{E}^0$.

4.5 Validation of the CATE models

As it is well established in the causal inference literature, the Conditional Average Treatment Effect (CATE) coincides with both the Conditional Average Treatment Effect on the Treated (CATT) and the Conditional Average Treatment Effect on the Untreated (CATU) when the conditioning set includes all explanatory variables (those satisfying the conditional independence assumption). This is shown in Figure 8, where we represent the estimated

heterogeneous treatment effects for January obtained by using the T-Learner for $CATE_{\{t_s-1,t_s\}}$, $CATU$ and $CATT$. The latter coincides with our estimator $\hat{\alpha}_i$. Finally, we also report the CATT obtained with $\hat{\hat{\alpha}}_i$.[29]

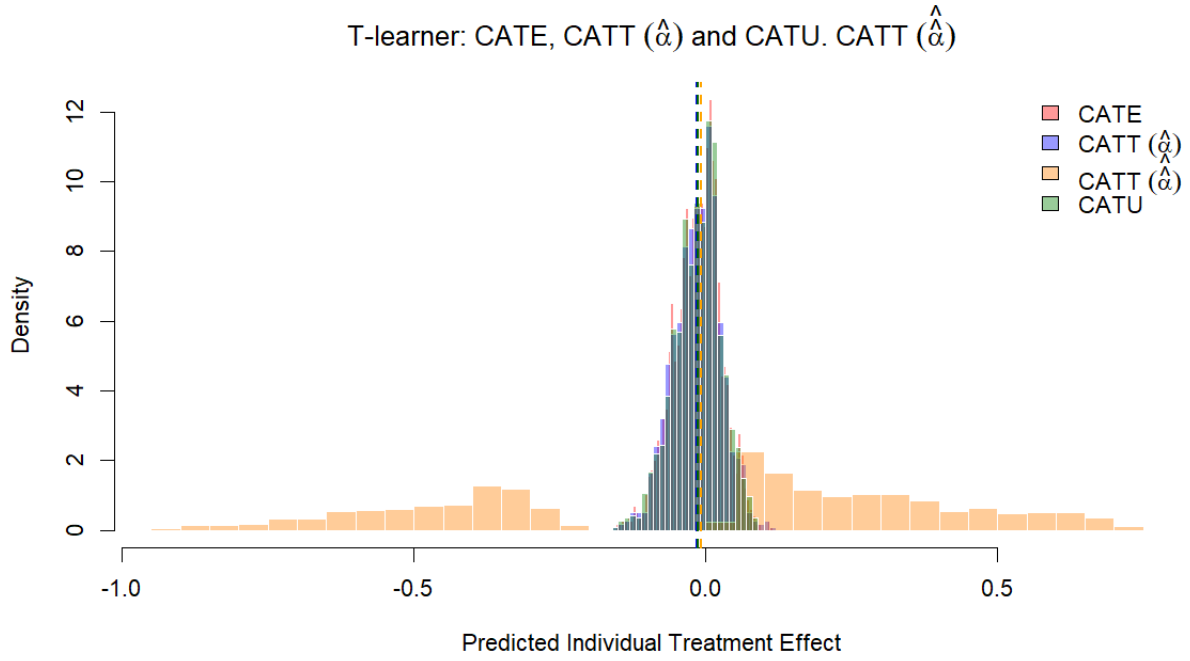


Figure 8: Heterogeneous treatment effects estimated in the A sample and predicted in the M sample for the month of January.

For the four estimators, we use Random Forest to estimate the potential outcomes. In our context, the distribution of observable characteristics is identical for firms that were observed exporting in 2018 (control group) and those exporting in 2019 (treated group). In Figure 8, this implies that the density of the $CATT$, $CATU$ and $CATE_{\{t_s-1,t_s\}}$ estimated with the T-Learner are the same. In this paragraph, we will compare results obtained with the T-Learner and other widely used estimators focusing on the $CATE_{\{t_s-1,t_s\}}$.

The primary aim of this section, which can be interpreted as a robustness check of the previous results, is to follow the procedure described by Chernozhukov et al. (2024) to apply the methodology of Chernozhukov et al. (2023) to a research design characterized by the Conditional Independence Assumption. We will compare the results obtained with SPE (and CA) with those obtained here by using GATES and CLAN. Finally, we will also estimate the Best Linear Predictor (BLP) of the CATE, a methodology to statistically test if a CATE estimator captures any statistically significant indication of treatment effect heterogeneity.

As an additional robustness check, we will also compare the results obtained with our T-Learner with those obtained by using other commonly adopted methods in the literature of causal inference (Chernozhukov et al., 2024), namely: the S-Learner, the R-Learner, the

[29] Notice that we have exactly the same result obtained in Figure 7 for January: the estimators based on Y–SUM produce a biased estimated distribution of the conditional treatment effects.

DR-Learner and the Generalized Random Forest (that we also call Causal Forest). An overview of the adopted learners can be found in Appendix F.

Throughout the section, we will randomly divide the observations in an auxiliary (A), consisting in 60% of observations, and a main (M) sample as recommended by Chernozhukov et al. (2023).[30] The models used to identify CATE are estimated using A and are tested in M. Chernozhukov et al. (2023) show that under mild regularity conditions the estimated parameters are normally distributed, conditional on the random sample split on A. To take into account the uncertainty originating from the splitting procedure, they also propose a multisplitting inference strategy in which the researcher repeat the estimation using different random division in splits and report the median of those estimates, along with the medians of the corresponding confidence intervals an p-values.[31]

Following Chernozhukov et al. (2023), we start by statistically testing the capacity of our estimators to capture treatment effect heterogeneity by using the BLP. As previously mentioned, Chernozhukov et al. (2024) propose a strategy to apply their methodology in a non-experimental setting. This approach relies on the pseudo-outcome—computed on the main sample M—of the AIPW-DML estimator (using cross-fitting with 5-Folds and Generalized Random Forest), which is reported again below:

$$\overset{ATE}{\{t_s,t_s-1\}} = \underbrace{\hat{Y}^1_{\{t_s,t_s-1\}} - \hat{Y}^0_{\{t_s,t_s-1\}}}_{\text{outcome predictions}} + \underbrace{\frac{D_{\{t_s,t_s-1\}}(Y_{\{t_s,t_s-1\}} - \hat{Y}^1_{\{t_s,t_s-1\}})}{\hat{e}(X_{\{t_s-1,t_s-2\}})} - \frac{(1-D_{\{t_s,t_s-1\}})(Y_{\{t_s,t_s-1\}} - \hat{Y}^0_{\{t_s,t_s-1\}})}{1-\hat{e}(X_{\{t_s-1,t_s-2\}})}}_{\text{weighted residuals}} \quad (17)$$

where, as a reminder, $\hat{e}(\cdot)$ represents the propensity score.

If the CATE model $\hat{\Delta}(X_{\{t_s-1,t_s-2\}})$ is well-specified, then the best linear predictor of the true CATE using the variables $(1, \hat{\Delta}(X_{\{t_s-1,t_s-2\}}))$ should yield a statistically significant coefficient on $\hat{\Delta}(X_{\{t_s-1,t_s-2\}})$. Given that $\mathbb{E}[\tilde{Y}^{ATE}|X] = \Delta(X)$, the BLP we estimate takes the following form:

$$(\beta_1, \beta_2) = \arg\min_{b_1,b_2} \mathbb{E}\left[\left(\tilde{Y}^{ATE}_{\{t_s-1,t_s\}} - b_1 - b_2(\hat{\Delta}(X_{\{t_s-1,t_s-2\}}) - \mathbb{E}[\hat{\Delta}(X_{\{t_s-1,t_s-2\}})])\right)^2\right]$$

where $\hat{\Delta}(X_{\{t_s-1,t_s-2\}})$ is the CATE estimated in A and predicted in M.

Under regularity conditions, indeed, the coefficient β_2 converges in population to

$$\beta_2 = \frac{\text{Cov}(\Delta(X_{\{t_s-1,t_s-2\}}), \hat{\Delta}(X_{\{t_s-1,t_s-2\}}))}{\text{Var}(\hat{\Delta}(X_{\{t_s-1,t_s-2\}}))}$$

In an ideal case where $\hat{\Delta}(X_{\{t_s-1,t_s-2\}})$ is perfectly aligned with the true CATE, we expect $\beta_2 = 1$. Hence, a statistically significant and positive β_2 provides evidence that the CATE model captures heterogeneous treatment effects.

For each month, Table 6 presents the median of the estimated β_2 coefficients of the

[30] Chernozhukov et al. (2023) require that the auxiliary sample be larger than the main sample.

[31] The results of this section using a single split are also available upon request.

CATE across 100 (A, M) splits with the relative median p-value for all the months and the mentioned CATE estimators (we also provide the R^2 of the regression).[32] The results show that for the months of April and May, all learners achieve highly significant β_2 coefficients (p-values ≤ 0.001), with magnitudes closer to 1 compared to other months—indicating that these models successfully detect meaningful treatment effect heterogeneity during those periods. In these months, the T-, R-, DR-learners, and the Generalized Random Forest exhibit comparable performance, with the T-Learner attaining slightly larger coefficients. In contrast, in earlier months (January–March), and late in the year (September–December), all learners yield low and statistically insignificant coefficients as expected, since in these periods the effect is constant and estimated to be zero. In the mid-year months (June–August) the estimated β_2 coefficients are moderately significant. Overall, Table 6 confirms the findings from previous sections, where we observed a sizable and heterogeneous COVID-19 impact between April and July, followed by a gradual recovery with decreasing heterogeneity beginning in August and a return to negligible constant effects by October.

[32]Following Chernozhukov et al. (2023), the goodness-of-fit of a CATE estimator $\hat{\Delta}(X_{\{t_s-1,t_s-2\}})$ in the BLP framework is indexed by the signal strength measure $\Xi := |\beta_2|^2 \cdot \text{Var}(\hat{\Delta}(X_{\{t_s-1,t_s-2\}})) = \text{Corr}^2(\Delta(X_{\{t_s-1,t_s-2\}}), \hat{\Delta}(X_{\{t_s-1,t_s-2\}})) \cdot \text{Var}(\Delta(X_{\{t_s-1,t_s-2\}}))$. Maximizing Ξ is equivalent to maximizing the correlation between the estimated and true CATE functions, which in practice corresponds to maximizing the R^2 of the BLP regression. Thus, the learner with the highest R^2 in a period is the one picking-up an higher degree of treatment effect heterogeneity.

Month	S-learner	T-learner	R-learner	DR-learner	Causal RF
	-3.132	-0.361	-0.184	-0.143	-0.474
Jan	(0.443)	(0.408)	(0.549)	(0.551)	(0.588)
	$R^2 = 0.00003$	$R^2 = 0.00035$	$R^2 = 0.00018$	$R^2 = 0.00018$	$R^2 = 0.00014$
	-0.466	-0.164	0.095	0.137	0.044
Feb	(0.522)	(0.566)	(0.596)	(0.663)	(0.545)
	$R^2 = 0.00024$	$R^2 = 0.00017$	$R^2 = 0.00014$	$R^2 = 0.00009$	$R^2 = 0.00017$
	2.023	0.236	0.169	0.158	0.665
Mar	(0.487)	(0.506)	(0.608)	(0.592)	(0.357)
	$R^2 = 0.00024$	$R^2 = 0.00022$	$R^2 = 0.00013$	$R^2 = 0.00014$	$R^2 = 0.00042$
	3.933	1.395	1.352	1.379	1.824
Apr	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	$R^2 = 0.0223$	$R^2 = 0.02493$	$R^2 = 0.02168$	$R^2 = 0.02176$	$R^2 = 0.02186$
	3.304	1.193	1.295	1.306	1.733
May	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	$R^2 = 0.01023$	$R^2 = 0.01486$	$R^2 = 0.01431$	$R^2 = 0.01484$	$R^2 = 0.01382$
	3.393	0.702	0.842	0.822	1.415
Jun	(0.06)	(0.014)	(0.005)	(0.007)	(0.003)
	$R^2 = 0.00161$	$R^2 = 0.0029$	$R^2 = 0.00345$	$R^2 = 0.00325$	$R^2 = 0.00395$
	3.751	0.591	0.534	0.516	1.161
Jul	(0.142)	(0.072)	(0.127)	(0.171)	(0.044)
	$R^2 = 0.00102$	$R^2 = 0.00149$	$R^2 = 0.0010$	$R^2 = 0.0008$	$R^2 = 0.00135$
	5.443	0.685	0.961	0.956	1.471
Aug	(0.058)	(0.04)	(0.008)	(0.007)	(0.024)
	$R^2 = 0.00172$	$R^2 = 0.00292$	$R^2 = 0.0033$	$R^2 = 0.0033$	$R^2 = 0.00231$
	0.985	0.183	0.184	0.141	0.331
Sep	(0.659)	(0.614)	(0.531)	(0.574)	(0.549)
	$R^2 = 0.0000$	$R^2 = 0.0001$	$R^2 = 0.0000$	$R^2 = 0.0001$	$R^2 = 0.0001$
	3.641	0.363	0.382	0.447	0.769
Oct	(0.301)	(0.323)	(0.336)	(0.268)	(0.297)
	$R^2 = 0.0004$	$R^2 = 0.0004$	$R^2 = 0.0005$	$R^2 = 0.0000$	$R^2 = 0.0004$
	1.411	0.068	0.116	0.106	0.379
Nov	(0.563)	(0.627)	(0.693)	(0.689)	(0.556)
	$R^2 = 0.0001$	$R^2 = 0.0001$	$R^2 = 0.0000$	$R^2 = 0.0000$	$R^2 = 0.0001$
	2.387	0.214	0.429	0.411	0.561
Dec	(0.42)	(0.579)	(0.29)	(0.303)	(0.444)
	$R^2 = 0.0002$	$R^2 = 0.0001$	$R^2 = 0.0004$	$R^2 = 0.0004$	$R^2 = 0.0002$

Table 6: Coefficient β_2 with their estimated median p-value across 100 A-M splits for the BLP considering different estimators of the CATE. The same analysis has been done with a single split without significant changes. Results are available upon request. P-values are clustered at the individual level.

To evaluate heterogeneity in treatment effects using our estimated models $\hat{\Delta}((X_{\{t_s-1,t_s-2\}}))$, we also use the *Group Average Treatment Effects (GATES)* following Chernozhukov et al. (2023). We slice the distribution of the estimated CATE into K parts and we are interested

in the average effect of firms within each slice. Table 7 displays the latter averages. Formally, for a partition of the support of $\hat{\Delta}(X_{\{t_s-1,t_s-2\}})$ into 4 quartile-based groups $G_k := \{\Delta \in I_k\}$, the GATE for group G_k is defined as

$$\gamma_k := \mathbb{E}[\tilde{Y}^{ATE}_{\{t_s-1,t_s\}} \mid G_k].$$

Situations where the γ_ks are similar indicate that no systematic heterogeneity was detected.

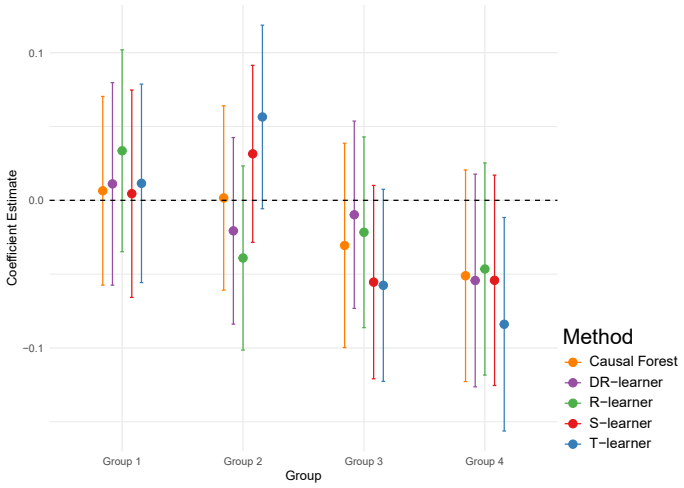
	Group	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
S-Learner	**1**	0.005	-0.005	-0.032	-0.368	-0.263	-0.119	-0.085	-0.085	-0.026	-0.04	-0.011	-0.026
		(0.981)	(0.981)	(0.975)	(0.001)	(0.004)	(0.003)	(0.024)	(0.178)	(0.772)	(0.966)	(0.983)	(0.979)
	2	-0.016	-0.005	-0.003	-0.171	-0.176	-0.097	-0.054	-0.056	-0.027	-0.021	0.000	-0.012
		(0.982)	(0.986)	(0.987)	(0.003)	(0.005)	(0.006)	(0.063)	(0.144)	(0.786)	(0.878)	(0.980)	(0.985)
	3	-0.012	-0.008	0.001	-0.108	-0.133	-0.062	-0.034	-0.034	-0.024	-0.014	-0.004	-0.004
		(0.985)	(0.984)	(0.987)	(0.000)	(0.005)	(0.005)	(0.077)	(0.164)	(0.733)	(0.799)	(0.982)	(0.988)
	4	-0.028	-0.005	0.011	-0.082	-0.052	-0.028	-0.025	-0.025	-0.021	0.005	-0.011	-0.007
		(0.975)	(0.978)	(0.983)	(0.000)	(0.006)	(0.003)	(0.078)	(0.178)	(0.773)	(0.979)	(0.987)	(0.982)
T-Learner	**1**	0.006	-0.003	-0.025	-0.387	-0.286	-0.120	-0.096	-0.096	-0.029	-0.031	-0.007	-0.025
		(0.987)	(0.986)	(0.979)	(0.000)	(0.003)	(0.001)	(0.023)	(0.124)	(0.799)	(0.879)	(0.988)	(0.982)
	2	0.000	-0.003	-0.016	-0.175	-0.173	-0.104	-0.046	-0.046	-0.019	-0.023	-0.006	-0.016
		(0.986)	(0.983)	(0.981)	(0.002)	(0.005)	(0.003)	(0.063)	(0.163)	(0.786)	(0.899)	(0.980)	(0.984)
	3	-0.012	-0.004	-0.005	-0.096	-0.123	-0.066	-0.026	-0.026	-0.028	-0.005	-0.011	-0.011
		(0.988)	(0.985)	(0.984)	(0.002)	(0.002)	(0.003)	(0.017)	(0.177)	(0.774)	(0.774)	(0.983)	(0.983)
	4	-0.028	0.003	0.016	-0.077	-0.047	-0.022	-0.019	-0.019	-0.016	0.006	-0.013	-0.003
		(0.977)	(0.987)	(0.983)	(0.001)	(0.020)	(0.004)	(0.082)	(0.182)	(0.776)	(0.777)	(0.981)	(0.990)
R-Learner	**1**	0.006	-0.015	-0.031	-0.382	-0.304	-0.133	-0.087	-0.087	-0.022	-0.042	-0.014	-0.042
		(0.983)	(0.980)	(0.975)	(0.004)	(0.004)	(0.004)	(0.063)	(0.115)	(0.656)	(0.636)	(0.982)	(0.966)
	2	-0.004	0.002	-0.005	-0.179	-0.172	-0.101	-0.046	-0.046	-0.017	-0.011	-0.002	0.001
		(0.982)	(0.981)	(0.983)	(0.000)	(0.003)	(0.004)	(0.061)	(0.162)	(0.688)	(0.989)	(0.989)	(0.984)
	3	-0.019	-0.015	0.004	-0.105	-0.107	-0.068	-0.024	-0.024	-0.029	-0.015	0.000	0.001
		(0.979)	(0.986)	(0.988)	(0.001)	(0.006)	(0.005)	(0.063)	(0.176)	(0.777)	(0.768)	(0.988)	(0.989)
	4	-0.022	-0.002	0.005	-0.084	-0.053	-0.013	-0.033	-0.033	-0.015	0.006	-0.019	-0.009
		(0.983)	(0.984)	(0.985)	(0.005)	(0.006)	(0.010)	(0.076)	(0.174)	(0.783)	(0.564)	(0.982)	(0.989)
DR-Learner	**1**	0.002	-0.014	-0.029	-0.383	-0.303	-0.134	-0.091	-0.091	-0.014	-0.041	-0.015	-0.032
		(0.987)	(0.983)	(0.976)	(0.006)	(0.000)	(0.000)	(0.028)	(0.171)	(0.782)	(0.986)	(0.984)	(0.975)
	2	-0.007	0.001	0.000	-0.168	-0.159	-0.106	-0.052	-0.055	-0.021	-0.013	-0.002	-0.006
		(0.985)	(0.983)	(0.985)	(0.003)	(0.001)	(0.000)	(0.06)	(0.166)	(0.689)	(0.986)	(0.984)	(0.982)
	3	-0.013	-0.012	0.000	-0.113	-0.105	-0.061	-0.029	-0.02	-0.023	-0.009	-0.006	-0.003
		(0.988)	(0.986)	(0.988)	(0.005)	(0.006)	(0.002)	(0.077)	(0.172)	(0.780)	(0.967)	(0.986)	(0.987)
	4	-0.023	-0.002	0.001	-0.081	-0.054	-0.014	-0.031	-0.031	-0.021	0.011	-0.015	-0.007
		(0.982)	(0.986)	(0.982)	(0.000)	(0.004)	(0.002)	(0.075)	(0.175)	(0.644)	(0.878)	(0.985)	(0.986)
Generalized Random Forest	**1**	-0.004	-0.002	-0.036	-0.388	-0.303	-0.142	-0.098	-0.098	-0.022	-0.046	-0.015	-0.028
		(0.982)	(0.989)	(0.971)	(0.002)	(0.007)	(0.000)	(0.022)	(0.122)	(0.979)	(0.964)	(0.986)	(0.978)
	2	-0.008	-0.014	0.000	-0.166	-0.161	-0.111	-0.055	-0.055	-0.026	-0.015	-0.020	-0.016
		(0.986)	(0.989)	(0.985)	(0.000)	(0.007)	(0.002)	(0.056)	(0.156)	(0.976)	(0.984)	(0.983)	(0.984)
	3	-0.014	-0.005	0.002	-0.108	-0.131	-0.057	-0.035	-0.035	-0.025	-0.004	-0.016	-0.007
		(0.984)	(0.984)	(0.985)	(0.001)	(0.002)	(0.006)	(0.071)	(0.171)	(0.679)	(0.981)	(0.981)	(0.982)
	4	-0.021	-0.002	0.012	-0.081	-0.041	-0.010	-0.009	-0.009	-0.016	0.012	0.011	0.003
		(0.981)	(0.984)	(0.982)	(0.006)	(0.005)	(0.006)	(0.085)	(0.185)	(0.682)	(0.986)	(0.985)	(0.989)

Table 7: Median GATES coefficients with their estimated median p-value across 100 A-M splits for the GATEs considering different estimators of CATE. The same analysis was performed with a single split with no significant changes. The results are available on request.

