Assessing the Impact of COVID-19 on Trade: a Machine Learning Counterfactual Analysis

Marco Dueñas[1], Federico Nutarelli[1], Víctor Ortiz[1], Massimo Riccaboni[1] and Francesco Serti[1]

[1]IMT School for Advanced Studies Lucca

EEA-ESEM Bocconi Milan

August 22, 2022

Motivation

- International trade has been severely affected by the health crisis and the related containment policies originated during the COVID-19 pandemic.
- Global trade, which is typically more volatile than output, has shown the biggest fall since the 2009 global financial crisis.
- The export collapse has been the consequence of a demand-side shock coming from destination markets that was accompanied also by a domestic supply-side shock (Baldwin and Tomiura, 2020).
- This domestic supply-side effect is reinforced by a supply-side contagion via importing/supply chains.
- Supply disruptions in imported intermediate inputs are likely to hurt also exports performance.

Contributions (I)

- This paper aims to estimate the causal effect of the 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.
  1. All firms are directly and/or indirectly exposed to the effects of the COVID-19 crisis making impossible to find a control group to build a counterfactual non-COVID-19 scenario.
  2. The economy-wide impact of the shock and the complex interdependencies between firms and products across sectors and countries makes difficult to identify the main patterns through which the shock has affected firm-level trade.

Contributions (II)

- By interpreting exporters' dynamics as a complex learning process, we investigate the effectiveness of different Machine Learning (ML) techniques in predicting Colombian firms' trade status.
- We predict the probability of Colombian firms to survive in the export market under two different scenarios: a COVID-19 setting and a non-COVID-19 counterfactual situation.
- 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.
- We use these predictions to estimate the causal effect of the COVID-19 shock at the individual firm level.

Contributions (III)

- In the literature using ML counterfactuals (Cerqua and Letta, 2020; Fabra et al., 2020), it is common to estimate causal effects by comparing the counterfactual predictions with the observed outcome in case of treatment
- We instead follow Chernozhukov et al. (2020) by using ML techniques both to reconstruct firm potential outcomes in case of no treatment and to predict the outcomes in the treatment scenario.
- Finally, we study the heterogeneity of the COVID-19 effects according to firms' characteristics. 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.
- By adapting recent Causal Machine learning tools (Chernozhukov et al. 2018, 2020) to our setting, we use the estimated effects stemming from our ML counterfactual empirical model to classify firms in two groups, the most and the least affected by the COVID-19 shock, and then we compare their average characteristics.

Data

1. Monthly export transactions data reported at the Colombian Customs Office (Dirección de Impuestos y Aduanas Nacionales, DIAN) for 2018, 2019, and 2020, considering:
    - 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.
    - the free on board value of the transaction in US dollars.
    - import value and country of origin
2. The COVID-19 Government Response Tracker data for 2020 (Monthly Government Measures Indexes).

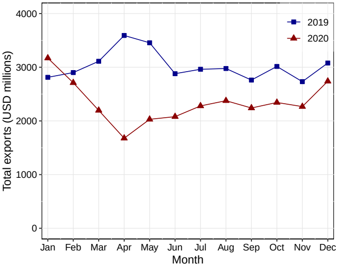
Variables considered (I)

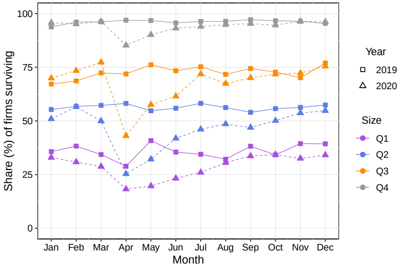
- We consider different features of exporters, according to their monthly trading: the total export (and import) value, the number of products (*NP*), the number of export destinations (*ND*), the number of import origin countries (*NO*), a set of dummies according the Colombian-department from which the product comes from, a set of dummies for the means of transportation used, a set of the dummies classifying the product sector (HS-chapter), and the product industry (HS-section).
- Moreover, we build two sets of dummy variables indicating whether a firm has experience exporting in specific destinations and product sector in a given month of the last year. 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.

Variables considered (II)

- To measure the COVID-19 demand and supply shock, we use the information on government contention measures, which consists of four indexes (ranging from 0 to 100) representing the strength of the measures taken by countries to contain the COVID-19 outbreak.
- 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.
- We build two variables at the firm level for each 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 firm i "Containment Index$_{i,j,z}$", with $j = \{E, H, G, S\}$ and $z = \{\text{Imp}, \text{Exp}\}$.

Evolution of total exports

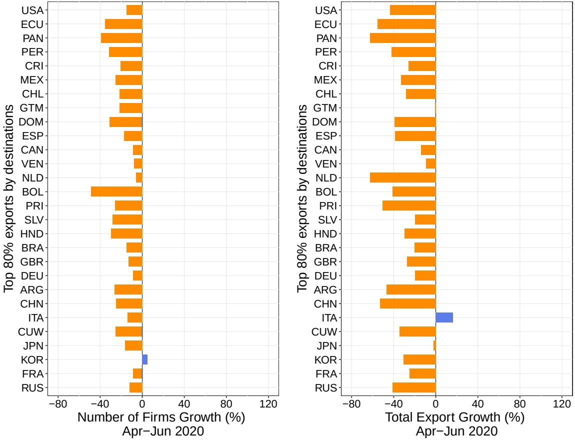




- The total monthly value of exports in 2020 is significantly reduced.
- The lockdown measures had a severe impact between April and June.
- The COVID-19 outbreak affected all firms regardless of their size, even if the effect appears stronger for smaller firms.
- Larger firms seem to be less affected and recover faster to the survival rates observed in 2019.

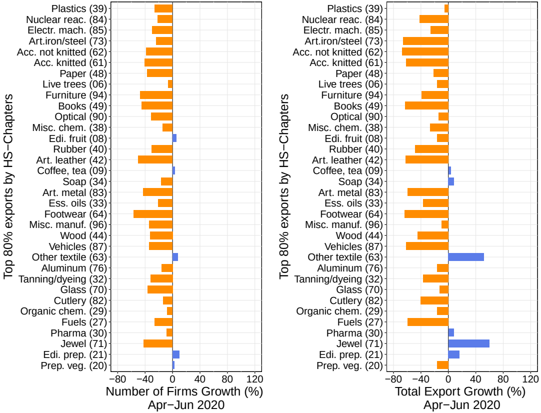
Entry-exit dynamics by destination

- There is a severe drop of the number of exporting firms and the volume of exports in practically all destinations.



Entry-exit dynamics by product chapter

- Negative shocks specially for sectors:
  - Footwear (HS64)
  - Leather Articles (HS42)
  - Furniture (HS94)
  - Books (HS49)
  - Articles of Metal (HS83)
  - Knitted and Not-Knitted Accessories (HS61-62)
  - Vehicles (HS87)
  - Articles of Iron or Steel (HS73)



The impact of the COVID-19 shock on Colombian firms' export has been heterogeneous across sectors and destinations.

Methodology (I)

Important considerations

- The main identification task is to build a counterfactual non-COVID-19 outcome for firms in 2020.
- Unfortunately, one cannot select any subset of untreated Colombian firms as a control group because this treatment is affecting all firms during 2020.
- To build this firm-level counterfactual, we use the information on firms' exporting behavior available for periods before the crisis.
- The main assumption is that we can learn what would have happened in the months of 2020 without the COVID-19 by exploiting the observed firm behavior in 2018-2019. Empirical Strategy

Methodology (II)

We use Machine Learning techniques to build the counterfactual scenario for the 2020 firms' level outcomes by using pre-pandemic information on firms' export behavior and firms' characteristics (see also: Cerqua and Letta (2020) and Fabra et al. (2020).)

- The outcome (*success*) that we want to analyze is whether a firm that was exporting in a given month in 2019 will export again in the same month of 2020.
- We build two different machines:
  - **Shock Unaware Machine** (SUM): it is the counterfactual machine, which does not consider the COVID-19 information.
  - **Shock Aware Machine** (SAM): it is fully aware of all the available information related to the COVID-19 scenario (i.e., firms behaviour in 2020 and measures at economic, health and government level, summarized in the different Indexes).

Methodology (III)

- We train SUM by using the characteristics of exporters observed in 2018 to explain their export behavior in 2019.
- We choose the best performing predictive SUM model (out of sample) using cross-validation techniques (i.e., K-fold method).
- We apply the "best performing" SUM (trained using data in 2018-2019) to predict the 2020 outcome for firms exporting in 2019 (the counterfactual).
- The SAM machine instead considers the exporters operating in the market in 2019 and use their observed dynamics in 2020: it is just a probabilistic picture of what actually happened.
- Therefore, we construct the counterfactual by using the SUM to predict export behavior in 2020 of firms exporting in 2019, and we compare these counterfactual predictions with those obtained by the SAM.

$$\hat{\alpha}_i = \hat{Y}_i^{SAM} - \hat{Y}_i^{SUM}. \tag{1}$$

Methodology (IV)

- We compute average treatment effects by month and by subsamples defined according to firm characteristics, and we calculate bootstrapped standard errors.
- Our estimator will be unbiased if the expected values of the prediction error of the SUM and of the SUM are the same in the relevant subsample.
- In order to uncover the heterogeneity of the effects much more fully, we estimate sorted effects: a collection of estimated partial effects sorted in increasing order and indexed by percentiles (Chernozhukov et al., 2018).
- First, we order the estimated individual specific treatment effects and calculate their percentiles.