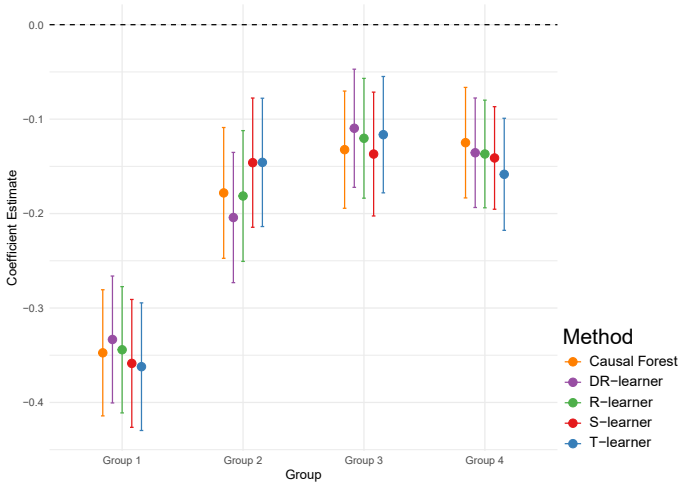
The results presented in Table 7 corroborate the findings from the main analysis. In

particular, the placebo tests for the first three months of the year are successful (i.e., the estimated GATES are all equal and not statistically significant) and the GATES estimates are substantially larger in magnitude and statistically different during the COVID-19 months—specifically April, May, and June—with some residual evidence in July. In other words, the probability of exporting declined significantly across all subgroups during these months, with the impact varying by group, thereby confirming the presence of meaningful treatment effect heterogeneity.

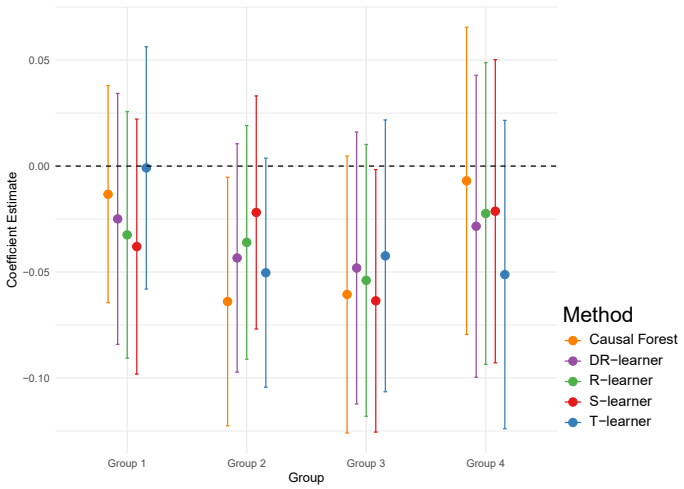
Among the different learners, the T-Learner and DR-Learner tend to exhibit more pronounced variation in GATE values across quantiles, suggesting stronger sensitivity to heterogeneity in treatment effects. The Generalized Random Forest also captures heterogeneity effectively, especially in the upper quantiles, while the S-Learner generally shows less differentiation across groups, indicating limited responsiveness to heterogeneity. Overall, the T-Learner performs competitively, aligning well with the observed patterns of export decline and offering reliable subgroup-specific estimates. Figure 9 provides a visual representation of the GATES for the different estimators of CATE with the associated confidence level for the first split (which are very similar to the more robust multisplit results in Table 7).



January



April



October

Figure 9: GATES estimates from January, April, and October 2020. The results are shown for the four quartiles according to CATE. In each graph, the colored bars are from left to right for Generalized Random Forest (orange), DR-learner (purple), R-learner (green), S-learner (red) and T-learner (blue). Figure Appx.11 in Appendix F shows the estimates for all months of 2020.

Having the BLP and GATES analyses confirmed the substantial heterogeneity detected by the SPE, it becomes particularly interesting to further explore the sources of these differences following a procedure that is exactly the same of the one used in the previous CA analyses. What is different here is the inference part. We repeat 100 times the split in A and M samples, together with the calculation of the first and third quartiles of the distribution of the estimated CATE and the comparison of the mean characteristics of the firms contained in the tails of the distribution. Then we aggregate the results by taking the median of the estimated difference in means and the corresponding p-values. More formally, let X_{t_s-1} denote the vector of observed covariates. The CLAN compares the average covariate values between the "least affected group" G_1 and the "most affected group" G_K, as defined by the GATES framework.

Table 8 presents the results of the CLAN across various estimators. The results confirm the findings from the CA analysis reported in Table 5, offering an additional diagnostic.

The CLANs in Table 8 highlights a consistent pattern of treatment effect heterogeneity across industrial characteristics, export behavior, and time dummies. Notably, the difference in means for sectors such as Agriculture, Textile, and Wood are significant across all learners, with p-values close to zero. For instance, the Textile sector shows systematically higher representation among more affected firms, especially under the R- and DR-Learners, suggesting that sectoral affiliation plays a key role in shaping heterogeneity. This is broadly aligned with the CA findings, where Textile also shows positive and statistically significant estimate.

Temporal patterns are also consistent across methods. Months corresponding to the COVID-19 peak (April–June) exhibit strong and positive CLAN differences, particularly under the Generalized Random Forest. For example, April differences range from 0.30 to 0.36 across learners, all significant at the 1% level. These closely mirror the CADiff estimates for the same months, validating the robustness of the temporal heterogeneity captured by both approaches.

Differences in trade intensity are particularly stark. CLAN reveals substantial negative differences in Value Exported (log) and Value Imported (log) across all estimators. These results suggest that more severely affected firms are systematically smaller traders and confirm the previous findings from CA.

Moreover, similar to the case of diversification, analyzing both the export and import sides supports and reinforces the previous finding that greater experience in importing inputs from abroad reduces the severity of the shock.

	T-Learner			**S-Learner**			**R-Learner**			**DR-Learner**			**Generalized Random Forest**		
Outcome variable	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Agriculture	-0.0653			-0.0385	0.0912		-0.0916			-0.0730			-0.0534		
	(0.0000)			(0.0000)	(0.0000)		(0.0000)			(0.0000)			(0.0000)		
Chemicals	-0.007			-0.0253	-0.0112		-0.0229			-0.0225			-0.0230		
	(0.5916)			(0.0036)	(0.4618)		(0.0014)			(0.0085)			(0.0125)		
Manufacturing	-0.019			-0.0283	-0.1183		-0.0144			-0.0175			-0.0141		
	(0.0186)			(0.0000)	(0.0000)		(0.0307)			(0.0380)			(0.1655)		
Metals	-0.013			-0.0195			-0.0107			-0.0175			-0.0195		
	(0.1525)			(0.0140)			(0.0919)			(0.0290)			(0.0162)		
Special	-0.0008			-0.0034			0.0015			-0.0003			-0.0009		
	(1.0000)			(1.0000)			(0.6307)			(1.0000)			(1.0000)		
Textile	0.0939			0.0707			0.0996			0.0932			0.0769		
	(0.0000)			(0.0000)			(0.0000)			(0.0000)			(0.0000)		
Wood	0.0228			0.0111			0.0297			0.0277			0.0176		
	(0.0000)			(0.1662)			(0.0000)			(0.0000)			(0.0039)		
Air	0.1347	0.1158	0.2035	0.1118		0.2768	0.1345	0.1108	0.1775	0.1280	0.1055	0.1729	0.1254	0.1044	0.3090
	(0.0000)	(0.0000)	(0.0000)	(0.0000)		(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Land	0.0051	0.0049	0.0029	-0.0176		0.0124	-0.0067	-0.0057	-0.0092	-0.0053	-0.0023	-0.0076	-0.0106	-0.0041	-0.0135
	(0.5944)	(0.7076)	(0.2341)	(0.0521)		(0.1008)	(0.3091)	(0.3712)	(0.3296)	(0.4692)	(0.8440)	(0.3065)	(0.2261)	(0.4418)	(0.4765)
Sea	-0.1585	-0.1319	-0.2459	-0.1469		-0.2187	-0.1929	-0.1610	-0.2777	-0.1720	-0.1394	-0.2643	-0.1687	-0.1430	-0.4199
	(0.0000)	(0.0000)	(0.0000)	(0.0000)		(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
January	-0.1314	-0.1316		-0.1124	-0.1402		-0.0694	-0.0676		-0.1307	-0.1319		-0.1856	-0.1866	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
February	-0.1259	-0.1277		-0.1149	-0.1302		-0.0820	-0.0860		-0.1319	-0.1352		-0.1961	-0.1979	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
March	-0.1202	-0.1200		-0.1530	-0.1187		-0.1292	-0.1309		-0.1316	-0.1311		-0.1806	-0.1792	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
April	0.2963	0.3004		0.3354	0.2897		0.3014	0.3074		0.3052	0.3118		0.3641	0.3691	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
May	0.2786	0.2791		0.2225	0.2893		0.2804	0.2822		0.2855	0.2868		0.3515	0.3539	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
June	0.1301	0.1293		0.1211	0.1453		0.1546	0.1519		0.1313	0.1293		0.1791	0.1746	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
July	0.0256	0.0263		0.0115	0.0111		-0.0391	-0.0388		0.0295	0.0275		0.0372	0.0354	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
August	0.0016	0.0014		0.01323	0.00254		0.0190	0.0187		-0.0004	-0.0018		0.0006	0.0001	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
September	-0.0589	-0.0580		-0.0662	-0.0896		-0.0407	-0.0388		-0.0636	-0.0624		-0.0784	-0.0765	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
October	-0.0796	-0.0803		-0.0657	-0.09877		-0.1297	-0.1332		-0.0829	-0.0843		-0.1016	-0.1016	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
November	-0.1157	-0.1177		-0.1091	-0.1675		-0.1103	-0.1106		-0.1139	-0.1146		-0.1672	-0.1658	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
December	-0.1001	-0.1000		-0.1341	-0.0908		-0.1548	-0.1543		-0.1047	-0.1045		-0.1344	-0.1238	
	(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)	
Number of export destination (ND)	-0.2390	-0.1800	-0.3055	-0.4465	-0.3560	-0.2855	-0.3315	-0.2238	-0.4854	-0.4096	-0.3187	-0.5247	-0.3652	-0.2952	-1.1601
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of import origins (NO)	-0.8198	-0.7726	-1.2243	-0.9680	-0.8261	-1.0672	-0.8946	-0.8183	-1.0788	-0.8927	-0.8075	-1.3310	-0.9045	-0.8278	-1.9191
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of exported products (NP)	0.0379	-0.2005	-0.2671	-0.3269	-0.2733	-0.3246	-0.0921	-0.2508	-0.4249	-0.0668	-0.2476	-0.3121	-0.1759	-0.2649	-0.3483
	(0.5970)	(0.0393)	(0.0458)	(0.0066)	(0.0030)	(0.0438)	(0.4009)	(0.0033)	(0.0000)	(0.4806)	(0.0098)	(0.0072)	(0.4934)	(0.01099)	(0.01129)
Value Exported (log)	-0.5692	-0.5133	-0.9142	-0.6218	-0.5611	-1.9233	-0.7167	-0.6382	-1.0998	-0.6545	-0.5854	-1.1236	-0.7247	-0.6451	-1.8854
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Value Imported (log)	-1.3158	-1.2947	-2.1863	-1.7081	-1.5344	-1.1924	-1.7072	-1.6628	-2.7229	-1.5690	-1.4955	-2.6176	-1.5667	-1.4588	-3.1997
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Deviation from sectoral mean		✓	✓		✓	✓		✓	✓		✓	✓		✓	✓
Deviation from monthly mean			✓			✓			✓			✓			✓

Table 8: CLANs with 100 splits for S-Learner, T-Learner, R-Learner, DR-Learner, Generalized Random Forest. We report the median of joint p-values.

5 Concluding discussion

In this paper, we show the potential of ML techniques for building counterfactuals, identifying the most affected subpopulations and the sources of treatment effect heterogeneity in scenarios where a credible control group is unavailable and it is difficult to define ex-ante the intensity of the shock for each unit.

In the application we consider, we concentrate on the effects of an economy-wide shock such as COVID-19 on a firm's export behaviour. Using data from the Colombian customs office, we estimate that, during 2020, on average, the COVID-19 shock decreased a firm's probability of surviving in the export market by about 15 to 20 percentage points in April and May and by approximately 5 to 8 percentage points in June and July. By analysing the estimated treatment effect distribution, we reveal that these average results hide considerable heterogeneity. For example, in April 2020, we find that for some exporters COVID-19 decreased their survival probability by 55 percentage points. We identify the firms most and least affected by COVID-19 and we compare their characteristics by integrating the Sorted Partial Effects methodology with our causal ML approach. We emphasize how the integration into global value chains on the import side, both in terms of the number of countries from which a firm sources and the value of imports, is an important factor of resilience for exporters facing the COVID-19 shock.

From a methodological point of view, we show practitioners how to apply the generic ML tools proposed by Chernozhukov et al. (2023) to a context in which there is no control group available; we suggest how to use in-time placebo tests to check the credibility of counterfactual estimates; finally, we provide evidence indicating that in the Sorted Partial Effects analysis, in which the focus lies on the tails of the distribution of the treatment effects, it is critical to correct the estimation error arising from the necessarily imperfect reconstruction of the unobservable counterfactual.

While this method is specifically designed for analyzing the heterogeneous impacts of economy-wide shocks, there exists potential utility in employing this approach in less extreme situations where policies or shocks may exhibit unobservable spillovers that are challenging to model in advance. In such contexts, our empirical framework proves advantageous in detecting these potential heterogeneous indirect effects, as it circumvents the need for a priori identification of a control group of untreated units.

Finally, in this paper we also demonstrate that ML methods can be applied successfully to predict firms' trade potential. We consider ML methods a promising tool to assist firms and public agencies in their export decision-making processes. The bulk of countries possesses export promotion agencies whose objective is to sustain firms' internationalization activities by lowering the costs of information acquisition (Broocks and Van Biesebroeck, 2017; Munch and Schaur, 2018). Given that exporters' dynamics can be understood as a complex learning process dense of interdependencies (complementarity or substitutability) between products and destination markets (from the perspective of technology/knowledge, local tastes,

legal requirements, and marketing and distribution costs) and that ML techniques have been successfully applied to predict firm performances in such settings, we believe that an important avenue for future research is to test the effectiveness of using these techniques and firm-level data to build recommendation systems. These systems could help firms identify their latent comparative advantages and provide export diversification and differentiation recommendations.

Table 5: Estimated differences in means of the estimated treatment effect and other covariates between the group of more affected and the group of less affected firms ($CADiff$) applying the classification analysis to the $Y - SUM$ estimates

Outcome variable	$\beta_{1,f}^{(1)}$	$\beta_{1,f}^{(2)}$	$\beta_{1,f}^{(3)}$
TE	-1.0910	-1.0930***	-1.0710
Agriculture	-0.0616		
Chemicals	-0.0192		
Manufacturing	0.0112		
Metals	0.0109		
Special	0.0059		
Textile	0.0486		
Wood	0.0041		
Air	0.0411	0.0271	0.0289
Land	0.0086	0.0062	0.0068
Sea	-0.0482	-0.0321	-0.0343
Jan	-0.0190	-0.0189	
Feb	-0.0242	-0.0237	
Mar	-0.0181	-0.0181	
Apr	0.0631	0.0630	
May	0.0620	0.0612	
Jun	0.0166	0.0167	
Jul	0.0033	0.0028	
Aug	-0.0050	-0.0053	
Sep	-0.0169	-0.0167	
Oct	-0.0216	-0.0208	
Nov	-0.0218	-0.0222	
Dec	-0.0183	-0.0181	
Number of export destinations (ND)	0.3310	0.3470	0.3306
Number of import origins (NO)	0.0350	-0.0595	-0.1077
Number of exported products (NP)	0.6050	0.4670	0.4275
Containment Index Stringency Export	-0.2280	-0.0264	0.9690
Containment Index Stringency Import	-4.2180	-4.4910	-0.0520
Value Exported (log)	-0.2700	-0.2760	-0.1800
Value Imported (log)	-0.0910	0.0296	0.0040
Deviation from sectoral mean		✓	✓
Deviation from monthly mean			✓

Notes: column 1 does not include sector or month variables in the regression; column 2 includes sector in the regression, and, column 3 includes both the sector and month variables. *** means significant at 1%, ** at 5%, * at 10%. Standard errors are obtained by bootstrapping the whole estimation process and joint p-values are adjusted to take into account the simultaneous testing of all the variables.

References

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

Albornoz, F., H. F. C. Pardo, G. Corcos, and E. Ornelas (2012). Sequential exporting. *Journal of International Economics 88*(1), 17–31.

Antras, P. and D. Chor (2022). *Global Value Chains*, Volume 5. Elsevier.

Antràs, P., S. J. Redding, and E. Rossi-Hansberg (2023). Globalization and pandemics. *American Economic Review 113*(4), 939–981.

Arkolakis, C., S. Ganapati, and M.-A. Muendler (2021). The extensive margin of exporting products: A firm-level analysis. *American Economic Journal: Macroeconomics 13*(4), 182–245.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Baiardi, A. and A. A. Naghi (2024). The value added of machine learning to causal inference: Evidence from revisited studies. *The Econometrics Journal 27*(2), 213–234.

Baldwin, R. and E. Tomiura (2020). Thinking ahead about the trade impact of COVID-19. *Economics in the Time of COVID-19 59*.

Bargagli-Stoffi, F. J., J. Niederreiter, and M. Riccaboni (2021). Supervised learning for the prediction of firm dynamics. In *Data Science for Economics and Finance: Methodologies and Applications*, pp. 19–41. Springer International Publishing Cham.

Berthou, A. and S. Stumpner (2024). Trade under lockdown. *Journal of International Economics 152*, 104013.

Bonadio, B., Z. Huo, A. A. Levchenko, and N. Pandalai-Nayar (2020). Global supply chains in the pandemic. National Bureau of Economic Research, No. w27224.

Breinlich, H., V. Corradi, N. Rocha, M. Ruta, J. Santos Silva, and T. Zylkin (2022). Machine learning in international trade research-evaluating the impact of trade agreements.

Broocks, A. and J. Van Biesebroeck (2017). The impact of export promotion on export market entry. *Journal of International Economics 107*, 19–33.

Buhl-Wiggers, J., J. T. Kerwin, J. Muñoz-Morales, J. Smith, and R. Thornton (2024). Some children left behind: Variation in the effects of an educational intervention. *Journal of Econometrics 243*(1-2), 105256.

Cepeda-López, F., F. Gamboa-Estrada, C. León-Rincón, and H. Rincón-Castro (2019). Colombian liberalization and integration to world trade markets: Much ado about nothing. *Borradores de Economía; No. 1065*.