Methodology (V)

- Second, we classify firms as highly affected and weakly affected by COVID-19 according to whether their estimated individual effects are lower than the 25th percentile or greater than the 75th percentile of the distribution of the estimated treatment effects, respectively.
- Third, we test which are the characteristics on which these two group of firms differ on average (difference in means of firm characteristics).
- We use the bootstrap to calculate standard errors of the difference in means and we calculate joint p-values that account for simultaneous inference (cause we are simultaneously testing many hypotheses, many differences in means).

ML Models (I)

- The prediction performance out of sample of our empirical models is of fundamental importance because our identification strategy is based on the ability to reconstruct a counterfactual that is in practice out of sample, because it is unobserved.
- Our approach recognizes that this is a complex task because
  - we have a very high number of potential explanatory variables
  - the existence of complex interdependencies between firms, and products and destinations that are difficult to know ex-ante.

ML Models (II)

- In such a situation, an approach that is based on the maximization of the accuracy of in-sample predictions will be prone to overfitting.
- Instead, ML techniques have been shown to constitute the best way to choose the optimal positioning on bias-variance trade-off for (out of sample) predictive tasks.
- We compare four different models: *Logit, Logit-Ridge, Logit-LASSO* Logit, and *Random Forest (RF)* RF.

Model Selection

- We compare the out of sample predictive performance of the four different models.
- We focus on various statistics summarizing the predictive power of the models.
  - **Root Mean Squared Error:** The closer to 0, the more accurate is the model
    $$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}$$
  - **Area Under the receiver operating Curve (AUC):** Varies between 0.5 and 1, where 0.5 means that we predict randomly and 1 that the model predicts correctly all the individuals. ▸ AUC

- Table 1 reports the accuracy of the estimates obtained studying the probability of exporting in 2019 for the population of 2018 exporters by using (5 folds) cross-validation.
- Logit-LASSO and RF models are the best performers.

Table 1: Goodness of Fit: SUM, 2018/19

	AUC				RMSE			
	Logit-LASSO	Logit-Ridge	Random Forest	Logit	Logit-LASSO	Logit-Ridge	Random Forest	Logit
Jan	0.73	0.53	0.73	0.59	0.40	0.45	0.41	0.64
Feb	0.70	0.50	0.71	0.58	0.41	0.45	0.41	0.64
Mar	0.70	0.56	0.71	0.57	0.41	0.44	0.41	0.65
Apr	0.73	0.59	0.73	0.60	0.40	0.43	0.40	0.63
May	0.72	0.52	0.71	0.59	0.40	0.44	0.41	0.64
Jun	0.71	0.50	0.72	0.59	0.40	0.45	0.41	0.64
Jul	0.73	0.50	0.73	0.55	0.40	0.45	0.40	0.66
Aug	0.70	0.51	0.72	0.58	0.41	0.45	0.40	0.64
Sep	0.72	0.50	0.71	0.58	0.41	0.45	0.40	0.64
Oct	0.73	0.58	0.74	0.58	0.40	0.44	0.41	0.64
Nov	0.71	0.51	0.72	0.57	0.41	0.45	0.41	0.64
Dec	0.70	0.50	0.71	0.58	0.41	0.45	0.41	0.64

SUM Models Performance, 2019/20

- The models of Table 2 are also trained with exporters characteristics in 2018 and their observed outcome in 2019.
- However, these models are tested using the set of exporters of 2019 and their observed outcome in 2020, the COVID-19 year.