Cerqua, A. and M. Letta (2020). Local economies amidst the COVID-19 crisis in Italy: a tale of diverging trajectories. *COVID Economics* (60), 142–171.

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.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2023). Fischer-schultz lecture: Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in india. Technical report, Working Papers hal-04238425, HAL.

Chernozhukov, V., I. Fernández-Val, and Y. Luo (2018). The sorted effects method: discovering heterogeneous effects beyond their averages. *Econometrica 86*(6), 1911–1938.

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*.

Curth, A. and M. Van der Schaar (2021). Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1810–1818. PMLR.

de Lucio, J., R. Mínguez, A. Minondo, and A. F. Requena (2020, December). Impact of COVID-19 containment measures on trade. Working Papers 2101, Department of Applied Economics II, Universidad de Valencia.

Dehling, H. and W. Philipp (2002). Empirical process techniques for dependent data. In *Empirical process techniques for dependent data*, pp. 3–113. Springer.

Deryugina, T., G. Heutel, N. H. Miller, D. Molitor, and J. Reif (2019). The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review 109*(12), 4178–4219.

Drees, H. and H. Rootzén (2010). Limit theorems for empirical processes of cluster functionals.

Eslava, M., J. Tybout, D. Jinkins, C. Krizan, J. Eaton, et al. (2015). A search and learning model of export dynamics. In *2015 Meeting Papers*, Number 1535. Society for Economic Dynamics.

Espitia, A., A. Mattoo, N. Rocha, M. Ruta, and D. Winkler (2021). Pandemic trade: COVID-19, remote work and global value chains. *The World Economy*.

Evenett, S. (2020). Sicken thy neighbour: The initial trade policy response to COVID-19. *The World Economy 43*(4), 828–839.

Fabra, N., A. Lacuesta, and M. Souza (2022). The implicit cost of carbon abatement during the COVID-19 pandemic. *European Economic Review 147*, 104165.

Felbermayr, G. and H. Görg (2020). Implications of COVID-19 for globalization. *KIELER BEITRÄGE ZUR*, 3.

Garavito, A., A. L. Garavito-Acosta, E. Montes-Uribe, J. H. Toro-Córdoba, C. Agudelo-Rivera, V. A. Corredor-Alfonso, Á. D. Carmona, M. M. Collazos-Gaitan, C. González-Sabogal, M. D. Hernandez-Bejarano, et al. (2020). Ingresos externos corrientes de colombia: desempeño exportador, avances y retos. *Revista Ensayos Sobre Política Económica; No. 95, julio 2020. Pág.: 1-81*.

Grossman, G. M., E. Helpman, and H. Lhuillier (2021). Supply Chain Resilience: Should Policy Promote Diversification or Reshoring? National Bureau of Economic Research, No. w29330.

Hale, T., A. Petherick, T. Phillips, and S. Webster (2020). Variation in government responses to COVID-19. *Blavatnik school of government working paper 31*, 2020–11.

Halpern, L., M. Koren, and A. Szeidl (2015). Imported inputs and productivity. *American Economic Review 105*(12), 3660–3703.

Keele, L. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis 23*(3), 313–335.

Kennedy, E. H. et al. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497 5*.

Knaus, M. C. (2022). Double machine learning-based programme evaluation under

unconfoundedness. *The Econometrics Journal 25*(3), 602–627.

Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, Volume 2033. Springer.

Kramarz, F., J. Martin, and I. Mejean (2020). Volatility in the small and in the large: The lack of diversification in international trade. *Journal of international economics 122*, 103276.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences 116*(10), 4156–4165.

Lafrogne-Joussier, R., J. Martin, and I. Mejean (2022). Supply shocks in supply chains: Evidence from the early lockdown in China. *IMF Economic Review*, 1–46.

Liu, X., E. Ornelas, and H. Shi (2021). The Trade Impact of the COVID-19 Pandemic. *The World Economy*.

Magnan, N., V. Hoffmann, N. Opoku, G. G. Garrido, and D. A. Kanyam (2021). Information, technology, and market rewards: Incentivizing aflatoxin control in Ghana. *Journal of Development Economics 151*, 102620.

McKenzie, D. and D. Sansone (2019). Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria. *Journal of Development Economics 141*, 102369.

Mirzaei, E., V. R. Kostic, A. Maurer, et al. An empirical bernstein inequality for dependent data in hilbert spaces and applications. In *The 28th International Conference on Artificial Intelligence and Statistics*.

Munch, J. and G. Schaur (2018). The effect of export promotion on firm-level performance. *American Economic Journal: Economic Policy 10*(1), 357–87.

Navas, A., F. Serti, and C. Tomasi (2020). The role of the gravity forces on firms' trade. *Canadian Journal of Economics/Revue canadienne d'économique 53*(3), 1059–1097.

Nie, X. and S. Wager (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika 108*(2), 299–319.

Nitsch, V. (2022). COVID-19 and international trade: Evidence from New Zealand. *Economics Letters*, 110627.

Ocampo, J. A. and C. Tovar (2000). Colombia in the classical era of 'inward-looking development', 1930–74. In E. Cárdenas, J. Ocampo, and R. Thorp (Eds.), *An economic history of twentieth-century Latin America*, pp. 239–281. Springer.

Okui, R. and T. Yanagi (2019). Panel data analysis with heterogeneous dynamics. *Journal of Econometrics 212*(2), 451–475.

Pierce, J. R. and P. K. Schott (2012). A concordance between ten-digit US Harmonized System Codes and SIC/NAICS product classes and industries. *Journal of Economic and Social Measurement 37*(1-2), 61–96.

Probst, P., M. N. Wright, and A.-L. Boulesteix (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery 9*(3), e1301.

Sadhanala, V. and R. J. Tibshirani (2019). Additive models with trend filtering. *The Annals of Statistics 47*(6), 3032–3068.

Salditt, M., T. Eckes, and S. Nestler (2024). A tutorial introduction to heterogeneous treatment

effect estimation with meta-learners. *Administration and Policy in Mental Health and Mental Health Services Research 51*(5), 650–673.

Singh, R. S. (1975). On the glivenko–cantelli theorem for weighted empiricals based on independent random variables. *The Annals of Probability*, 371–374.

Tsybakov, A. B. and A. B. Tsybakov (2009). Nonparametric estimators. *Introduction to Nonparametric Estimation*, 1–76.

Van Biesebroeck, J., E. Yu, and S. Chen (2015). The impact of trade promotion services on Canadian exporter performance. *Canadian Journal of Economics/Revue canadienne d'économique 48*(4), 1481–1512.

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

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.

World Bank (2020). Global Economic Prospects, June 2020. Washington, DC: World Bank. Available at: https://openknowledge.worldbank.org/handle/10986/33748.

Zuluaga, Ó. I., C. G. Cano, J. J. Echavarría, F. Tenjo, J. P. Zárate, and J. D. Uribe (2009). Informe de la Junta Directiva al Congreso de la República – Marzo de 2009. Banco de la República de Colombia.

A Appendix - The Colombian economy amidst the COVID-19 crisis

Colombia exports little compared to other countries in Latin America with similar development levels. In recent years, the share of total exports of Colombian GDP has oscillated around 15%, well below other countries in the region such as Chile and Mexico (Cepeda-López et al., 2019). Colombia started to open its economy in the 1990s with several market-oriented reforms to liberalize financial and capital markets. Although the Colombian economy was still relatively closed during most of the twentieth century (Ocampo and Tovar, 2000), it has been strongly affected by the global financial crisis in 2008-2009 (Zuluaga et al., 2009). Nowadays, despite the growing number of trade agreements, partners, and volume of trade, the integration of Colombia into world trade markets is still modest.

The main reason behind Colombia's poor performance is the low diversification of trade, with a prevalence of primary products, because of the relative abundance of natural resources and low-skilled labor. Besides, the emergence of raw products derived from mining has gained a larger share in total exports, reducing the importance of other products such as coffee, bananas, flowers, some labor-intensive manufactures, and petrochemicals (Garavito et al., 2020).

Since the outbreak of the COVID-19 pandemic, Colombia implemented early measures to contain the spread of COVID-19 and prepare the health system and mitigate the economic and social impact. The Colombian government issued non-compulsory requests for remote working to private companies on February 24, 2020; schools and universities were closed on March 16. On March 25, when there were fewer than a dozen deaths, the government implemented a complete and mandatory lockdown until April 13. During this period, only a few essential activities – such as health services, public services, communications, banking and financial services, food production, pharmaceuticals, and cleaning and disinfection products – were excluded. The partial lockdown implementation–between April 27 and May 11–allowed a gradual restoration of mobility, enabling a set of non-essential activities under security guidelines and protocols to guarantee social distancing. Most manufacturing activities were gradually allowed at this stage, while non-authorized activities were restricted to sell their products through electronic commerce platforms. Finally, from May 28, restrictions to the services sector have been lifted, and on September 1, the government announced the confinement end, and airports were opened.

To better cope with the emergency, Colombian authorities have introduced transitory provisions to secure international trade of essential products. Along with the lockdown measures, medicines, supplies, and equipment in the health sector had zero tariffs for six months. Besides, the export and re-export of these products were forbidden. There was a zero-tariff from April 7 to June 30 for raw materials such as maize, sorghum, soybeans, and soybean cake.

During the second half of 2020, the Colombian government implemented several economic recovery policies. These included unconditional cash transfers through programs such as *Ingreso Solidario*, *Familias en Acción*, *Jóvenes en Acción*, and the *Día sin IVA* (VAT-free day), alongside support from Family Compensation Funds (*Cajas de Compensación*), benefiting millions of households living in poverty or vulnerable conditions.

In terms of sectoral and business support, various credit lines and incentives were introduced. For example, financial assistance was provided to the country's main commercial airline through the *Fondo de Mitigación de Emergencias* (FOME), in order to preserve air connectivity. The government also launched credit lines through local development banks, most notably *Bancóldex*, such as *Colombia Responde*, initially targeted at the tourism, aviation, and entertainment sectors. These lines were later extended to other industries, offering reduced interest rates and guarantees, particularly for micro and small enterprises. Additional instruments such as *Reactívate* were introduced to help SMEs across all sectors finance the implementation of biosecurity protocols. A significant share of the incentives focused on so-called "strategic sectors" such as mining, infrastructure, and construction, with fiscal benefits and targeted subsidies aimed at stimulating investment and employment. However, our analysis does not aim to assess the effectiveness of these recovery measures. Indeed, many of them were not specifically targeted at exporting companies. Therefore, we do not create a counterfactual scenario depicting outcomes without such recovery policies, which can be explored in future research.

B Appendix - Data

Table Appx.1: Predictors for exporters' success

Variable	Description	Source
	Models: SUM and SAM	
NP, ND, NO	Number of products exported by, number of destinations where a company exports, and number of import origin countries for an exporter in a given month, respectively.	Authors' own elaboration from Colombian Customs Office (DIAN).
HH_p, HH_d	Product-Herfindahl Index, and Destination-Herfindahl Index. Measure the concentration of products at 6-digit HS, and the concentration at destination by company-month, respectively.	Authors' own elaboration from the Colombian Customs Office (DIAN).
Total value (exports)	Free on board value of the export transaction in US dollars for each company-month.	Colombian Customs Office (DIAN)
Total value (imports)	Free on board value of the import transaction in US dollars for each company-month.	Colombian Customs Office (DIAN)
Size	4 class dummies classifying firms according to the quartiles of the firm-level (Q1, Q2, Q3 and Q4) distribution of the total monthly value of exports (in ln).	Authors' own elaboration
Destination	Factor variable with one level (dummy variable) for each destination country where Colombian exporters operate by month.	Colombian Customs Office (DIAN)
Origin	Factor variable with one level (dummy variable) for each import origin country, where Colombian exporters operate by month.	Colombian Customs Office (DIAN)
Continent	Factor variable with one level (dummy variable) for each continent where Colombian exporters operate.	Authors' own elaboration
Department	Factor variable with one level (dummy variable) for each department (region) in Colombia from which companies operate.	Colombian Customs Office (DIAN)
Means of Transportation	4 class dummies indicating the means of transportation a company uses to perform a transaction (land, sea, air, others).	Colombian Customs Office (DIAN)
Sector	99 class dummies for the classification of company products by 2-digit HS code (corresponding to HS-chapters).	Authors' own elaboration
Industry	22 class dummies indicating the industries (HS-sections) where companies operate.	Authors' own elaboration from the Colombian Customs Office (DIAN).
Sector Experience	Factor variable with one level (dummy variable) for each sector. Takes value 1 in all periods after a company exports for the first time in a given sector (reflecting past experience in a sector).	Authors' own elaboration from the Colombian Customs Office (DIAN).
Destination Experience	Factor variable with one level (dummy variable) for each destination. Takes value 1 in all periods after a company exports for the first time in a given destination (reflecting past experience in a destination).	Authors' own elaboration from the Colombian Customs Office (DIAN).
Exporter (importer) Experience	Variable counting the accumulated value exported (imported) in the last twelve months.	Authors' own elaboration from the Colombian Customs Office (DIAN).
	Models: SAM (COVID-19 variables)	
Containment Economic Index	We consider the Economic Index from Hale et al. (2020) that provides a measure of the strength of the economic policies set in place to deal with the pandemic (such as income support and debt relief) for each country in the world. It ranges from 0 to 100. At the firm level we define two variables, one at the export and one at the import side, by taking a weighted average of these country level scores according to the monthly value of exports(imports) that a company ships(source) in every country.[1]	Hale et al. (2020) and Colombian Customs Office (DIAN).
Containment Government Index	We consider the Government Index from Hale et al. (2020) that measures the strictness of 'lockdown' style policies that primarily restrict people's behavior. It ranges from 0 to 100. At the firm level we define two variables, one at the export and one at the import side, by taking a weighted average of these country level scores according to the monthly value of exports(imports) that a company ships(source) in every country.	Hale et al. (2020) and Colombian Customs Office (DIAN).
Containment Health Index	We consider the Health Index from Hale et al. (2020) that combines 'lockdown' restrictions and closures with measures such as testing policy and contact tracing, short-term investment in healthcare, as well as investments in vaccine. Ranges from 0 to 100. At the firm level, we define two variables, one at the export and one at the import side, by taking a weighted average of these country-level scores according to the monthly value of exports(imports) that a company ships (source) in every country.	Hale et al. (2020) and Colombian Customs Office (DIAN).
Containment Stringency Index	We consider the Stringency Index from Hale et al. (2020) that records how the response of governments has varied over all indicators, becoming stronger or weaker over the course of the outbreak. Ranges from 0 to 100. At the firm level we define two variables, one at the export and one at the import side, by taking a weighted average of these country level scores according to the monthly value of exports(imports) that a company ships(source) in every country.	Hale et al. (2020) and Colombian Customs Office (DIAN).
	Models: SUM and SAM (variables only for Logit, Logit-LASSO, Logit-Ridge and SVM)	
Size*Industry	Factor variables with 5 levels for each industry. It takes the value 1 if the firm size is Q1, the value 2 if the firm size is Q2, the value 3 if the size is Q3 and the value 4 if the size is Q4 while operating in a particular industry. However, it takes the value 0 if a company is not active in this industry (regardless of size).	Authors' own elaboration based on the Colombian Customs Office (DIAN).
Size*Sector	Factor variables with 5 levels for each sector. It takes the value 1 if the firm size is Q1, the value 2 if the firm size is Q2, the value 3 if the size is Q3 and the value 4 if the size is Q4 while operating in a specific sector. However, it takes the value 0 if a company is not active in this sector (regardless of size).	Authors' own elaboration based on the Colombian Customs Office (DIAN).
Size*Means of Transportation	Factor variables with 5 levels for each sector. It takes the value 1 when the company size is Q1, value 2 when the company size is Q2, value 3 when the size is Q3, and value 4 when the size is Q4 while operating using a given means of transportation. However, it takes value 0 if a company is not operating using this means of transportation (for any size level).	Authors' own elaboration from the Colombian Customs Office (DIAN).
Size*Destination	Factor variables with 5 levels for each sector. It takes the value 1 when the company size is Q1, value 2 when the company size is Q2, value 3 when the size is Q3, and value 4 when the size is Q4 while operating in a given destination. However, it takes value 0 if a company is not operating in this destination (for any size level).	Authors' own elaboration from the Colombian Customs Office (DIAN).

* https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data

[1] When an exporter does not import, we impute the corresponding internal Index (Economic, Government, Health, and Stringency) of Colombia to create the corresponding import side Index.

[2] Only the variables and interactions listed in this table were used in the analysis (no second or higher degree polynomial function). Interactions were removed in tree-based models (XGBoost and Random Forest).

Table Appx.2: Sector-Industry mapping

Section (Industry)	Industry Name	HS-Chapter (Sector)
1	Live Animals/ Animal Products	1-5
2	Vegetable Products	6-14
3	Animal or Vegetable Fats/Oils	15
4	Prepared Foodstuffs	16-24
5	Mineral Products	25-27
6	Products of Chemical Industries	28-38
7	Plastics, Rubber	39-40
8	Raw Hides, Skins and Leather	41-43
9	Wood	44-46
10	Paper	47-49
11	Textile	50-63
12	Footwear	64-67
13	Art. of Stone, Cement	68-70
14	Jewelries	71
15	Base Metals	72-83
16	Machinery Equipment	84-85
17	Vehicles	86-89
18	Precision Instruments	90-92
19	Arms	93
20	Miscellaneous Manufactured Articles	94-96
21	Works of Art	97
22	Special Classification Provisions	98-99

Source: Author's elaboration using Pierce and Schott (2012) tables.

C Appendix - Descriptive Statistics

The left panel in Figure Appx.1 shows the evolution of total monthly exports during 2019 and 2020. The total monthly value of exports in 2020 is significantly lower than the one observed for the corresponding month in 2019, except for January and February. The lockdown measures implemented to contain the COVID-19 outbreak in Colombia and abroad had a severe impact between April and June–the value in April 2020 is half of the one observed in April 2019 (47%).

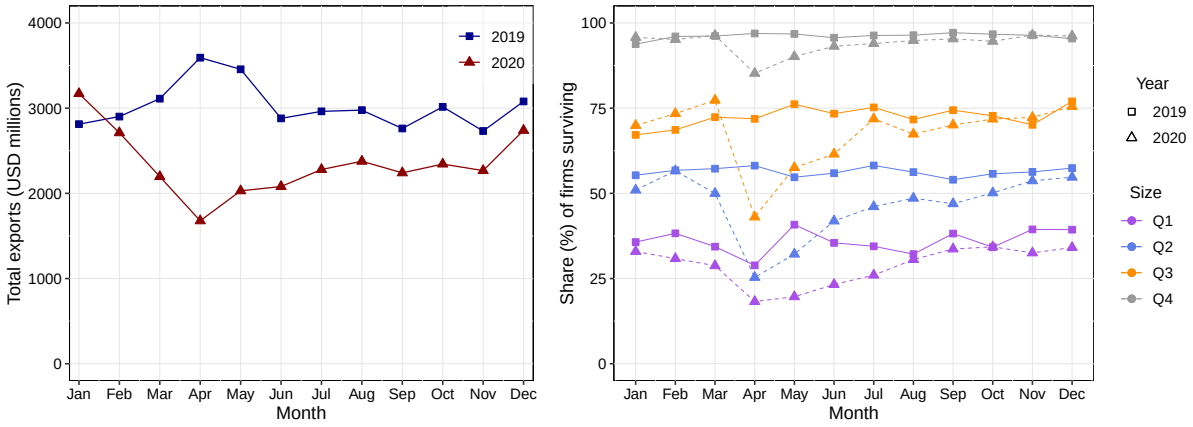


Figure Appx.1: The evolution of total exports (left) and the proportion of surviving exporting firms in year t among those exporting in year $t-1$ within size classes defined at $t-1$ (right). Firm size classes are based on the quartiles of the firms' exports (in ln) distribution in a given month.

In a typical month, large firms get a lion's share of the total exports. A regular pattern in looking at customs data is that more prominent exporters trade for many months and ship more frequently than smaller firms, which make only a few shipments. The right panel in Figure Appx.1 shows the proportion of surviving exporting firms in year t among those exporting in year $t-1$, by size classes defined at $t-1$. Comparing the figures for 2020 with those for 2019, it seems that the COVID-19 outbreak affected all firms regardless of their size. However, the effect looks proportionally stronger for small firms (Q1 and Q2 of the distribution). In contrast, larger firms are less affected and recover faster than the survival rates observed in 2019.

In the following of this Appendix C, we show the growth patterns of the number of exporters and export volumes between 2019 and 2020 (and, as a comparison, between 2018 and 2019) segmented by country of destination and product sector, offering further insights into the heterogeneous impacts of the COVID-19 pandemic on Colombian exports.