Table 2: Goodness of Fit: SUM, 2019/20

	AUC				RMSE			
	Logit-LASSO	Logit-Ridge	Random Forest	Logit	Logit-LASSO	Logit-Ridge	Random Forest	Logit
Jan	0.72	0.53	0.72	0.49	0.41	0.45	0.41	0.75
Feb	0.69	0.50	0.69	0.56	0.41	0.45	0.42	0.64
Mar	0.72	0.54	0.73	0.59	0.40	0.44	0.41	0.63
Apr	0.67	0.56	0.66	0.51	0.48	0.50	0.49	0.70
May	0.69	0.51	0.69	0.60	0.46	0.48	0.46	0.63
Jun	0.68	0.50	0.68	0.59	0.43	0.47	0.44	0.63
Jul	0.70	0.50	0.69	0.59	0.42	0.46	0.43	0.63
Aug	0.68	0.51	0.69	0.58	0.42	0.45	0.43	0.63
Sep	0.69	0.50	0.70	0.59	0.42	0.45	0.42	0.63
Oct	0.71	0.59	0.70	0.60	0.42	0.45	0.43	0.63
Nov	0.71	0.51	0.71	0.59	0.41	0.45	0.41	0.63
Dec	0.69	0.50	0.69	0.58	0.42	0.46	0.42	0.63

SUM Models Performance, 2019/20 (II)

- If the functions representing the relationship between explanatory variables and the outcome in absence of the pandemic are sufficiently similar for the pre-pandemic year and 2020, we expect that the accuracy in the first three months of 2020 (when arguably no relevant COVID-19 effect is in place in Colombia) to be similar to that one which is observed during the same months of 2019.
- Indeed, during these months, the accuracy of Logit-LASSO and RF remains unchanged, as expected, compared to the accuracy obtained in Table 1.
- After April, the accuracy obtained in Table 2 is lower because it refers to the ability of a model trained without using any COVID-19 information to predict outcomes under a COVID-19 shock scenario.

SAM Models Performance, 2019/20

- Models in Table 3 are trained and tested with the universe of exporters in 2019 and their observed outcomes in 2020.
- The accuracy of the predictions is very similar to the one obtained with the SUM for 2019 and for the first three months of 2020.

Table 3: Goodness of Fit: SAM, 2019/20

	AUC				RMSE			
	Logit-LASSO	Logit-Ridge	Random Forest	Logit	Logit-LASSO	Logit-Ridge	Random Forest	Logit
Jan	0.73	0.58	0.74	0.50	0.41	0.45	0.41	0.71
Feb	0.70	0.50	0.70	0.49	0.41	0.46	0.42	0.70
Mar	0.73	0.50	0.73	0.50	0.40	0.46	0.40	0.71
Apr	0.74	0.66	0.73	0.52	0.42	0.47	0.42	0.69
May	0.76	0.74	0.77	0.50	0.41	0.46	0.41	0.71
Jun	0.73	0.69	0.73	0.48	0.42	0.46	0.42	0.72
Jul	0.73	0.63	0.72	0.51	0.41	0.45	0.42	0.69
Aug	0.72	0.50	0.72	0.53	0.41	0.46	0.42	0.69
Sep	0.71	0.50	0.70	0.55	0.42	0.47	0.42	0.67
Oct	0.72	0.50	0.71	0.52	0.42	0.46	0.42	0.70
Nov	0.72	0.52	0.72	0.49	0.41	0.45	0.41	0.71
Dec	0.71	0.51	0.70	0.51	0.41	0.45	0.42	0.70

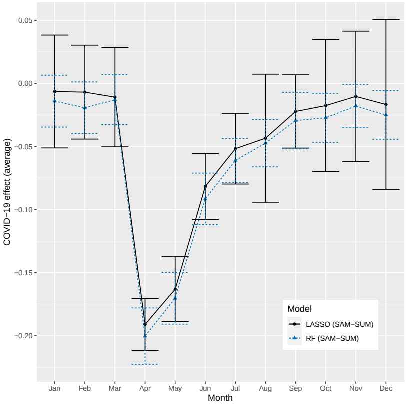
Evaluation of the COVID-19 effect

In-Time Placebo

- Suppose during first months of 2020 firms are not affected by COVID-19 shock $\rightarrow$ Comparing SAM and SUM predictions is a falsification test (Abadie et al., 2015).
- If we estimate economically significant effect of COVID-19 before the actual shock happened $\rightarrow$ our model mechanically predicts a COVID-19 effect even when it is not expected.
- We apply this placebo study also conditioning on exogenous (observed in 2019) firms' characteristics.
- We interpret these placebo studies as robustness checks of our results on treatment effect heterogeneity.

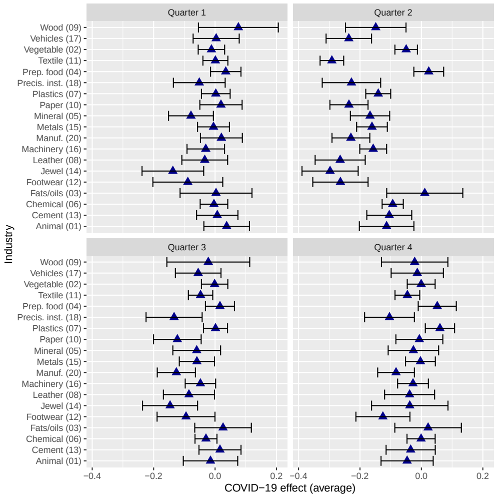
Average Treatment Effect by Month

We use the Logit-LASSO predicted probabilities to estimate the average monthly effect of the COVID-19 shock as the monthly average of $\hat{\alpha}_i = \hat{Y}_i^{SAM} - \hat{Y}_i^{SUM}$.



- **In-Time Placebo** $\rightarrow$The probabilities obtained from the SUM and the SAM are almost identical on average from January to March (March 25, 2020, the Colombian government implemented the lockdown).

Average Treatment Effect by Industry

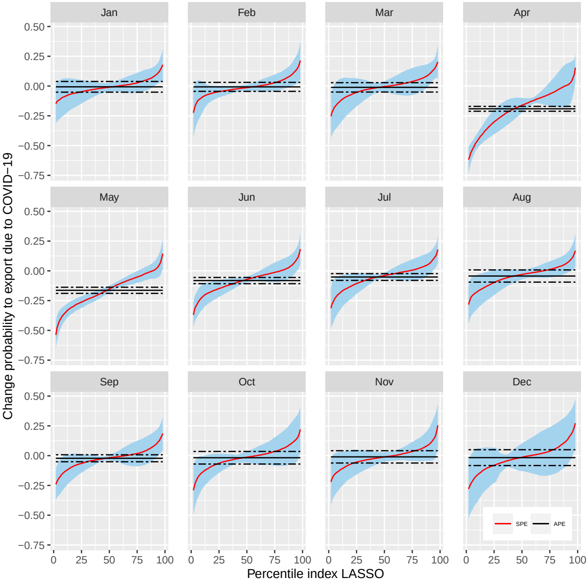


- The effects at the industry level are negative in general, but there are industries more affected than others.

Sorted effects by month (I)

- The next figure shows the estimated Sorted Partial Effects (SPE) by month, which are just the percentiles of the estimated sorted individual treatment effects, and 95% confidence intervals with blue bands (in black as a reference the average partial effects, APE).
- The main result is that we find significant treatment effect heterogeneity just for the months of April, May and, to a lesser extent, June, when statistically significant negative values are reported just in the left tail of the distribution.
- Instead, starting from July the confidence intervals of the SPEs intersect those of APE.
- Very importantly, we can also observe how the SPEs do almost coincide with the APEs in pre-pandemic months, suggesting that our methodology is robust also in the tails of the distribution of treatment effects.

Sorted effects by month (II)



Determinants of TE heterogeneity (I)

- To identify the determinants of treatment effect heterogeneity, we focus on the difference in means of the the main explanatory variables across the most and least affected groups (according to whether their estimated individual treatment effects are lower than the first quartile or greater than the third quartile, respectively).
- 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 indicating whether a firm belongs to the group of the most affected firms (in left tail of the distribution of the effects, with negative effects).
- Then, we also provide the difference in adjusted means once we have controlled for firm sector and both for firm sector and month of the year.
- The variables that we consider to explore the sources of COVID-19 treatment effect heterogeneity are firm characteristics observed in 2019 (the year before receiving the treatment): the industry, the means of transportation, the months when firms operate, the number of export destinations (ND), of import origins (NO), and of products (NP) exported. We also consider as dependent variables the weighted Containment Stringency Index that exporters face when exporting and importing.

Determinants of TE heterogeneity (II)

Outcome variable	$\beta^1_{1,f}$	$\beta^2_{1,f}$	$\beta^3_{1,f}$
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^{**}	
ND	-0.1990	-0.1640	-0.2480
NO	-1.7470	-1.9820^{***}	-2.4440^{**}
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			✓

Determinants of TE heterogeneity (III)

- The most affected exporters (those located in the first SPE quartile distribution) experienced a decrease in the probabilities to export between 27.9*p.p.* and 31.3*p.p.* lower than the one experienced by the least affected firms (those located in the third SPE quartile distribution).
- The share of Textile firms among the most affected 2019 exporters is 16*p.p.* higher with respect to the one estimated for the group of the least affected firms. Similarly, we find the presence of 2.9*p.p.* more wood exporters among the most affected than among the least affected firms.
- There are more exporters using the air among the most affected than among the least affected firms. However, there are less exporters using the sea to ship goods among the most affected than among the least affected firms.