Figure Appx.2 shows, separately for the first and second quarter of a year, the percentage of firms that survive, enter or exit the export market and their corresponding shares of total exports. Thus, for a given quarter in 2019 and the corresponding quarter in 2020, we label each firm as *exiting* when it is present in 2019 and absent in 2020, *entrant* when it is absent in 2019 and present in 2020, and *surviving* when it is present in both years. We average the total value exported by each firm during the same quarter of two different years. Then, we sum the individual average value exported according to the firms' status. It turns out that surviving firms play an essential role in explaining total exports: they are around half of the total number of firms in both quarters and account for about 90% of the total export value. The volume lost, during the second quarter

of 2020, due to exiting firms is around 5% (assuming they would have exported in 2020 similar export volumes as observed in 2019). Entrant firms almost made up this 5% loss. Despite this, the firms' composition that participates in exports is very different. The number of existing firms in the second quarter of 2020 is much higher than the share of the first quarter of 2020 and the share of 2019 in the same period of the year.

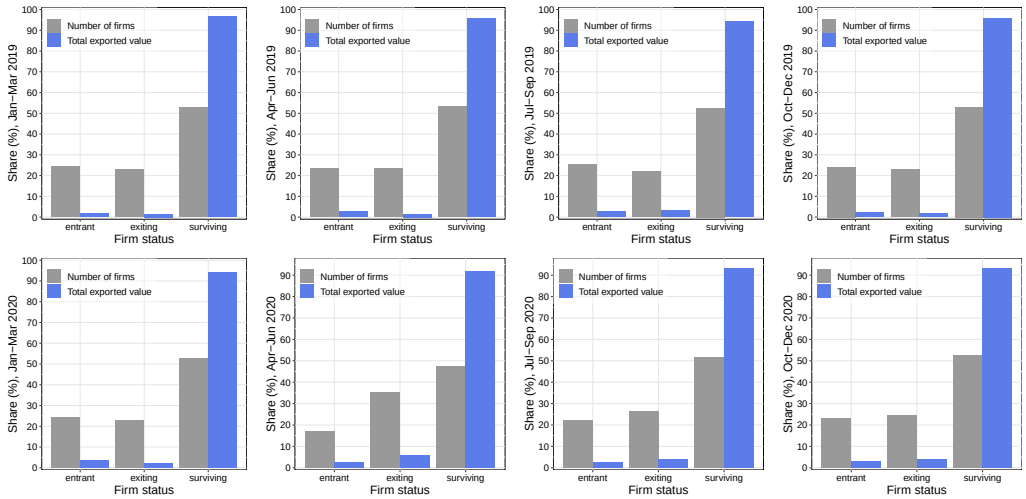


Figure Appx.2: Entry-exit dynamics of firms and total export value by firms that drop, enter or stay active, in 2019 (upper part of the figure) and in 2020 (bottom part of the figure) by quarters. Firm status is defined by looking at the previous year.

Figures Appx.3 and Appx.4 show the growth of the total number of exporters and the growth of the total volume of exports between 2019 and 2020, by country of destination and product sector. We consider the first and the second quarter separately, and we select destinations and product sectors that account for 80% of the total exporters in 2019. In both figures, the product sectors and the destinations are arranged by importance from top to bottom.

Figure Appx.3 shows that, compared to the first quarter of 2020, the second quarter of the year is characterized by a severe and pervasive drop in the number of exporting firms and the volume of exports in almost all the destinations reported. Figure Appx.5 shows that the same drop is not observed during the second quarter of 2019. During the third and fourth quarters of 2020, the value exported experienced more volatility than the number of firms. Nevertheless, the latter did not recover to the growth rates of the first quarter of the year.

Exports by product sectors in the second quarter of 2020 (see Figure Appx.4) reveal a generalized decrease in the number of exporting firms and trade values, while the first quarter exhibits very heterogeneous patterns. The sectors that appear to be more severely affected in the second quarter are Footwear (HS64), Leather Articles (HS42), Furniture (HS94), Books (HS49), Articles of Metal (HS83), Knitted and Not-Knitted Accessories (HS61-62), Vehicles (HS87) and Articles of Iron or Steel (HS73). Interestingly, these sectors are relatively more labor-intensive in Colombia, and therefore they could be susceptible to disruptions connected to social distancing. Finally, only for Coffee and Tea (HS08), Other textiles (HS63), and Jewelries (HS71) exports in value significantly grew in the second quarter. Instead, in terms of the number of exporting firms, no product sectors exhibit notable positive dynamics. During the third and fourth quarters of 2020, there is a rapid

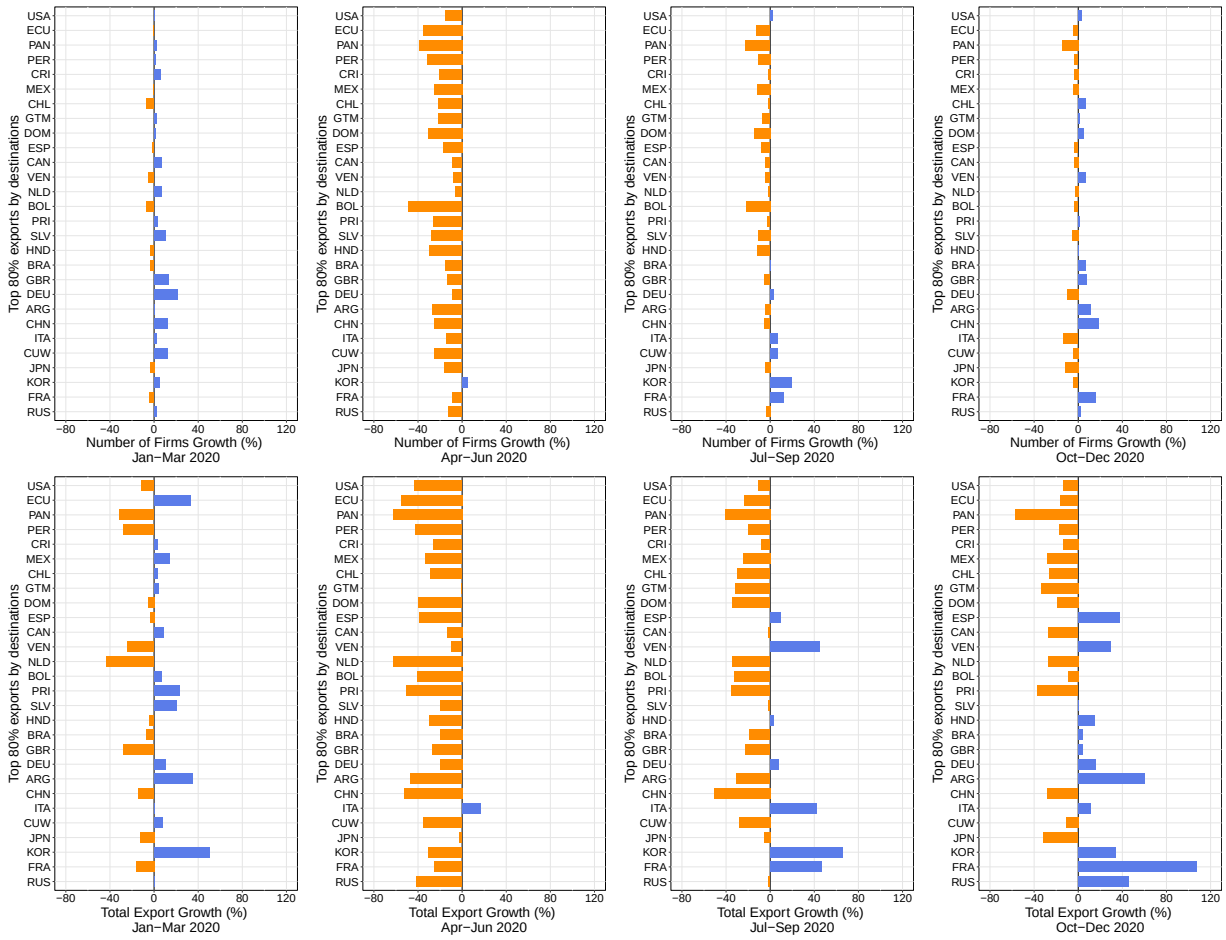


Figure Appx.3: The growth of the total number of exporters and the total value of exports by the destination country for the four quarters of 2020. Orange bars represent negative growth and blue bars positive growth. Destination countries are sorted from top to bottom accordingly to their importance in the share of the number of exporters in 2019.

back to normality in both the growth of value exported and in the number of exporters' growth rate by sector. Figure Appx.6 shows the same figures for 2019, suggesting that in periods without relevant shocks – such as the ones of the first quarter of 2020 – the changes in exports are also very heterogeneous, but there are not such extreme changes.

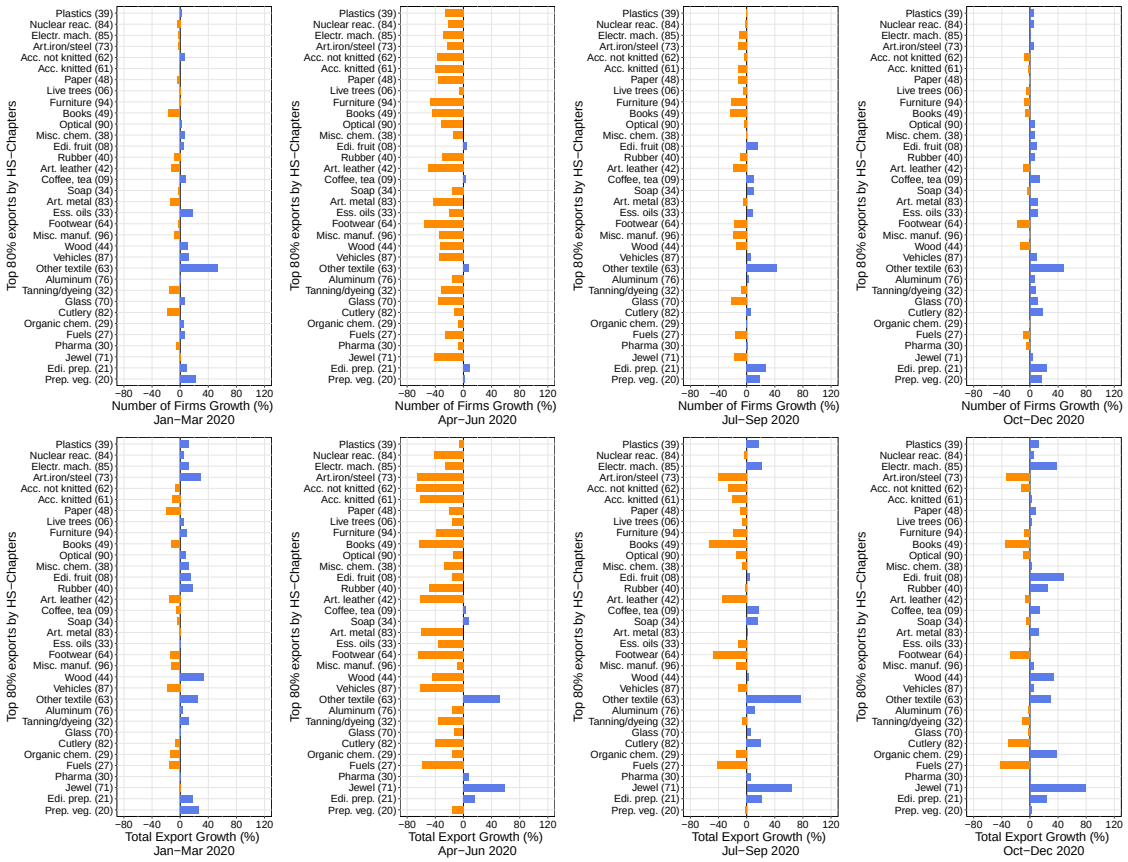


Figure Appx.4: The growth of the total number of exporters and the total value of exports by sector for the four quarters of 2020. Orange bars represent export reductions and blue bars positive export growth. Product sectors are sorted from top to bottom according to their importance in the share of the number of exporters in 2019. Product sectors correspond to the chapters of the HS code in parenthesis and the full name of the chapters is shortened to improve readability.

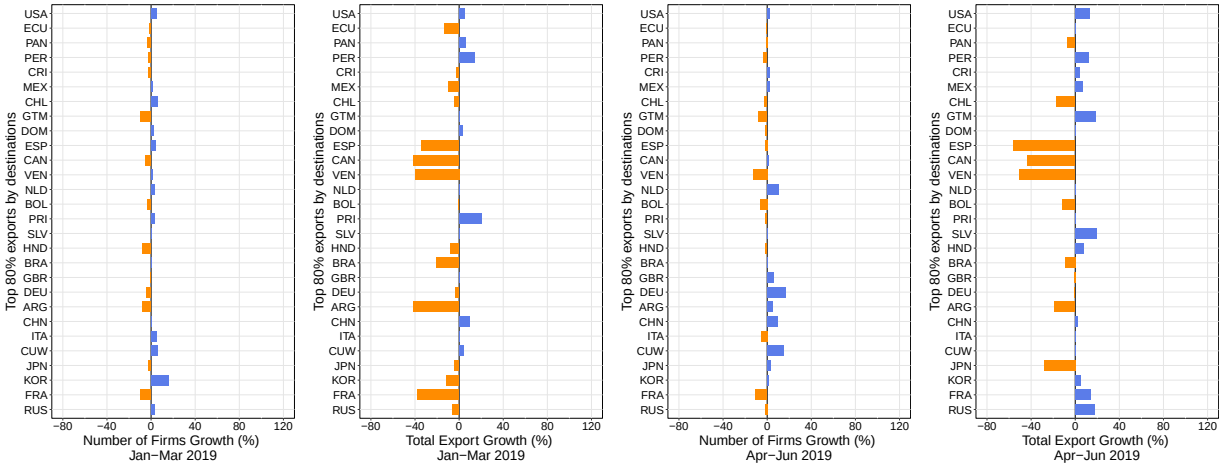


Figure Appx.5: The growth of the total number of exporters and the total value of exports by destination country for the first and the second quarters of 2019. Orange bars represent negative growth and blue bars positive growth. Destination countries are sorted from top to bottom accordingly with their importance in the share of number of exporters in 2019.

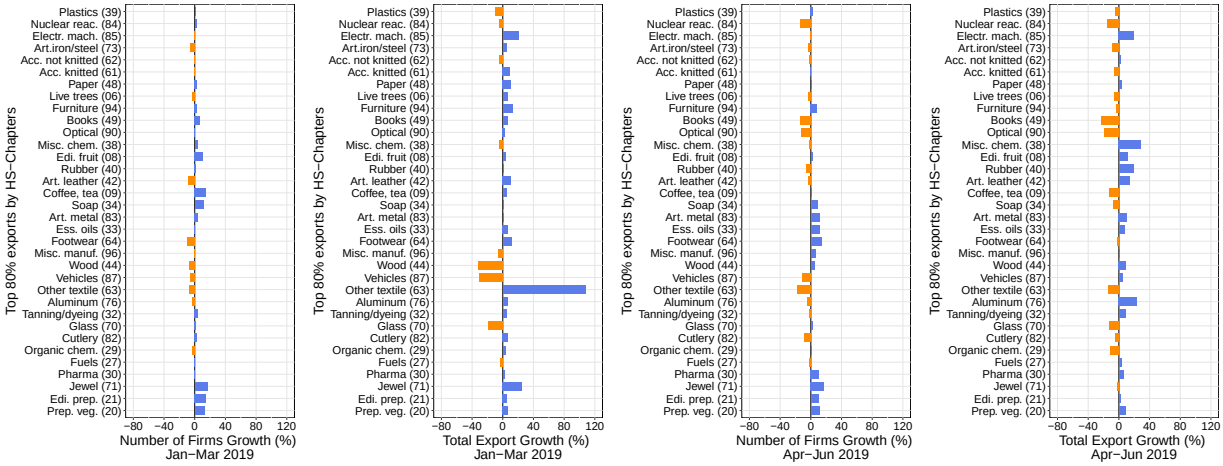


Figure Appx.6: The growth of the total number of exporters and the total value of exports by sector for the first and the second quarters of 2019. Orange bars represent export reductions and blue bars positive export growth. Product sectors are sorted from top to bottom accordingly with their importance in the share of number of exporters in 2019. Product sectors correspond to the chapters of the HS-code in parenthesis, the full name of the chapters is shortened to improve readability.

D Appendix: On the of convergence the T-Learner in our scenario

The estimator $\hat{\alpha}_i = \hat{f}^1(X_{i,t_s-1}) - \hat{f}^0(X_{i,t_s-1})$ is a T-learner estimator (Künzel et al., 2019), where the two potential outcome functions are estimated separately. While the T-learner lacks Neyman orthogonality, it has been heuristically observed in the literature that when the two potential outcome functions are estimated at heterogeneous rates, the overall estimation error of the T-learner is often practically dominated by the slower component (Künzel et al., 2019; Wager and Athey, 2018; Curth and Van der Schaar, 2021). In particular, Curth and Van der Schaar (2021) formally derive that the mean squared error (MSE) of the T-learner estimator is asymptotically bounded by the sum of the squared errors of the two nuisance regressions, implying that the overall convergence rate is determined by the slower component when the nuisance estimators converge at heterogeneous rates.

In the following we want to show that the only threaten to identification is the SUM.
It is easy to see that $\hat{\hat{\alpha}}$ identifies the CATE if $\mathbf{E}[\mathcal{E}^0_{t_s}(X_{t_s-1})] = 0$ from Eq. 7.
In order to analyze the same for $\hat{\alpha}$ we need to take an additional step and examine the differences between SUM and SAM through the lens of empirical process theory, as discussed in Tsybakov and Tsybakov (2009). Indeed, from Eq. 12, we observe that the main threat to identification arises from the difference in prediction errors between SUM and SAM when considering the estimator $\hat{\alpha}$. The objective is to determine whether the convergence rates to the true functions f^0 differ between SUM and SAM, and if so, to identify which of the two converges more slowly, thereby posing a greater threat to identification.
In the following discussion, $(X_{1,t-1}; Y_{1,t}), \ldots, (X_{i,t-1}; Y_{i,t}), \ldots (X_{N,t-1}; Y_{N,t}))$ are assumed to be i.i.d. [33]. In what follows we will assume that temporal dependence does not induce cross-sectional dependence in a "pooled cross-sectional" sense for providing the intuition [34]. Furthermore, we will focus on estimation tasks for penalized regressions. We assume a moderate sparsity s in

[33] Notice that the analysis can be easily extended to the case of independent but not identically distributed random variables following Singh (1975)

[34] The results can be extended to the panel case (Dehling and Philipp, 2002; Mirzaei, Kostic, Maurer, et al., Mirzaei et al.; Okui and Yanagi, 2019). Specifically, since our estimation procedures for SUM and SAM rely on different time periods, we should account for the fact that (X_t, Y_t) follows a weakly dependent stochastic process satisfying a strong mixing condition. Namely, the cross-sectional independence assumption should be enriched by an α-mixing condition of the form:

$$\alpha_k = \sup_{A \in \mathcal{F}_t, B \in \mathcal{F}_{t-k}} |P(A \mid B) - P(A)| \to 0 \quad \text{as } k \to \infty. \tag{18}$$

where $\mathcal{F}_t$ is the sigma-algebra generated by (X_t, Y_t). $P(A)$ represents the probability that a firm's characteristics and outcome at time t, denoted as $(X_{i,t}, Y_{i,t})$, fall into a given set (S, T) without conditioning on past values. $P(A \mid B)$ represents the probability that $(X_{i,t}, Y_{i,t})$ falls into (S, T), given that in an earlier period, the firm's characteristics and outcome were in a different set (S', T'), i.e., $(X_{i,t-k}, Y_{i,t-k}) \in (S', T')$. This ensures that temporal dependence decays over time as k –the lag– grows, allowing empirical process results to hold even when observations are not strictly i.i.d. (Drees and Rootzén, 2010). Moreover, it is possible to show that a form of the Glivenko-Cantelli theorem is valid also for panel data (Okui and Yanagi, 2019). The major results of empirical process theory are reproduced if, additionally to the above, we assume that (X_t, Y_t) follow a weakly stationary process and that the estimation bias in time-dependent quantities does not dominate asymptotic results.

order to extend our results to LASSO, Ridge and ElasticNet types of models. In particular $||f^0(\cdot)||_n, ||f^1(\cdot)||_n \geq n^{-1/w}$ for $0 < w < 2$. The reader is referred to Koltchinskii (2011) for further details. Finally, notice that an advantage of the following study is that it adapts to nonparametric models.