Determinants of TE heterogeneity (IV)

- 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 to ND, NP, and, in the first column, to NO.
- However, once we control for sector and therefore, inter alia, for the fact that some sector has relatively more diversification potential, we find that the most affected Colombian exporters tend to import from 1.98 less countries in 2019 than the least affected firms.
- The most affected Colombian exporters face on average a higher Export (Import) Containment Stringency Index with respect to the one faced by least affected firms.
- Finally, the least affected firms exported (imported) more value in 2019 than the most affected firms. Therefore, Colombian exporters trading in larger volumes (in value) are more resilient under a COVID-19 scenario.

Concluding remarks

- Our study contributes to the understanding of the effects of COVID-19 on international trade.
- We exploit pre-2020 information and ML methods to reconstruct the counterfactual of the 2020 firm-level outcomes and to study treatment effect heterogeneity.
- This paper suggest that when no control group is available, ML can be used to build a more accurate counterfactual with respect to traditional models (e.g., Logit).
- Predicting a counterfactual is the main task to identify causal effects, and ML has been shown to be extremely useful for doing predictions.
- This is the first paper showing that ML can be successfully applied to predict firms' trade potential.
- This also opens up the possibility to use ML to assist firms and public agencies in their decision-making processes.

Thanks for your attention

Appendix

App. 1: Empirical Strategy

Back

- We denote the potential outcome and the regressors under the scenario $d \in \{0,1\}$ for firm i at time t as Y_{it}^{d} and X_{it}^{d}, where d is an indicator variable for the presence of COVID-19.
- The first step of the analysis is to estimate the counterfactual outcome in 2020: $Y_{i,2020}^{0}$.
- In particular, we will use the outcomes and covariates observed in 2018 and 2019 to reconstruct Y_{2020}^{0} under the following assumptions (we omit i):

App. 1: Empirical Strategy

(i) Both covariates and outcomes of 2018 and 2019 are not affected by the pandemic:

$$Y_t = Y_t^0 = Y_t^1, \quad X_t = X_t^0 = X_t^1 \quad \text{for} \quad t = 2018, 2019. \tag{2}$$

(ii) Define $Y_t^0 = f_t^0(X_{t-1}^0) + u_t^0$, where $f_t^0(\cdot)$ is a generic model or function representing the relationship between explanatory variables and the outcome in absence of the pandemic such that $\mathbf{E}[Y_t^0|X_{t-1}^0] = f_t^0(X_{t-1}^0)$.
Under (i), for $t = 2019$ we have that $Y_{2019} = f_{2019}^0(X_{2018}) + u_{2019}^0$ such that $\mathbf{E}[Y_{2019}|X_{2018}] = f_{2019}^0(X_{2018})$.
The second assumption states that the function f_t^0 does not depend on t, i.e. it is stable over the two considered years:

$$f_{2019}^0 = f_{2020}^0 = f^0 \tag{3}$$

- Under the above assumptions, we can write $Y^0_{2020} = f^0(X_{2019}) + u^0_{2020}$, such that $\mathbf{E}[Y^0_{2020}|X_{2019}] = f^0(X_{2019})$, and we can use data on 2018 and 2019 to estimate $Y^0_{2019} = f^0(X_{2018}) + u^0_{2019}$ and retrieve $\hat{f}^0$.
- By applying this invariant estimated function to the covariates of 2019 we can obtain the predictions for the counterfactual (without COVID-19) outcome in 2020:

$$\hat{Y}^0_{2020} = \hat{f}^0(X_{2019}) = Y^0_{2020} - \overbrace{\mathcal{E}^0_{2020}(X_{2019})}^{\textit{Prediction error}} - \overbrace{u^0_{2020}}^{\textit{Orthogonal error}} \tag{4}$$

App. 1: Empirical Strategy

- In general, the estimated counterfactual outcome in 2020, $\hat{Y}^0_{2020}$, will not be a perfect estimate for Y^0_{2020} because $\hat{f}^0$ will not be a perfect estimate of f^0 thus producing a prediction error, which in the formula above we have denoted with $\mathcal{E}^0_{2020}(X_{2019}) = f^0(X_{2019}) - \hat{f}^0(X_{2019})$, and because of the existence of other determinants of the outcome that are orthogonal to the covariates, which in the formula above are contained in u^0_{2020}.
- The inaccuracy coming from the estimation of f^0, that can vary according to a firm's characteristics X_{2019}, will be reduced by experimenting with different ML techniques and using the one associated with the best out-of-sample performance.

App. 1: Empirical Strategy

- Finally, we obtain the $\hat{Y}^0_{2020}$ by estimating $Y_{2019} = f^0(X_{2018}) + u^0_{2019}$ on entire set of 2018 exporters (also in this case month by month) and, as shown in (4), applying the estimated function $\hat{f}^0$ to the set of 2019 exporters.
- Given that during the first three months of 2020 Colombia was in practice not exposed to COVID-19 (and therefore $Y_{2020} = Y^0_{2020}$), if assumption (3) holds we expect that in those months the accuracy of the predictions $\hat{Y}_{2019}$ obtained in the cross-validation step for 2019 will be very similar to those of $\hat{Y}^0_{2020}$ for 2020.

App. 1: Empirical Strategy

- Following Cerqua and Letta (2020) and Fabra et al. (2020), we define as an estimator of the individual-specific COVID-19 effect α the simple comparison of the observed outcome under COVID-19 in 2020 with the estimated counterfactual outcome for a given firm:

$$\hat{\hat{\alpha}} = Y_{2020} - \hat{Y}^{0}_{2020}. \tag{5}$$

- Eq. (5) provides the full distribution of treatment effects.
- All the parameters of interest of the paper are obtained by computing (conditional) averages and quantiles of such distribution.

App. 1: Empirical Strategy

- Starting from Eq. (5), by taking the expected value of the individual treatment effect $\hat{\hat{\alpha}}$ for those units with $X_{2019} = x_{2019}$, we can define the following estimator of the conditional average treatment effect (CATE; the average effect for those units with $X_{2019} = x_{2019}$)

$$\begin{aligned} \mathbf{E}[\hat{\hat{\alpha}}|X_{2019} = x_{2019}] &= \mathbf{E}[(Y_{2020} - Y^0_{2020}) - \mathcal{E}^0_{2020} - u^0_{2020}|X_{2019} = x_{2019}] = \\ &= \underbrace{\Delta(X_{2019} = x_{2019})}_{CATE} - \mathbf{E}[\mathcal{E}^0_{2020}|X_{2019} = x_{2019}] - \underbrace{\mathbf{E}[u^0_{2020}|X_{2019} = x_{2019}]}_{=0 \text{ by assumption}}, \end{aligned} \tag{6}$$

where,

$$\Delta(X_{2019} = x_{2019}) = \mathbf{E}[Y_{2020} - Y^0_{2020}|X_{2019} = x_{2019}].$$

- Therefore $\mathbf{E}[\hat{\hat{\alpha}}_i]$ will identify the unconditional average treatment effect, $\mathbf{E}[\Delta(X_{2019})] = \Delta$, if on average the prediction error is zero: $\mathbf{E}[\mathcal{E}^0_{2020}] = 0$.
- The conditional average treatment effect, $\Delta(X_{2019} = x_{2019})$, will be identified by $\mathbf{E}[\hat{\hat{\alpha}}_i|X_{2019} = x_{2019}]$ if on average the prediction error will be zero in the relevant sub-sample: $\mathbf{E}[\mathcal{E}^0_{2020}|X_{2019} = x_{2019}] = 0$.

App. 1: Empirical Strategy

- Now let's decompose the outcome observed in 2020 in presence of the pandemic, Y^1_{2020}, in a generic model or function $f^1(X^1_{2019})$, which represents the relationship between explanatory variables and the outcome during the pandemic, and other determinants of the outcome, u^1_{2020}, that are orthogonal to the covariates

$$Y^1_{2020} = f^1(X^1_{2019}) + u^1_{2020}, \quad s.t.\ \mathbf{E}[Y^1_{2020}|X^1_{2019}] = f^1(X^1_{2019}). \quad (7)$$

- Given that $Y^1_{2020} = Y_{2020}$ and $X^1_{2019} = X_{2019}$, then

$$Y_{2020} = f^1(X_{2019}) + u^1_{2020}, \quad s.t.\ \mathbf{E}[Y_{2020}|X_{2019}] = f^1(X_{2019}). \quad (8)$$

App. 1: Empirical Strategy

- At this point, we can define an alternative estimator of the individual-specific COVID-19 effect α as the comparison of the predicted outcome under COVID-19 in 2020 with the estimated counterfactual outcome for a given firm:

$$\hat{\alpha} = \hat{Y}_{2020} - \hat{Y}^{0}_{2020}, \tag{9}$$

where $\hat{Y}_{2020} = \hat{f}^{1}(X_{2019}) = Y_{2020} - \mathcal{E}^{1}_{2020} - u^{1}_{2020}$. We call "Shock Aware Machine" (SAM) the model that we use to predict Y_{2020} (and the predictions $\hat{Y}_{2020}$ themselves)