Given the mentioned assumptions and for the sake of clarity in the notation, we will refer to $(X_{i,t-1}, Y_{i,t})$ as simply (X_i, Y_i) so that we can rewrite the potential outcomes as $Y_i^d = f^d(X_i)$. In what follows, we will assume that the variables at our disposal are sub-gaussian [35]. We now define the classes associated with SAM and SUM as

$$\mathcal{F}_{\text{SAM}} = \{f : X \to [0,1] \text{ predicting } Y^1 \text{ from observed } (X_i, Y_i^1)\}$$
$$\mathcal{F}_{\text{SUM}} = \{f : X \to [0,1] \text{ predicting } Y^0 \text{ from } X_i \text{ only (no observed labels)}\}$$

We briefly investigate the complexity of the latter classes in what follows. As aforementioned, given Assumption 2, it is possible in our context to use $\hat{f}^0_{2019}$ on X_{2020} to estimate Y^0_{2020}, whose labels are, however, unknown. Hence, although the estimation of $\hat{f}^0_{2019}$ is achieved through supervised learning, the validity of $\hat{Y}^0_{2020}$ is guaranteed only if Assumption 2 holds. This is because the latter task is inherently unsupervised, as the labels of $\hat{Y}^0_{2020}$ remain unobserved. In other words, while the SAM is directly supervised by the observed data (X_i, Y_i^1), its function class $\mathcal{F}_{\text{SAM}}$ is constrained to fit the empirical distribution of observed outcomes, the SUM must approximate counterfactuals without direct labels, meaning that its function space $\mathcal{F}_{\text{SUM}}$ must be broader to accommodate all possible mappings from X to Y^0. More formally, if the function classes are parameterized by some hypothesis space Θ, the cardinality of $\mathcal{F}$ can be linked to the dimension of Θ. Suppose: $\mathcal{F}_{\text{SAM}}$ is parameterized by Θ_{SAM}. and $\mathcal{F}_{\text{SUM}}$ is parameterized by Θ_{SUM}. Since the SUM does not observe Y^0 directly, its parameter space must include extra degrees of freedom to account for unobserved variability. This means:

$$\dim(\Theta_{\text{SUM}}) > \dim(\Theta_{\text{SAM}}).$$

The latter is strictly related to the concept of complexity of $\mathcal{F}_{\text{SUM}}$ and $\mathcal{F}_{\text{SAM}}$.

The complexity of a function class is quantified by its metric entropy, given by $\log N(\varepsilon, \mathcal{F}, L_1(Q))$, where $N(\varepsilon, \mathcal{F}, L_1(Q))$ is the covering number, i.e., the minimal number of functions required to approximate all $f \in \mathcal{F}$ within an error ε. Since SUM must account for a broader range of possible counterfactual relationships, its function space is necessarily larger, implying that $N(\varepsilon, \mathcal{F}_{\text{SUM}}, L_1(Q)) > N(\varepsilon, \mathcal{F}_{\text{SAM}}, L_1(Q))$, and taking the logarithm yields

$$\log N(\varepsilon, \mathcal{F}_{\text{SUM}}, L_1(Q)) > \log N(\varepsilon, \mathcal{F}_{\text{SAM}}, L_1(Q))$$

Since we assumed moderate sparsity, we can be confident that the function classes of both SUM and SAM are not excessively complex, ensuring that theoretical bounds can still be established.

[35] The latter assumption is not too much restrictive as it includes all normal random variables, all bounded random variables and all random variables for which even moments exist and satisfy $E[X^{2k}] \leq \frac{(2k)!}{2^k k!}\xi^{2k}$ for $k = 1, 2, 3, \ldots$ and some parameter $\xi \geq 0$.

In the analysis of empirical risk minimization under complexity constraints, the convergence rate of the estimator $\hat{f}$ to the true function f_0 is governed by the metric entropy of the function class $\mathcal{F}$. Given the metric entropy bound $\log N(\delta, \mathcal{F}, \|\cdot\|) \leq C\delta^{-w}$, for $C > 0$, classical results from empirical process theory (see Sadhanala and Tibshirani (2019)) establish that the expected rate of convergence satisfies:

$$\|\hat{f} - f_0\|_n^2 = O(n^{-2/(2+w)}).$$

Applying this result to the function classes of SUM and SAM, we recover their respective entropies as satisfying:

$$\log N(\delta, \mathcal{F}_{\text{SUM}}, \|\cdot\|) \geq C_{\text{SUM}}\delta^{-w_{\text{SUM}}}, \quad \log N(\delta, \mathcal{F}_{\text{SAM}}, \|\cdot\|) \leq C_{\text{SAM}}\delta^{-w_{\text{SAM}}}.$$

Since SUM estimates counterfactuals without direct supervision, it must approximate a wider range of functional relationships, leading to a function class $\mathcal{F}_{\text{SUM}}$ with a larger covering number and a lower complexity exponent $w_{\text{SUM}} < w_{\text{SAM}}$. Consequently, the corresponding rates of convergence are:

$$\|\hat{f}_{\text{SUM}} - f_0\|_n^2 = O(n^{-2/(2+w_{\text{SUM}})}), \quad \|\hat{f}_{\text{SAM}} - f_0\|_n^2 = O(n^{-2/(2+w_{\text{SAM}})}). \tag{19}$$

Since $w_{\text{SUM}} < w_{\text{SAM}}$, it follows that:

$$n^{-2/(2+w_{\text{SUM}})} \gg n^{-2/(2+w_{\text{SAM}})},$$

which formally confirms that the convergence rate of SUM is slower than that of SAM. We can summarize all of this in a Lemma.

Lemma 1 (Identification Threat from the SUM under regularized ML)**.** *Consider the set up of Section 2. Let $\hat{f}_{SAM}$ and $\hat{f}_{SUM}$ denote the estimators obtained via penalized regression (e.g., LASSO, Ridge) with moderate sparsity s, where inputs are i.i.d. sub-Gaussian. Let $\mathcal{F}_{SAM}$ and $\mathcal{F}_{SUM}$ be the corresponding function classes, with metric entropies satisfying:*

$$\log N(\delta, \mathcal{F}_{SUM}, \|\cdot\|) \geq C_{SUM}\delta^{-w_{SUM}},$$
$$\log N(\delta, \mathcal{F}_{SAM}, \|\cdot\|) \leq C_{SAM}\delta^{-w_{SAM}}, \quad \textit{for } 0 < w_{SUM} < w_{SAM} < 2.$$

Then the mean squared convergence rates of the two estimators are given by:

$$\|\hat{f}_{SUM} - f^0\|_n^2 = O(n^{-2/(2+w_{SUM})}), \quad \|\hat{f}_{SAM} - f^1\|_n^2 = O(n^{-2/(2+w_{SAM})}).$$

Consequently, the SUM estimator converges more slowly than the SAM estimator with

$$n^{-2/(2+w_{SUM})} \gg n^{-2/(2+w_{SAM})},$$

Proof Let $\mathcal{F}$ be a function class such that $\log N(\delta, \mathcal{F}, \|\cdot\|_n) \leq C\delta^{-w}$ for some $C > 0$ and $w \in (0, 2)$. Then standard empirical process theory (see Koltchinskii (2011)) yields the convergence

rate:

$$\|\hat{f} - f_0\|_n^2 = O(n^{-2/(2+w)}).$$

Applying this to SAM and SUM yields:

$$\|\hat{f}_{\text{SAM}} - f^1\|_n^2 = O(n^{-2/(2+w_{\text{SAM}})}), \quad \|\hat{f}_{\text{SUM}} - f^0\|_n^2 = O(n^{-2/(2+w_{\text{SUM}})}).$$

Since $w_{\text{SUM}} < w_{\text{SAM}}$, we have:

$$\frac{2}{2 + w_{\text{SUM}}} > \frac{2}{2 + w_{\text{SAM}}},$$

so the exponent in the rate for SUM is smaller, and the convergence rate is slower. Hence,

$$\|\hat{f}_{\text{SUM}} - f^0\|_n^2 \gg \|\hat{f}_{\text{SAM}} - f^1\|_n^2.\blacksquare$$

E Appendix: Robustness checks with panel cross validation and alternative ML methods

In this appendix, we report the results of a series of robustness checks that pursue a threefold goal: (1) the robustness of our scenario in the case of a longer panel before the shock; (ii) the robustness of our scenario when more machine learning algorithms than those given in the main text are used; (iii) the robustness of our scenario when different performance metrics are used in the validation step.

We start by combining the datasets that contain the relevant information on Colombian companies from 2014 to 2018[36]. Since we are dealing with a panel, the validation process is more complicated than the strategy chosen in the main text. We consider two possible strategies for splitting the panel:

1. The first possibility (*panel-split 1*) is to repeat the same approach as for Fabra et al. (2022). Namely, take the features x_{t+k} to predict y_{t+k} in the training, where $k = 1, 2, \ldots, K$ is rolling and K is the size of the training. Then use x_{K+1} as validation to predict y_{K+1} using the trained function.

   To clearly distinguish between the structure of the dataset and the actual time of the modeled behavior, we introduce the following notation. Let s denote the *observation year*, i.e. the year in which observations are recorded in the dataset. Let $t = s + 1$ denote the *effective year*, which represents the time period to which the outcome variable Y_t refers. In our data set, the outcome Y_t, which was recorded in the year s, reflects the firm's export behavior in the year t. Instead, $t = s$ applies to the characteristics. This means that we observe tuples $(X_{t=s}, Y_t)$,

[36] To prepare the data set for the training and validation of the model, the data is first split into predictor features and target variables for the training and validation set. Irrelevant columns, such as identifiers and date-related fields, are removed from the predictors, while the target variable remains unchanged. Categorical features are converted to numeric form using binary indicator variables and any missing values in the data are filled in to ensure consistency. The numeric characteristics (except for the binary indicators) are then transformed to obtain a consistent scale for all variables, usually by adjusting for a common central tendency and dispersion. This standardization step ensures that all continuous features contribute equally during model training.

where Y_t encodes the behavior in the year $s+1$. For the sake of simplicity, we consider a generic prediction task with a fixed target year t (e.g. 2016) and a fixed calendar month (e.g. January). In this context, the training data is constructed from firms observed in the two previous years $s-2$ and $s-1$, i.e. we use observations where the features X_{s-2}, X_{s-1} and outcomes Y_{s-1}, Y_s. Since the export behavior in s is encoded in Y_{t-1}, recorded in $s-1$ and accordingly linked to firm characteristics from the year $s-1$, our data would in principle allow us to learn a mapping from X_s to Y_t. However, the model of Fabra et al. (2022) – requires learning a mapping from X_s to Y_s in both training and validation. To this end, we have replaced the target Y_t with the export behavior from the same year in which the input features are recorded, i.e. Y_s. In practice, this means that the monthly values of Y_t are overwritten with those of $Y_s := Y_{t-1}$, forcing a common future behavior in the training data. This trains the model to predict an imputed target — it learns the mapping $X_s \to \widetilde{Y}_t$, where the tilde means that Y_t has been replaced by Y_s. The validation features correspond to X_t for a fixed calendar month (e.g. January). The corresponding validation targets are Y_t[37]
To summarize, the model is trained on $X_s \to \widetilde{Y}_t$ and evaluated on $X_t \to Y_t$[38].
The procedure described above can be recursively generalized [39].

2. The second strategy (*panel-split 2*) stays with the current version of the dataset, without distinguishing between the observation year and the effective year, as in *panel-split 1*, noting that (1) the trained model should predict the export results for the following year and (2) the implementation is slightly different from that of Fabra et al. (2022). To be consistent with Fabra et al. (2022), our training procedure rolls forward in time due to the way Y was recorded in our dataset, but the variable used as Y_{train} corresponds to what Fabra et al. (2022) labeled as $Y_{validation}$.
This strategy can be further subdivided into alternative data splitting schemes that fall into two broad categories: those that exploit the temporal dimension of the dataset and those that exploit the structure of the panel (i.e., both the cross-sectional and temporal dimensions). Since the former approach is more commonly used in our reference literature (Cerqua and Letta, 2020; Fabra et al., 2022), we follow this approach in our analysis. Accordingly, we describe the latter here only briefly so as not to overburden the reader. Nevertheless, we emphasize that a more comprehensive approach to training and cross-validation in panel data should ideally also make use of the cross-sectional dimension. We leave this extension to future research and refer the reader to the online appendix for a possible direction in this regard.

[37] In order to create a consistent validation set, the inputs X_t recorded in $s=t$ and the targets Y_t recorded in $s-1$ are merged using an inner join on firm identifier and the calendar month.

[38] As an illustration, consider the case where the target year is $t = 2016$, which corresponds to an observation year $s = 2015$, with data recorded in a fixed month (e.g. January). The training data is drawn from the two previous years $s = 2014$ and $s = 2015$ using inputs X_s and targets Y_t. Following Fabra et al. (2022), the target Y_t is replaced by Y_s, i.e. by the export behavior from the same year as the input features. Specifically, Y_{2016} is overwritten with monthly values from Y_{2015}, which leads to an imputed target $\widetilde{Y}_{2016}$. The model is thus trained on $X_{2015} \to \widetilde{Y}_{2016}$ and then applied to predict Y_{2017} from X_{2016}.

[39] For example, for 2017, we keep 2015 and 2016 in the training set, impute the corresponding values of the export (i.e. Y of 2014 to 2015, Y of 2015 to 2016) and use X_{2017} to predict Y_{2017} (i.e. Y corresponding to 2016).

The first strategy, which utilizes the temporal dimension, slightly modifies the Fabra et al. (2022) procedure by performing a grid search approach. We call this strategy *panel-split 2 (i)*. The idea is to choose *ex-ante* a set of, say K_{ML}, possible combinations of hyperparameters for the different ML algorithms (the subscript indicates that the number of hyperparameter combinations depends on the ML algorithm). The process consists of training each of the K_{ML} models in turn, following a "year-forward chaining" approach. Let $\{t_0, t_1, \ldots, t_T\}$ represent the available years in the dataset, where $t_0 := t - L$ is the start year and T is the end year. At each step s, the training set includes all years $\{t_0, t_1, \ldots, t_{s-1}\}$ to predict the target variable in year t_s, and the validation set consists of t_s to predict t_{s+1}. For example, in the first iteration, the model is trained with the input features $\mathbf{X}_{t_0}$ and the target variable $\mathbf{y}_{t_1}$ and then validated with the features $\mathbf{X}_{t_1}$ to predict $\mathbf{y}_{t_2}$. The hyperparameters are selected by minimizing the RMSE of the validation predictions. In the following iterations, the training set is expanded step by step: in iteration s the model is trained on $\{\mathbf{X}_{t_0}, \mathbf{X}_{t_1}, \ldots, \mathbf{X}_{t_{i-1}}\}$ to predict $\{\mathbf{y}_{t_1}, \mathbf{y}_{t_2}, \ldots, \mathbf{y}_{t_s}\}$ and validated using $\mathbf{X}_{t_s}$ to predict $\mathbf{y}_{t_{s+1}}$. This chaining process is continued until the last year t_T, recording the RMSEs (AUCs and BACCs) for each validation step. The hyperparameters that result in the lowest overall RMSE across all validation sets are selected as optimal.

Algorithm 1 shows an example with two machine learning models, LASSO and Ridge regression, using a "year-forward chaining" approach, where each model is evaluated for four different regularization strengths ($\lambda \in \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$). Assume that the data set spans three years (2014, 2015, 2016) and predictions are made for the target variable of the following year.

Algorithm 1 Year-Forward Chaining for Hyperparameter Selection

Require: Dataset $\{(\mathbf{X}_t, \mathbf{y}_{t+1})\}_{t=t_0}^{t_T}$;
ML models $\mathcal{M} = \{\text{LASSO}, \text{Ridge}\}$;
Hyperparameters $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$

for each model $m \in \mathcal{M}$ **do**
  for each $\lambda \in \Lambda$ **do**
    for iteration $i = 1$ to $T - 1$ **do**
      Define training years: $\{t_0, t_1, \ldots, t_{i-1}\}$
      Define validation year: t_i
      Train: Fit model m with λ on

$$\left\{(\mathbf{X}_{t_0}, \mathbf{y}_{t_1}), \ldots, (\mathbf{X}_{t_{i-1}}, \mathbf{y}_{t_i})\right\}$$

      Validate: Predict $\hat{\mathbf{y}}_{t_{i+1}}$ using $\mathbf{X}_{t_i}$ (if $\mathbf{y}_{t_{i+1}}$ is available)
      Evaluate: Compute RMSE, AUC, BACC between $\hat{\mathbf{y}}_{t_{i+1}}$ and $\mathbf{y}_{t_{i+1}}$
      Store performance metrics
    end for
  end for
end for
for each model $m \in \mathcal{M}$ **do**
  Aggregate validation metrics across iterations for each $\lambda \in \Lambda$
  Select $\lambda_m^* = \arg\min_\lambda \text{RMSE}_{\text{cumulative}}$
end for
return Optimal hyperparameters $\{\lambda^*_{\text{LASSO}}, \lambda^*_{\text{Ridge}}\}$

In the first iteration, the models are trained using the data from 2014 ($\mathbf{X}_{2014}$ as input and $\mathbf{y}_{2015}$ as target) and validated on 2015 ($\mathbf{X}_{2015}$) to predict $\mathbf{y}_{2016}$. For each combination of model (LASSO or Ridge) and λ, the predictions are evaluated using the RMSE, AUC and BACC and the results are recorded. Then the training set is extended by 2014 and 2015 ($\mathbf{X}_{2014}, \mathbf{X}_{2015}$ as input and $\mathbf{y}_{2015}, \mathbf{y}_{2016}$ as target), while 2016 ($\mathbf{X}_{2016}$) is used for validation to predict $\mathbf{y}_{2017}$. Again, the RMSE (AUC and BACC) is calculated for each model and each combination of λ and the results are recorded. At the end of these iterations, the RMSE (AUC, BACC) values from all validation steps are aggregated for each model and each λ value. The best hyperparameters for LASSO and Ridge are determined by selecting the λ values that minimize the cumulative RMSE across all validation steps.

The procedure we used in our exercise is based on *panel-split 2 (i)* with the only difference that the validation year is set to t_T (in our case the dataset containing the X_{2017} and Y_{2018}). We call this strategy *panel-split 2 (ii)*. In iteration s, the model is trained on $\{\mathbf{X}_{t_s}, \ldots, \mathbf{X}_{t_{T-2}}, \mathbf{X}_{t_{T-1}}\}$ to predict $\{\mathbf{y}_{t_1}, \mathbf{y}_{t_2}, \ldots, \mathbf{y}_{t_s}\}$ and validated with $\mathbf{X}_{t_s}$ to predict $\mathbf{y}_{t_{s+1}}$. This chaining process

continues until the last year t_T, recording the RMSEs for each validation step.
Once the optimal λs are calculated for each month and each training size via the procedure described above, we then estimate the RMSEs for each training size including the dataset containing X_{2017} and Y_{2018}[40], in the training set with the selected λs (*model estimation*). This makes it possible to learn the optimal coefficients of the model based on the validated λs. We finally use the estimated model to test the predictions obtained for Y_{2019} using X_{2018} (and save the test errors) (*Y-SUM placebo for 2019*).

The use of *panel-split 1* leads to unnecessary complexity, especially because the dependent variable in our data set corresponds to the following year, which makes practical implementation more difficult. In addition, the approach of Fabra et al. (2022) iteratively varies the validation year across different splits. In our view, this strategy reduces the predictive power of the T-learner, whose main goal is to accurately estimate the SUM for counterfactual inference. In contrast, setting the validation set to Y_{2018} provides a more stable and targeted framework. Given the temporal proximity between 2018 and 2019, using Y_{2018} for validation increases the likelihood that the selected hyperparameters generalize well to Y_{2019}, improving the estimation of the SUM function relevant for the counterfactual prediction.
For these reasons, we have opted for a modified version of the strategy in Fabra et al. (2022), in which the validation year remains fixed at Y_{2018}. This consideration has led us to adopt *panel-split 2 (ii)*, which addresses the limitations associated with *panel-split 1*. The reason why we used *panel-split 2 (ii)* and not *panel-split 2 (i)* is explained in more detail in the online appendix. The results using *panel-split 1* are available on request.

Results for panel case with alternative ML techniques

In this section, we report on the results we obtain when we apply the *panel-split 2 (ii)* approach.

Table Appx.3 shows the hyperparameter grid used for each ML method.

[40] This effectively means that the training set was constructed as follows:

- Training size 1: the training data contained X_{2017} and Y_{2018};
- Training size 2: the training data contained X_{2016}, X_{2017} and Y_{2017}, Y_{2018};
- Training size 3: the training data contained $X_{2015}, X_{2016}, X_{2017}$ and $Y_{2016}, Y_{2017}, Y_{2018}$.

The training size 4 was not included coherently with the validation, whose maximum size is necessarily 3

Model	Hyperparameters
LASSO	*lambda*: Values generated by combining: - Coarse grid: $\log_{10}$-spaced values in range $[10^{-4}, 10^{2}]$ with 20 points, - Fine grid: $\log_{10}$-spaced values in range $[10^{-1}, 10^{1}]$ with 50 points, - Unique combination of both grids.
Ridge	Same as LASSO (*lambda* values)
RandomForest	*n_estimators*: [25, 100, 200, 500], *max_features*: ['sqrt', 'log2', None], *max_depth*: [None, 3, 5, 7, 10], *max_leaf_nodes*: [3, 6, 9], *min_samples_split*: [2, 8]
XGBoost	*subsample*: [0.4, 0.5, 0.7, 0.9], *learning_rate*: [0.05, 0.1, 0.3, 0.5, 0.9], *max_depth*: range(3, 8), *n_estimators*: [100, 200]
SVM	*C*: $\log_{10}$-spaced values in range $[10^{-2}, 10^{3}]$, *kernel*: ['linear', 'rbf'], *gamma*: $\log_{10}$-spaced values in range $[10^{-3}, 10^{1}]$

Table Appx.3: Tuned hyperparameters for different machine learning models.

Results are presented for training sets created from one, two or three years prior to a fixed validation year. For further methodological details, the reader is referred to the previous section. A more detailed analysis of how the performance of the regularization methods changes with different training set sizes can be found in the online appendix.

Regularization techniques: LASSO and Ridge The following figure displays the difference Y-SUM with different training sizes using RMSE as validation criterion:

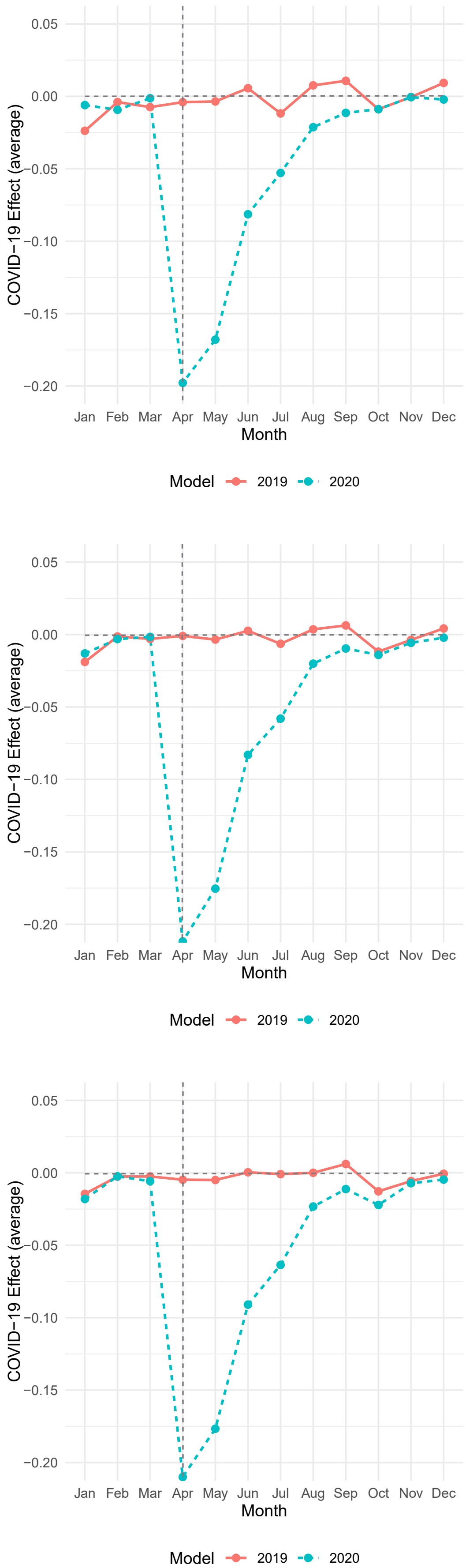


Figure Appx.7: Y-SUM using Logit-LASSO-CV for different training sizes: top: training size 1, mid: training size 2, bottom: training size 3

It is worth noting that in the main text, using standard cross-sectional cross-validation, we observe a very similar distribution of test errors for the period January to March 2020. This same distribution is also obtained when applying the SAM during the same period, indicating that the issue of bias in specific regions of the distribution of treatment effects has been effectively addressed.

SVM The same exercise has been repeated for SVM.
Figure Appx.8 summarizes the results for the placebo in 2019 when SVM is used.

Tree based methods (RF, XGB, BART) The first of tree based methods presented is RF. The hyperparameters for RF are chosen according to Probst et al. (2019). Figure Appx.9 summarizes the results for the placebo in 2019 when RF is used:

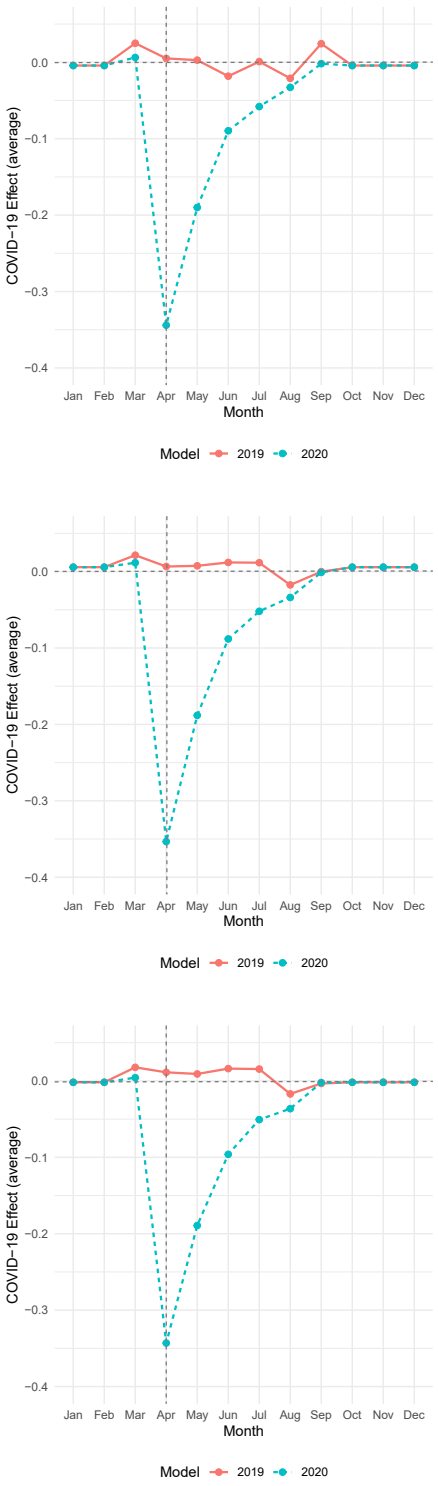


Figure Appx.8: Y-SUM using SVM for different training sizes for 2019 and 2020: top: training size 1, mid: training size 2, bottom: training size 3

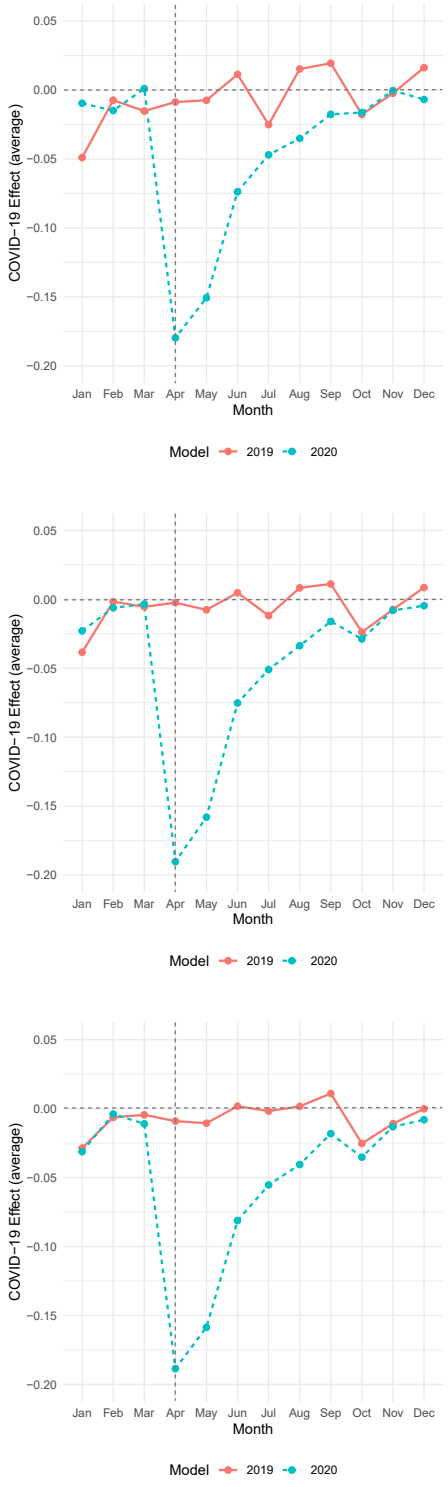


Figure Appx.9: Y-SUM using RF for different training sizes for 2019 and 2020. Top: training size 1, mid: training size 2, bottom: training size 3.

XGB results The following are the results for XGB:

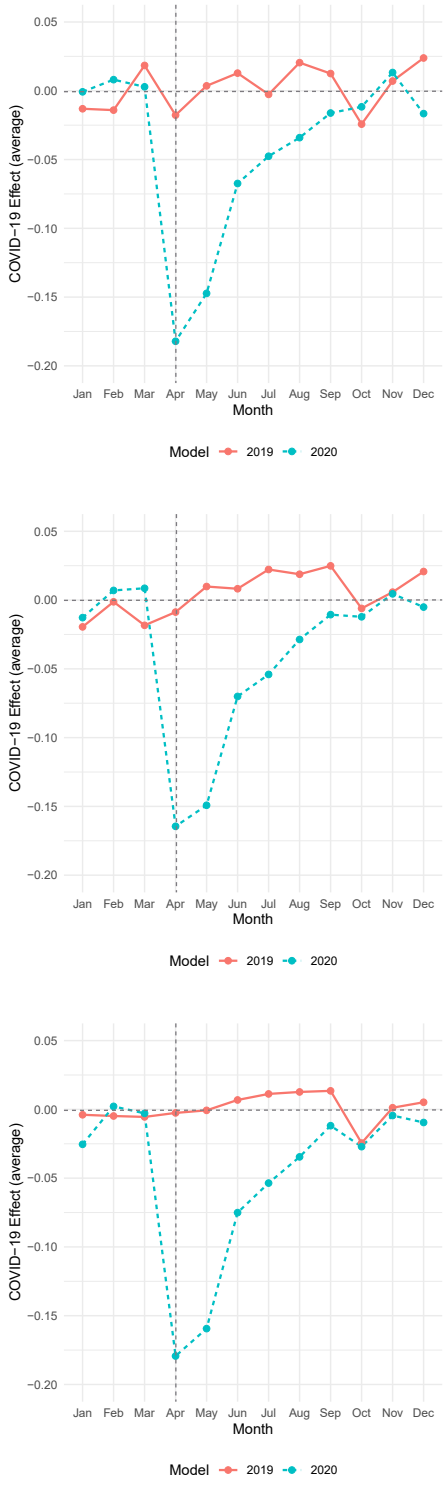


Figure Appx.10: Y-SUM using XGB for different training sizes for 2019 and 2020. Top: training size 1, mid: training size 2, bottom: training size 3.

Remark Figures Appx.7-Appx.10 present the results for both the placebo case—where no COVID-19 occurred—and for the year 2020. The estimator $\hat{\hat{\alpha}}$ was used because incorporating additional lagged years can only improve the counterfactual prediction (i.e., the SUM) in our context. Therefore, this is the only estimator tested. The results are consistent with the main analysis and fall within the same range of values. No significant effects were detected in the placebo

case. Furthermore, not only are the CATE estimates aligned with those reported in the main text, but the predictive performance of the various machine learning estimators is also comparable. A detail of the performance of the results of the various ML methods is Tables Appx.4 and Appx.5.

Month	Valid. SVM	Test SVM	Valid. RF	Test RF	Valid. XGB	Test XGB	Valid. LASSO	Test LASSO	Valid. Ridge	Test Ridge	Valid. Logit	Test Logit
	tr.size 1: 0.437	tr.size 1: 0.434	tr.size 1: 0.438	tr.size 1: 0.443	tr.size 1: 0.423	tr.size 1: 0.402	tr.size 1: 0.416	tr.size 1: 0.409	tr.size 1: 0.419	tr.size 1: 0.416	tr.size 1: 0.423	tr.size 1: 0.425
Jan	tr.size 2: 0.435	tr.size 2: 0.433	tr.size 2: 0.436	tr.size 2: 0.440	tr.size 2: 0.424	tr.size 2: 0.401	tr.size 2: 0.417	tr.size 2: 0.406	tr.size 2: 0.411	tr.size 2: 0.410	tr.size 2: 0.413	tr.size 2: 0.415
	tr.size 3: 0.433	tr.size 3: 0.430	tr.size 3: 0.437	tr.size 3: 0.438	tr.size 3: 0.423	tr.size 3: 0.400	tr.size 3: 0.415	tr.size 3: 0.404	tr.size 3: 0.412	tr.size 3: 0.405	tr.size 3: 0.424	tr.size 3: 0.426
	tr.size 1:0.453	tr.size 1: 0.443	tr.size 1: 0.431	tr.size 1: 0.437	tr.size 1: 0.414	tr.size 1: 0.408	tr.size 1: 0.418	tr.size 1: 0.411	tr.size 1: 0.413	tr.size 1: 0.415	tr.size 1: 0.421	tr.size 1: 0.424
Feb	tr.size 2: 0.448	tr.size 2: 0.445	tr.size 2: 0.429	tr.size 2: 0.434	tr.size 2: 0.410	tr.size 2: 0.406	tr.size 2: 0.415	tr.size 2: 0.405	tr.size 2: 0.403	tr.size 2: 0.405	tr.size 2: 0.416	tr.size 2: 0.420
	tr.size 3: 0.423	tr.size 3: 0.425	tr.size 3: 0.428	tr.size 3: 0.434	tr.size 3: 0.410	tr.size 3: 0.403	tr.size 3: 0.410	tr.size 3: 0.406	tr.size 3: 0.405	tr.size 3: 0.409	tr.size 3: 0.417	tr.size 3: 0.419
	tr.size 1:0.422	tr.size 1: 0.421	tr.size 1: 0.434	tr.size 1: 0.434	tr.size 1: 0.424	tr.size 1: 0.399	tr.size 1: 0.401	tr.size 1: 0.407	tr.size 1: 0.411	tr.size 1: 0.409	tr.size 1: 0.420	tr.size 1: 0.425
Mar	tr.size 2: 0.420	tr.size 2: 0.418	tr.size 2: 0.430	tr.size 2: 0.433	tr.size 2: 0.413	tr.size 2: 0.396	tr.size 2: 0.414	tr.size 2: 0.406	tr.size 2: 0.408	tr.size 2: 0.411	tr.size 2: 0.421	tr.size 2: 0.424
	tr.size 3: 0.411	tr.size 3: 0.420	tr.size 3: 0.430	tr.size 3: 0.431	tr.size 3: 0.407	tr.size 3: 0.396	tr.size 3: 0.398	tr.size 3: 0.500	tr.size 3: 0.409	tr.size 3: 0.412	tr.size 3: 0.418	tr.size 3: 0.421
	tr.size 1: 0.455	tr.size 1: 0.443	tr.size 1: 0.430	tr.size 1: 0.489	tr.size 1: 0.404	tr.size 1: 0.472	tr.size 1: 0.418	tr.size 1: 0.402	tr.size 1: 0.404	tr.size 1: 0.406	tr.size 1: 0.413	tr.size 1: 0.420
Apr	tr.size 2: 0.443	tr.size 2: 0.431	tr.size 2: 0.430	tr.size 2: 0.492	tr.size 2: 0.408	tr.size 2: 0.469	tr.size 2: 0.403	tr.size 2: 0.402	tr.size 2: 0.407	tr.size 2: 0.409	tr.size 2: 0.417	tr.size 2: 0.421
	tr.size 3: 0.433	tr.size 3: 0.428	tr.size 3: 0..429	tr.size 3: 0.491	tr.size 3: 0.401	tr.size 3: 0.469	tr.size 3: 0.412	tr.size 3: 0.398	tr.size 3: 0.399	tr.size 3: 0.401	tr.size 3: 0.411	tr.size 3: 0.416
	tr.size 1: 0.439	tr.size 1: 0.446	tr.size 1: 0.431	tr.size 1: 0.472	tr.size 1: 0.418	tr.size 1: 0.445	tr.size 1: 0.414	tr.size 1: 0.406	tr.size 1: 0.407	tr.size 1: 0.410	tr.size 1: 0.417	tr.size 1: 0.422
May	tr.size 2: 0.440	tr.size 2: 0.444	tr.size 2: 0.429	tr.size 2: 0.473	tr.size 2: 0.416	tr.size 2: 0.445	tr.size 2: 0.415	tr.size 2: 0.404	tr.size 2: 0.408	tr.size 2: 0.411	tr.size 2: 0.422	tr.size 2: 0.423
	tr.size 3: 0.437	tr.size 3: 0.432	tr.size 3: 0.428	tr.size 3: 0.472	tr.size 3: 0.419	tr.size 3: 0.445	tr.size 3: 0.412	tr.size 3: 0.402	tr.size 3: 0.401	tr.size 3: 0.404	tr.size 3: 0.421	tr.size 3: 0.425
	tr.size 1: 0.457	tr.size 1: 0.454	tr.size 1: 0.433	tr.size 1: 0.456	tr.size 1: 0.415	tr.size 1: 0.424	tr.size 1: 0.410	tr.size 1: 0.404	tr.size 1: 0.407	tr.size 1: 0.410	tr.size 1: 0.412	tr.size 1: 0.418
Jun	tr.size 2: 0.446	tr.size 2: 0.439	tr.size 2: 0.432	tr.size 2: 0.455	tr.size 2: 0.404	tr.size 2: 0.422	tr.size 2: 0.408	tr.size 2: 0.401	tr.size 2: 0.407	tr.size 2: 0.411	tr.size 2: 0.411	tr.size 2: 0.415
	tr.size 3: 0.440	tr.size 3: 0.433	tr.size 3: 0.432	tr.size 3: 0.456	tr.size 3: 0.399	tr.size 3: 0.421	tr.size 3: 0.407	tr.size 3: 0.400	tr.size 3: 0.402	tr.size 3: 0.408	tr.size 3: 0.414	tr.size 3: 0.417
	tr.size 1: 0.420	tr.size 1: 0.411	tr.size 1: 0.430	tr.size 1: 0.446	tr.size 1: 0.410	tr.size 1: 0.416	tr.size 1: 0.417	tr.size 1: 0.398	tr.size 1: 0.403	tr.size 1: 0.406	tr.size 1: 0.414	tr.size 1: 0.416
Jul	tr.size 2: 0.422	tr.size 2: 0.420	tr.size 2: 0.429	tr.size 2: 0.446	tr.size 2: 0.404	tr.size 2: 0.415	tr.size 2: 0.421	tr.size 2: 0.397	tr.size 2: 0.400	tr.size 2: 0.404	tr.size 2: 0.410	tr.size 2: 0.414
	tr.size 3: 0.417	tr.size 3: 0.418	tr.size 3: 0.433	tr.size 3: 0.446	tr.size 3: 0.403	tr.size 3: 0.413	tr.size 3: 0.417	tr.size 3: 0.394	tr.size 3: 0.406	tr.size 3: 0.408	tr.size 3: 0.408	tr.size 3: 0.410
	tr.size 1: 0.443	tr.size 1: 0.434	tr.size 1: 0.430	tr.size 1: 0.443	tr.size 1: 0.424	tr.size 1: 0.413	tr.size 1: 0.414	tr.size 1: 0.403	tr.size 1: 0.405	tr.size 1: 0.410	tr.size 1: 0.413	tr.size 1: 0.415
Aug	tr.size 2: 0.435	tr.size 2: 0.433	tr.size 2: 0.428	tr.size 2: 0.441	tr.size 2: 0.444	tr.size 2: 0.412	tr.size 2: 0.413	tr.size 2: 0.401	tr.size 2: 0.403	tr.size 2: 0.409	tr.size 2: 0.411	tr.size 2: 0.417
	tr.size 3: 0.430	tr.size 3: 0.433	tr.size 3: 0.427	tr.size 3: 0.441	tr.size 3: 0.428	tr.size 3: 0.410	tr.size 3: 0.411	tr.size 3: 0.402	tr.size 3: 0.405	tr.size 3: 0.407	tr.size 3: 0.414	tr.size 3: 0.419
	tr.size 1: 0.413	tr.size 1: 0.403	tr.size 1: 0.429	tr.size 1: 0.437	tr.size 1: 0.410	tr.size 1: 0.409	tr.size 1: 0.409	tr.size 1: 0.403	tr.size 1: 0.403	tr.size 1: 0.405	tr.size 1: 0.411	tr.size 1: 0.413
Sep	tr.size 2: 0.410	tr.size 2: 0.400	tr.size 2: 0.428	tr.size 2: 0.436	tr.size 2: 0.405	tr.size 2: 0.408	tr.size 2: 0.410	tr.size 2: 0.401	tr.size 2: 0.404	tr.size 2: 0.408	tr.size 2: 0.414	tr.size 2: 0.419
	tr.size 3: 0.411	tr.size 3: 0.412	tr.size 3: 0.428	tr.size 3: 0.436	tr.size 3: 0.405	tr.size 3: 0.407	tr.size 3: 0.409	tr.size 3: 0.400	tr.size 3: 0.400	tr.size 3: 0.403	tr.size 3: 0.408	tr.size 3: 0.412
	tr.size 1: 0.412	tr.size 1: 0.402	tr.size 1: 0.433	tr.size 1: 0.445	tr.size 1: 0.419	tr.size 1: 0.411	tr.size 1: 0.410	tr.size 1: 0.405	tr.size 1: 0.408	tr.size 1: 0.410	tr.size 1: 0.417	tr.size 1: 0.425
Oct	tr.size 2: 0.411	tr.size 2: 0.404	tr.size 2: 0.433	tr.size 2: 0.444	tr.size 2: 0.422	tr.size 2: 0.411	tr.size 2: 0.411	tr.size 2: 0.404	tr.size 2: 0.409	tr.size 2: 0.411	tr.size 2: 0.413	tr.size 2: 0.424
	tr.size 3: 0.414	tr.size 3: 0.402	tr.size 3: 0.432	tr.size 3: 0.443	tr.size 3: 0.408	tr.size 3: 0.410	tr.size 3: 0.411	tr.size 3: 0.403	tr.size 3: 0.403	tr.size 3: 0.405	tr.size 3: 0.415	tr.size 3: 0.422
	tr.size 1: 0.436	tr.size 1: 0.421	tr.size 1: 0.435	tr.size 1: 0.436	tr.size 1: 0.414	tr.size 1: 0.410	tr.size 1: 0.413	tr.size 1: 0.405	tr.size 1: 0.406	tr.size 1: 0.409	tr.size 1: 0.415	tr.size 1: 0.418
Nov	tr.size 2: 0.432	tr.size 2: 0.411	tr.size 2: 0.434	tr.size 2: 0.435	tr.size 2: 0.408	tr.size 2: 0.405	tr.size 2: 0.411	tr.size 2: 0.405	tr.size 2: 0.405	tr.size 2: 0.409	tr.size 2: 0.414	tr.size 2: 0.416
	tr.size 3: 0.425	tr.size 3: 0.412	tr.size 3: 0.434	tr.size 3: 0.434	tr.size 3: 0.401	tr.size 3: 0.403	tr.size 3: 0.411	tr.size 3: 0.403	tr.size 3: 0.405	tr.size 3: 0.408	tr.size 3: 0.412	tr.size 3: 0.415
	tr.size 1: 0.450	tr.size 1: 0.455	tr.size 1: 0.432	tr.size 1: 0.436	tr.size 1: 0.413	tr.size 1: 0.407	tr.size 1: 0.415	tr.size 1: 0.406	tr.size 1: 0.410	tr.size 1: 0.413	tr.size 1: 0.417	tr.size 1: 0.422
Dec	tr.size 2: 0.445	tr.size 2: 0.440	tr.size 2: 0.429	tr.size 2: 0.435	tr.size 2: 0.414	tr.size 2: 0.406	tr.size 2: 0.414	tr.size 2: 0.405	tr.size 2: 0.409	tr.size 2: 0.411	tr.size 2: 0.418	tr.size 2: 0.425
	tr.size 3: 0.430	tr.size 3: 0.431	tr.size 3: 0.428	tr.size 3: 0.435	tr.size 3: 0.403	tr.size 3: 0.405	tr.size 3: 0.414	tr.size 3: 0.404	tr.size 3: 0.405	tr.size 3: 0.409	tr.size 3: 0.415	tr.size 3: 0.420
	tr.size 1: 0.436	**tr.size 1: 0.430**	**tr.size 1: 0.432**	**tr.size 1: 0.448**	**tr.size 1: 0.415**	**tr.size 1: 0.418**	**tr.size 1: 0.413**	**tr.size 1: 0.404**	**tr.size 1: 0.410**	**tr.size 1: 0.411**	**tr.size 1: 0.414**	**tr.size 1: 0.422**
Overall	**tr.size 2: 0.432**	**tr.size 2: 0.426**	**tr.size 2: 0.430**	**tr.size 2: 0.447**	**tr.size 2: 0.414**	**tr.size 2: 0.416**	**tr.size 2: 0.412**	**tr.size 2: 0.403**	**tr.size 2: 0.408**	**tr.size 2: 0.413**	**tr.size 2: 0.412**	**tr.size 2: 0.417**
	tr.size 3: 0.425	**tr.size 3: 0.414**	**tr.size 3: 0.431**	**tr.size 3: 0.446**	**tr.size 3: 0.408**	**tr.size 3: 0.416**	**tr.size 3: 0.409**	**tr.size 3: 0.410**	**tr.size 3: 0.404**	**tr.size 3: 0.410**	**tr.size 3: 0.413**	**tr.size 3: 0.416**

Table Appx.4: RMSE for the different ML methods in the validation and test set for the panel cross-validation for 2020.

Month	Valid. SVM	Test SVM	Valid. RF	Test RF	Valid. XGB	Test XGB	Valid. LASSO	Test LASSO	Valid. Ridge	Test Ridge	Valid. Logit	Test Logit
	tr.size 1: 0.743	tr.size 1: 0.733	tr.size 1: 0.818	tr.size 1: 0.803	tr.size 1: 0.793	tr.size 1: 0.832	tr.size 1: 0.826	tr.size 1: 0.826	tr.size 1: 0.788	tr.size 1: 0.789	tr.size 1: 0.674	tr.size 1: 0.677
Jan	tr.size 2: 0.752	tr.size 2: 0.745	tr.size 2: 0.823	tr.size 2: 0.809	tr.size 2: 0.787	tr.size 2: 0.837	tr.size 2: 0.827	tr.size 2: 0.830	tr.size 2: 0.793	tr.size 2: 0.795	tr.size 2: 0.688	tr.size 2: 0.685
	tr.size 3: 0.755	tr.size 3: 0.754	tr.size 3: 0.808	tr.size 3: 0.816	tr.size 3: 0.791	tr.size 3: 0.838	tr.size 3: 0.818	tr.size 3: 0.831	tr.size 3: 0.795	tr.size 3: 0.802	tr.size 3: 0.694	tr.size 3: 0.701
	tr.size 1: 0.733	tr.size 1: 0.722	tr.size 1: 0.784	tr.size 1: 0.779	tr.size 1: 0.794	tr.size 1: 0.810	tr.size 1: 0.821	tr.size 1: 0.806	tr.size 1: 0.802	tr.size 1: 0.807	tr.size 1: 0.706	tr.size 1: 0.710
Feb	tr.size 2: 0.724	tr.size 2: 0.714	tr.size 2: 0.805	tr.size 2: 0.794	tr.size 2: 0.805	tr.size 2: 0.814	tr.size 2: 0.826	tr.size 2: 0.815	tr.size 2: 0.804	tr.size 2: 0.810	tr.size 2: 0.723	tr.size 2: 0.717
	tr.size 3: 0.730	tr.size 3: 0.725	tr.size 3: 0.804	tr.size 3: 0.793	tr.size 3: 0.804	tr.size 3: 0.821	tr.size 3: 0.821	tr.size 3: 0.816	tr.size 3: 0.811	tr.size 3: 0.816	tr.size 3: 0.755	tr.size 3: 0.763
	tr.size 1: 0.710	tr.size 1: 0.712	tr.size 1: 0.794	tr.size 1: 0.822	tr.size 1: 0.779	tr.size 1: 0.830	tr.size 1: 0.839	tr.size 1: 0.816	tr.size 1: 0.809	tr.size 1: 0.818	tr.size 1: 0.711	tr.size 1: 0.708
Mar	tr.size 2: 0.732	tr.size 2: 0.720	tr.size 2: 0.807	tr.size 2: 0.811	tr.size 2: 0.797	tr.size 2: 0.837	tr.size 2: 0.840	tr.size 2: 0.819	tr.size 2: 0.814	tr.size 2: 0.821	tr.size 2: 0.724	tr.size 2: 0.722
	tr.size 3: 0.736	tr.size 3: 0.733	tr.size 3: 0.802	tr.size 3: 0.815	tr.size 3: 0.808	tr.size 3: 0.837	tr.size 3: 0.842	tr.size 3: 0.500	tr.size 3: 0.500	tr.size 3: 0.500	tr.size 3: 0.500	tr.size 3: 0.500
	tr.size 1: 0.714	tr.size 1: 0.700	tr.size 1: 0.820	tr.size 1: 0.786	tr.size 1: 0.819	tr.size 1: 0.787	tr.size 1: 0.821	tr.size 1: 0.828	tr.size 1: 0.823	tr.size 1: 0.822	tr.size 1: 0.790	tr.size 1: 0.788
Apr	tr.size 2: 0.715	tr.size 2: 0.711	tr.size 2: 0.811	tr.size 2: 0.784	tr.size 2: 0.813	tr.size 2: 0.778	tr.size 2: 0.825	tr.size 2: 0.827	tr.size 2: 0.826	tr.size 2: 0.824	tr.size 2: 0.801	tr.size 2: 0.800
	tr.size 3: 0.722	tr.size 3: 0.710	tr.size 3: 0.814	tr.size 3: 0.782	tr.size 3: 0.826	tr.size 3: 0.792	tr.size 3: 0.826	tr.size 3: 0.835	tr.size 3: 0.833	tr.size 3: 0.830	tr.size 3: 0.801	tr.size 3: 0.802
	tr.size 1: 0.702	tr.size 1: 0.678	tr.size 1: 0.813	tr.size 1: 0.823	tr.size 1: 0.791	tr.size 1: 0.823	tr.size 1: 0.824	tr.size 1: 0.821	tr.size 1: 0.818	tr.size 1: 0.821	tr.size 1: 0.731	tr.size 1: 0.733
May	tr.size 2: 0.722	tr.size 2: 0.702	tr.size 2: 0.811	tr.size 2: 0.823	tr.size 2: 0.794	tr.size 2: 0.827	tr.size 2: 0.826	tr.size 2: 0.823	tr.size 2: 0.811	tr.size 2: 0.818	tr.size 2: 0.743	tr.size 2: 0.748
	tr.size 3: 0.718	tr.size 3: 0.700	tr.size 3: 0.807	tr.size 3: 0.820	tr.size 3: 0.793	tr.size 3: 0.830	tr.size 3 :0.824	tr.size 3: 0.825	tr.size 3: 0.823	tr.size 3: 0.828	tr.size 3: 0.750	tr.size 3: 0.757
	tr.size 1: 0.745	tr.size 1: 0.724	tr.size 1: 0.812	tr.size 1: 0.804	tr.size 1: 0.798	tr.size 1: 0.816	tr.size 1: 0.830	tr.size 1: 0.821	tr.size 1: 0.820	tr.size 1: 0.818	tr.size 1: 0.750	tr.size 1: 0.744
Jun	tr.size 2: 0.743	tr.size 2: 0.732	tr.size 2: 0.804	tr.size 2: 0.799	tr.size 2: 0.820	tr.size 2: 0.820	tr.size 2: 0.834	tr.size 2: 0.826	tr.size 2: 0.823	tr.size 2: 0.820	tr.size 2: 0.753	tr.size 2: 0.751
	tr.size 3: 0.733	tr.size 3: 0.745	tr.size 3: 0.799	tr.size 3: 0.790	tr.size 3: 0.827	tr.size 3: 0.820	tr.size 3: 0.833	tr.size 3: 0.828	tr.size 3: 0.830	tr.size 3: 0.828	tr.size 3: 0.758	tr.size 3: 0.753
	tr.size 1: 0.738	tr.size 1: 0.735	tr.size 1: 0.828	tr.size 1: 0.807	tr.size 1: 0.803	tr.size 1: 0.820	tr.size 1: 0.815	tr.size 1: 0.837	tr.size 1: 0.832	tr.size 1: 0.830	tr.size 1: 0.744	tr.size 1: 0.743
Jul	tr.size 2: 0.758	tr.size 2: 0.745	tr.size 2: 0.821	tr.size 2: 0.809	tr.size 2: 0.820	tr.size 2: 0.823	tr.size 2: 0.818	tr.size 2: 0.839	tr.size 2: 0.833	tr.size 2: 0.833	tr.size 2: 0.754	tr.size 2: 0.750
	tr.size 3: 0.777	tr.size 3: 0.760	tr.size 3: 0.795	tr.size 3: 0.806	tr.size 3: 0.823	tr.size 3: 0.825	tr.size 3: 0.817	tr.size 3: 0.839	tr.size 3: 0.840	tr.size 3: 0.842	tr.size 3: 0.756	tr.size 3: 0.754
	tr.size 1: 0.802	tr.size 1: 0.773	tr.size 1: 0.818	tr.size 1: 0.807	tr.size 1: 0.783	tr.size 1: 0.819	tr.size 1: 0.828	tr.size 1: 0.822	tr.size 1: 0.819	tr.size 1: 0.815	tr.size 1: 0.753	tr.size 1: 0.751
Aug	tr.size 2: 0.792	tr.size 2: 0.780	tr.size 2: 0.814	tr.size 2: 0.803	tr.size 2: 0.752	tr.size 2: 0.821	tr.size 2: 0.827	tr.size 2: 0.824	tr.size 2: 0.822	tr.size 2: 0.820	tr.size 2: 0.757	tr.size 2: 0.753
	tr.size 3: 0.812	tr.size 3: 0.783	tr.size 3: 0.817	tr.size 3: 0.800	tr.size 3: 0.775	tr.size 3: 0.824	tr.size 3: 0.827	tr.size 3: 0.824	tr.size 3: 0.827	tr.size 3: 0.823	tr.size 3: 0.801	tr.size 3: 0.798
	tr.size 1: 0.720	tr.size 1: 0.723	tr.size 1: 0.810	tr.size 1: 0.808	tr.size 1: 0.803	tr.size 1: 0.817	tr.size 1: 0.831	tr.size 1: 0.824	tr.size 1: 0.820	tr.size 1: 0.811	tr.size 1: 0.765	tr.size 1: 0.761
Sep	tr.size 2: 0.744	tr.size 2: 0.734	tr.size 2: 0.804	tr.size 2: 0.798	tr.size 2: 0.812	tr.size 2: 0.819	tr.size 2: 0.834	tr.size 2: 0.825	tr.size 2: 0.822	tr.size 2: 0.815	tr.size 2: 0.777	tr.size 2: 0.769
	tr.size 3: 0.767	tr.size 3: 0.745	tr.size 3: 0.803	tr.size 3: 0.795	tr.size 3: 0.815	tr.size 3: 0.820	tr.size 3: 0.834	tr.size 3: 0.827	tr.size 3: 0.827	tr.size 3: 0.822	tr.size 3: 0.798	tr.size 3: 0.790
	tr.size 1: 0.711	tr.size 1: 0.718	tr.size 1: 0.821	tr.size 1: 0.785	tr.size 1: 0.793	tr.size 1: 0.815	tr.size 1: 0.827	tr.size 1: 0.829	tr.size 1: 0.821	tr.size 1: 0.818	tr.size 1: 0.745	tr.size 1: 0.740
Oct	tr.size 2: 0.731	tr.size 2: 0.711	tr.size 2: 0.820	tr.size 2: 0.790	tr.size 2: 0.785	tr.size 2: 0.815	tr.size 2: 0.831	tr.size 2: 0.831	tr.size 2: 0.826	tr.size 2: 0.820	tr.size 2: 0.755	tr.size 2: 0.749
	tr.size 3: 0.728	tr.size 3: 0.737	tr.size 3: 0.816	tr.size 3: 0.789	tr.size 3: 0.818	tr.size 3: 0.818	tr.size 3: 0.831	tr.size 3: 0.833	tr.size 3: 0.830	tr.size 3: 0.822	tr.size 3: 0.783	tr.size 3: 0.776
	tr.size 1: 0.689	tr.size 1: 0.677	tr.size 1: 0.804	tr.size 1: 0.817	tr.size 1: 0.801	tr.size 1: 0.812	tr.size 1: 0.823	tr.size 1: 0.823	tr.size 1: 0.811	tr.size 1: 0.808	tr.size 1: 0.769	tr.size 1: 0.761
Nov	tr.size 2: 0.687	tr.size 2: 0.723	tr.size 2: 0.797	tr.size 2: 0.805	tr.size 2: 0.811	tr.size 2: 0.824	tr.size 2: 0.826	tr.size 2: 0.820	tr.size 2: 0.822	tr.size 2: 0.815	tr.size 2: 0.772	tr.size 2: 0.766
	tr.size 3: 0.704	tr.size 3: 0.717	tr.size 3: 0.793	tr.size 3: 0.802	tr.size 3: 0.827	tr.size 3: 0.828	tr.size 3: 0.828	tr.size 3: 0.823	tr.size 3: 0.822	tr.size 3: 0.820	tr.size 3: 0.775	tr.size 3: 0.770
	tr.size 1: 0.632	tr.size 1: 0.598	tr.size 1: 0.812	tr.size 1: 0.803	tr.size 1: 0.798	tr.size 1: 0.817	tr.size 1: 0.819	tr.size 1: 0.820	tr.size 1: 0.815	tr.size 1: 0.811	tr.size 1: 0.744	tr.size 1: 0.738
Dec	tr.size 2: 0.656	tr.size 2: 0.643	tr.size 2: 0.809	tr.size 2: 0.795	tr.size 2: 0.797	tr.size 2: 0.819	tr.size 2: 0.822	tr.size 2: 0.820	tr.size 2: 0.816	tr.size 2: 0.811	tr.size 2: 0.748	tr.size 2: 0.744
	tr.size 3: 0.697	tr.size 3: 0.645	tr.size 3: 0.810	tr.size 3: 0.797	tr.size 3: 0.820	tr.size 3: 0.821	tr.size 3: 0.822	tr.size 3: 0.821	tr.size 3: 0.820	tr.size 3: 0.818	tr.size 3: 0.755	tr.size 3: 0.749
	tr.size 1: 0.720	**tr.size 1: 0.710**	**tr.size 1: 0.811**	**tr.size 1: 0.799**	**tr.size 1: 0.795**	**tr.size 1: 0.810**	**tr.size 1: 0.824**	**tr.size 1: 0.821**	**tr.size 1: 0.814**	**tr.size 1: 0.810**	**tr.size 1: 0.733**	**tr.size 1: 0.732**
Overall	**tr.size 2: 0.701**	**tr.size 2: 0.700**	**tr.size 2: 0.811**	**tr.size 2: 0.795**	**tr.size 2: 0.798**	**tr.size 2: 0.815**	**tr.size 2: 0.827**	**tr.size 2: 0.824**	**tr.size 2: 0.815**	**tr.size 2: 0.811**	**tr.size 2: 0.744**	**tr.size 2: 0.742**
	tr.size 3: 0.755	**tr.size 3: 0.731**	**tr.size 3: 0.806**	**tr.size 3: 0.796**	**tr.size 3: 0.810**	**tr.size 3: 0.817**	**tr.size 3: 0.826**	**tr.size 3: 0.806**	**tr.size 3: 0.801**	**tr.size 3: 0.800**	**tr.size 3: 0.694**	**tr.size 3: 0.691**

Table Appx.5: AUC for the different ML methods in the validation and test set for the panel cross-validation for 2020.

F An overview of the alternative CATE estimators

The BLP, GATEs and CLAN analysis we presented has been performed by using the CATEs estimated by using various meta-learners and the *Generalized Random Forests* (CFs) method. CFs modifies a particular standard (predictive) machine learning method (Random Forests) so that it directly targets the estimation of the CATE. Instead, meta-learners operate through a multi-step procedure that break down the task of estimating CATE into several smaller sub-problems that can be addressed using any standard (predictive) machine learning technique. They typically involve the following steps:

1. Estimation of nuisance parameters: auxiliary components such as the the propensity scores are estimated using machine learning algorithms.

2. Construction of an objective function: The estimated nuisance components are then used to construct a tailored minimization problem whose solution targets the CATE function. This step is designed to isolate the heterogeneity in treatment effects while accounting for confounding.

3. Solution via machine learning: The resulting minimization problem is solved via machine learning.

4. Prediction of CATEs: Finally, the learned model is used to generate predictions of the CATE for each observational unit, thus enabling individualized causal effect estimation.

The meta-learners we have used are the S-Learner, the T-Learner, the R-Learner and the DR-Learner. They are all based on the assumption of strong ignorability. We will provide a brief overview in what follows. $m(1,x)$ denotes the conditional mean function under treatment, $m(0,x)$ the conditional mean function in absence of treatment, and $e(X)$ the propensity score.
The first meta-learner we explore is the *S-Learner*. It fits a single model in which the observed outcome is modeled as a function of the covariates and the treatment indicator variable. The resulting model is then used to obtain two predictions for each subject: under treatment and control. The CATE is then estimated by taking the differences between the two predictions: $\hat{\Delta}^s(x) = \hat{m}(1,x) - \hat{m}(0,x)$. Instead, the T-Learner employs two different models estimated separately on the treated and control samples, and then the CATE is obtained as a difference, as in the S-Learner.

The *R-Learner*, introduced by Nie and Wager (2021), builds on the partially linear model while allowing for covariate-specific treatment effects. In this setting, the potential outcome model is given by $Y^D = \Delta(X)D + g(X) + U^D, \quad \text{with } \mathbb{E}[U^D \mid D, X] = 0$, which implies that the observed outcome satisfies $Y = D\Delta(X) + g(X) + U^D$. By defining the outcome regression function $m(X) := \mathbb{E}[Y \mid X]$, the model can be transformed into a residualized form:

$$Y - m(X) = \Delta(X)(D - e(X)) + U^D.$$

This motivates the R-Learner objective:

$$\hat{\Delta}^{rl}(X) = \arg\min_{\Delta} \sum_{i=1}^{N} \left(Y_i - \hat{m}(X_i) - \Delta(X_i)(W_i - \hat{e}(X_i))\right)^2,$$

where the nuisance parameters $\hat{m}(X)$ and $\hat{e}(X)$ are estimated using high-quality machine learning methods, typically via cross-fitting [41] (Nie and Wager, 2021).
When one does not wish to impose linearity on $\Delta(X)$, the R-Learner objective can be rewritten as

$$\hat{\Delta}^{rl}(X) = \arg\min_{\Delta} \sum_{i=1}^{N} (D_i - \hat{e}(X_i))^2 \left(\frac{Y_i - \hat{m}(X_i)}{D_i - \hat{e}(X_i)} - \Delta(X_i)\right)^2.$$

This representation highlights that any supervised learning algorithm capable of handling weighted minimization problems can be employed (e.g. neural networks, random forests, and gradient boosting among others). In this formulation, the weights are $(D_i - \hat{e}(X_i))^2$, the pseudo-outcome is $\frac{Y_i - \hat{m}(X_i)}{D_i - \hat{e}(X_i)}$, and the original covariates X are used as features.
The *DR-Learner*, introduced by Kennedy et al. (2020), constructs a pseudo-outcome in the first stage, defined as

$$\tilde{Y}_{\text{ATE}} = \underbrace{\hat{m}(1, X) - \hat{m}(0, X)}_{\text{outcome predictions}} + \underbrace{\frac{D(Y - \hat{m}(1, X))}{\hat{e}(X)} - \frac{(1 - D)(Y - \hat{m}(0, X))}{1 - \hat{e}(X)}}_{\text{weighted residuals}},$$

similarly to Eq. (17) and targets the CATE function through the conditional expectation

$$\Delta(x) = \mathbb{E}\left[\tilde{Y}_{\text{ATE}} \mid X = x\right].$$

Since $\mathbb{E}[\tilde{Y}_{\text{ATE}} \mid X]$ is a conditional expectation function of random variable, it can be approximated using standard supervised learning techniques. The DR-Learner uses this pseudo-outcome as the dependent variable in a generic machine learning regression:

$$\hat{\Delta}^{\text{dr}}(X) = \arg\min_{\Delta} \sum_{i=1}^{N} \left(\tilde{Y}_{i,\text{ATE}} - \Delta(X_i)\right)^2.$$

The *Generalized Random Forests* (CFs) estimator cannot be properly considered a Meta-learner because it alters a specific ML method in such a way that it estimates the CATE. The development of *Generalized Random Forests* (CFs) has evolved through multiple stages. The Generalized Random

[41] A simple case arises, for instance, if we model the CATE as a linear function $\Delta(X) = X'\beta$, the minimization becomes

$$\hat{\beta}^{rl} = \arg\min_{\beta} \sum_{i=1}^{N} \left(Y_i - \hat{m}(X_i) - \tilde{X}_i'\beta\right)^2,$$

where $\tilde{X}_i = (D_i - \hat{e}(X_i))X_i$ are the so-called modified or pseudo-covariates. The estimated CATE is then given by $\hat{\Delta}^{rl}(x) = x'\hat{\beta}^{rl}$, noting that this differs from $\tilde{X}\hat{\beta}^{rl}$. In this case, obtaining $\hat{\beta}^{rl}$ reduces to a standard regression of the residualized outcome on the modified covariates, and shrinkage estimators such as Lasso can be readily applied. Importantly, the nuisance parameters can still be estimated using non-linear machine learning methods

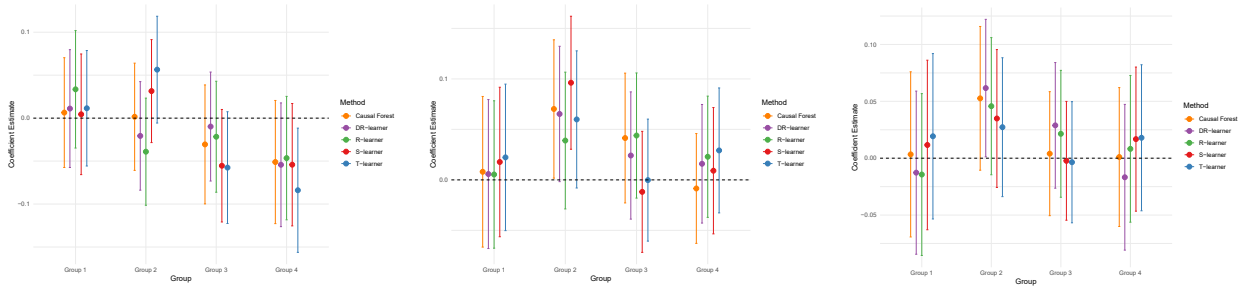
Forest introduced by Wager and Athey (2018) is constructed as an ensemble of *Causal Trees* and is primarily designed for experimental settings. Subsequently, Athey et al. (2019) extended this approach by proposing an approximation to the splitting rule of Causal Trees for binary random treatments, while also generalizing the method to observational settings and continuous treatments. Conceptually, modern Generalized Random Forests estimate CATEs via a localized, individualized residual-on-residual regression of the form

$$\hat{\Delta}^{\text{cf}}(x) = \arg\min_{\Delta} \left\{ \sum_{i=1}^{N} \pi_i(x) \left[(Y_i - \hat{m}(X_i)) - \Delta(x)(D_i - \hat{e}(X_i)) \right]^2 \right\},$$

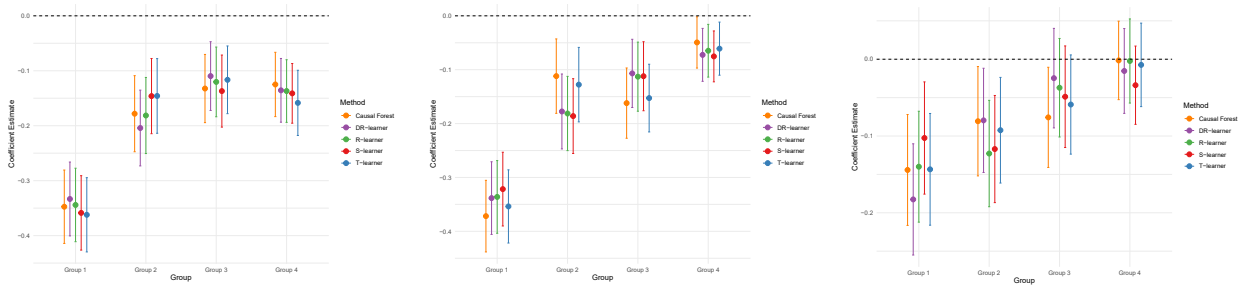
where $\pi_i(x)$ denotes the frequency with which the i-th training sample falls into the same leaf as the target sample x. This procedure represents a localized version of the partially linear estimator, with the nuisance components $\hat{m}(X)$ and $\hat{e}(X)$ being estimated in a preliminary step, typically through cross-fitting.

Finally, in the main text, for comparability with CFs, we use Random Forest in steps 1 and 3 of the meta-learners described above.

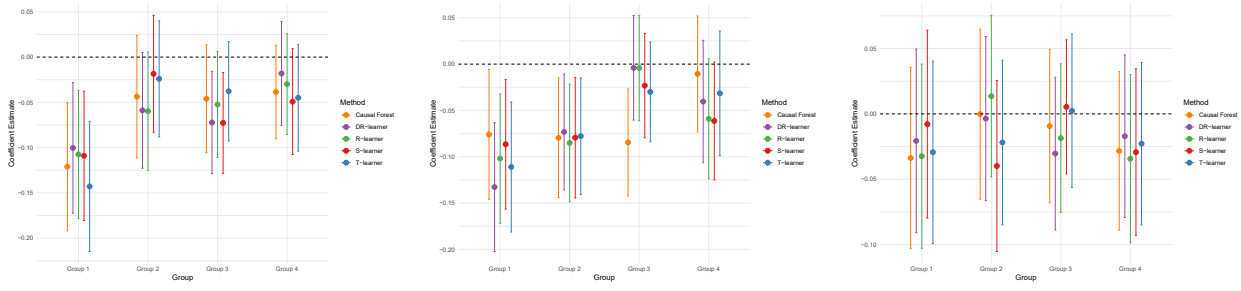
GATES estimates from January to December 2020 are summarized in Figure Appx.11.



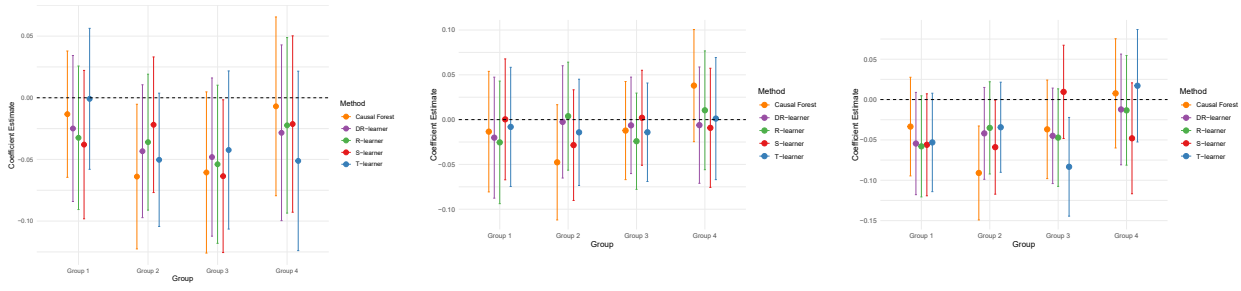
January February March



April May June



July August September



October November December

Figure Appx.11: GATES estimates from January to December 2020. The results are shown for the four quartiles according to CATE. In each graph, the colored bars are from left to right for Generalized Random Forest (orange), DR-learner (purple), R-learner (green), S-learner (red) and T-learner (blue).

G DML AIPW estimator

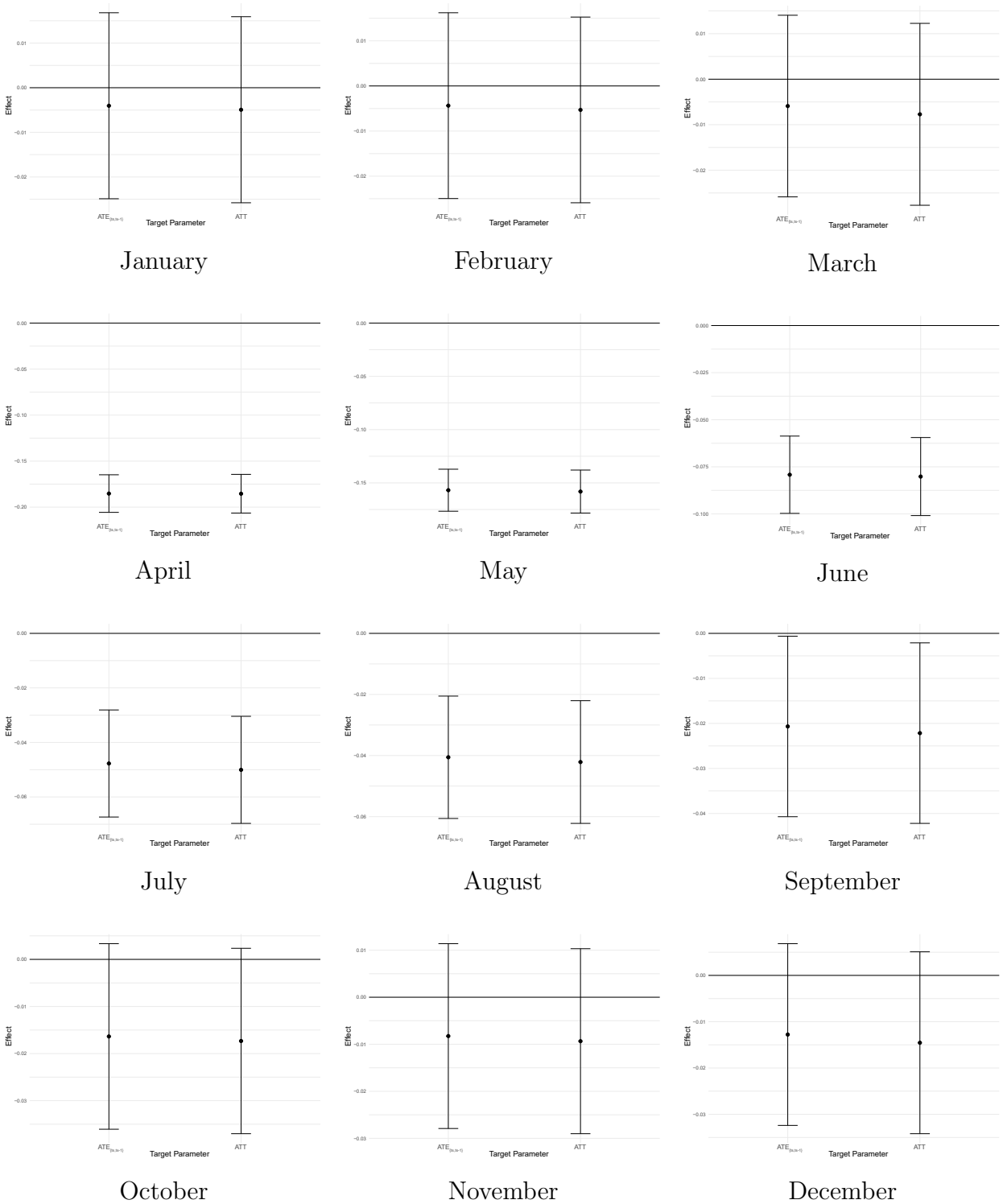
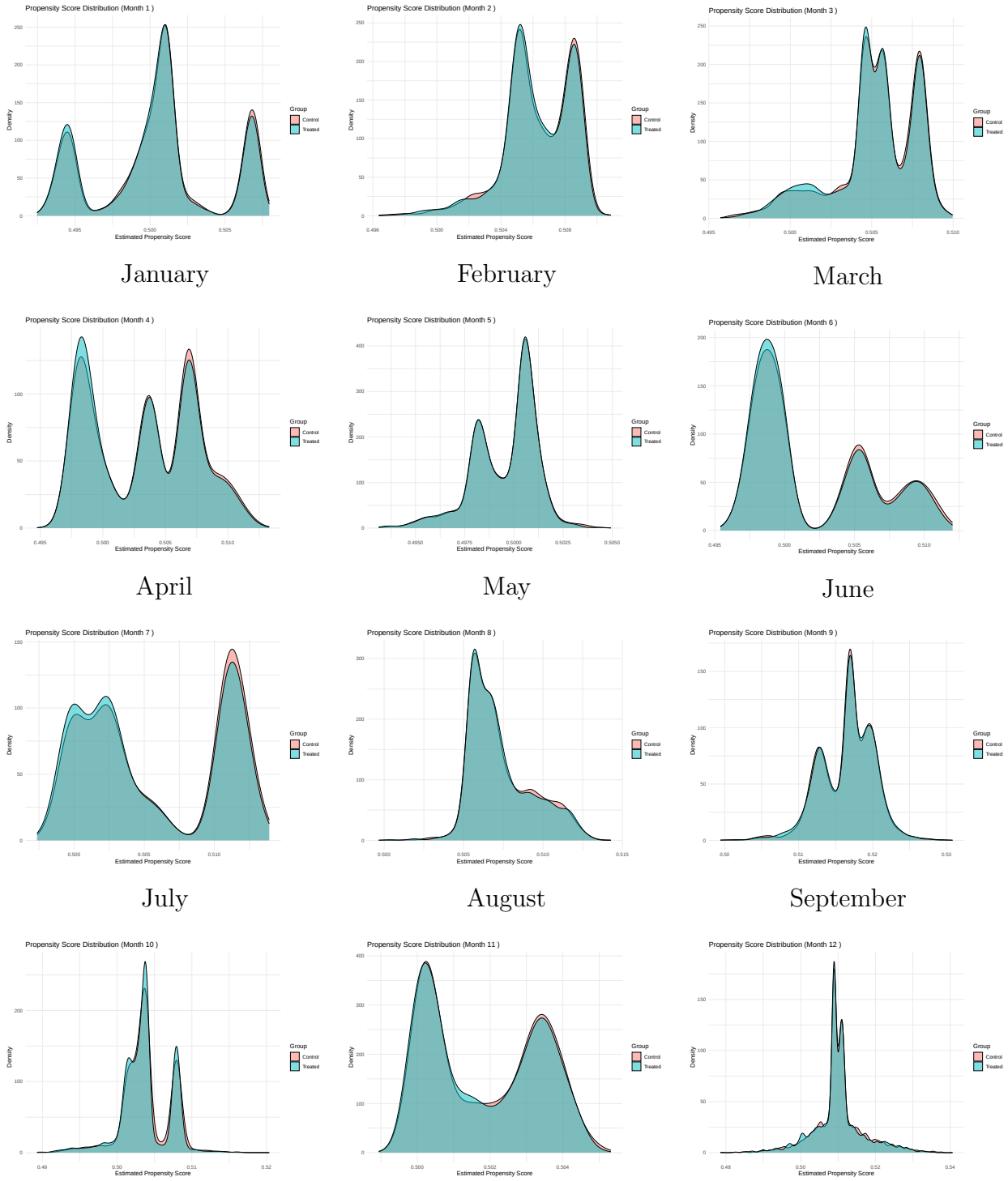


Figure Appx.12: Double Machine Learning (DML) AIPW estimates (with 5-fold cross-fitting and nuisance parameters estimated with Generalized Random Forest) from January to December 2020. $ATE_{\{t_s,t_{s-1}\}}$ refers to the ATE estimated by considering the combined cohort of treated (firms in t_s) and control firms (firms in t_{s-1}) as a unique sample. Notice also that $ATT = ATT_{\{t_s,t_{s-1}\}}$ because all the treated units are in t_s.

H Estimated Propensity Scores

The propensity score is defined in equation (3) as $P(D_{i,\{t_s,t_s-1\}} = 1|X_{i,\{t_s-1,t_s-2\}}) = e(X_{i,\{t_s-1,t_s-2\}})$, where $D_{i,\{t_s-1,t_s\}}$ is a dummy variable indicating whether an observation belongs to the treated group or to the control group, and $X_{i,\{t_s-1,t_s-2\}}$ are the corresponding explanatory variables. Therefore, the propensity score refers to the conditional probability of belonging to the cohort of firms observed in $t-s$ considering the unique sample that combines the cohort of treated (firms observed in t_s) and the cohort of control firms (firms observed in t_{s-1}). Following the methodology of Chernozhukov et al. (2018), a K-fold cross-fitting strategy is employed to estimate this quantity without overfitting and Generalized Random Forest Lerner is used.



January February March

April May June

July August September

October November December

Figure Appx.13: Propensity scores estimates from January to December using 5-Folds DML-AIPW based on Generalized Random Forest.