App. 1: Empirical Strategy

- Starting from Eq. (9), by taking the expected value of the individual treatment effect $\hat{\alpha}$ for those units with $X_{2019} = x_{2019}$, we can define the following alternative estimator of the conditional average treatment effect (for those units with $X_{2019} = x_{2019}$)

$$\begin{aligned}\mathbf{E}[\hat{\alpha}_i | X_{2019} = x_{2019}] =& \mathbf{E}[(Y_{2020} - Y^0_{2020}) - (\mathcal{E}^1_{2020} - \mathcal{E}^0_{2020}) - (u^1_{2020} - u^0_{2020}) | X_{2019} = x_{2019}] \\ =& \underbrace{\Delta(X_{2019} = x_{2019})}_{CATE} - \mathbf{E}[\underbrace{(\mathcal{E}^1_{2020} - \mathcal{E}^0_{2020})}_{\Delta\mathcal{E}} | X_{2019} = x_{2019}] - \\ & \mathbf{E}[u^1_{2020} - u^0_{2020} | X_{2019} = x_{2019}].\end{aligned} \tag{10}$$

- Therefore, $\mathbf{E}[\hat{\alpha}_i]$ will identify the unconditional average treatment effect, $\mathbf{E}[\Delta(X_{2019})] = \Delta$, if on average the difference in prediction errors is zero: $\mathbf{E}[\Delta\mathcal{E}] = 0$.
- The conditional average treatment effect, $\Delta(X_{2019} = x_{2019})$, will be identified by $\mathbf{E}[\hat{\alpha}_i | X_{2019} = x_{2019}]$ if on average the difference in prediction errors is zero in the relevant sub-sample: $\mathbf{E}[\Delta\mathcal{E} | X_{2019} = x_{2019}] = 0$.

App. 1: Empirical Strategy

- Given the definitions of SUM and SAM, to simplify the reasoning in the following we will refer to Equations (5) and (9) respectively as

$$\hat{\hat{\alpha}} = Y - \hat{Y}_{SUM} = Y - SUM. \tag{11}$$

$$\hat{\alpha} = \hat{Y}_{SAM} - \hat{Y}_{SUM} = SAM - SUM. \tag{12}$$

- The assumptions behind these identification results are not directly testable as they are expressed in terms of the expected values of the prediction error $\mathcal{E}^0_{2020}$ that is a function of the unobservable counterfactual Y^0_{2020}.
- The next table distinguishes the five different scenarios concerning the values of $\mathcal{E}^0_{2020}$ and $\mathcal{E}^1_{2020}$ that are relevant in determining whether applying the statistic $\mathbf{T}$ to $Y - SUM$ and $SAM - SUM$ is able to recover the corresponding treatment effect estimand (e.g., whether averaging the estimated individual treatment effects would recover the average treatment effect).

App. 1: Empirical Strategy

	$\mathbf{T}(SAM - SUM)$	$\mathbf{T}(Y - SUM)$
$\mathbf{T}[\mathcal{E}^1_{2020}] \neq 0$ and $\mathbf{T}[\mathcal{E}^0_{2020}] = 0$	X	✓
$\mathbf{T}[\mathcal{E}^1_{2020}] = \mathbf{T}[\mathcal{E}^0_{2020}] = 0$	✓	✓
$\mathbf{T}[\mathcal{E}^1_{2020}] = 0$ and $\mathbf{T}[\mathcal{E}^0_{2020}] \neq 0$	X	X
$\mathbf{T}[\mathcal{E}^1_{2020}] = \mathbf{T}[\mathcal{E}^0_{2020}] \neq 0$	✓	X
$\mathbf{T}[\mathcal{E}^1_{2020}] \neq \mathbf{T}[\mathcal{E}^0_{2020}] \neq 0$	X	X

Table 4: Identification of generic functions of the individual treatment effects, **T**, according to the corresponding value taken by the prediction errors.

App. 2: Logit, Ridge, LASSO

Back

- **Logit** estimates the parameters maximizing the following log-likelihood function:

$$l(\beta) = \sum_{i=1}^{n} [y_i x_i \beta - log(1 + e^{x_i \beta})]$$

- **Logit-Ridge** adds a L_2 penalty to $l(\beta)$, that shrinks the parameters towards zero, without actually setting any of them to zero:

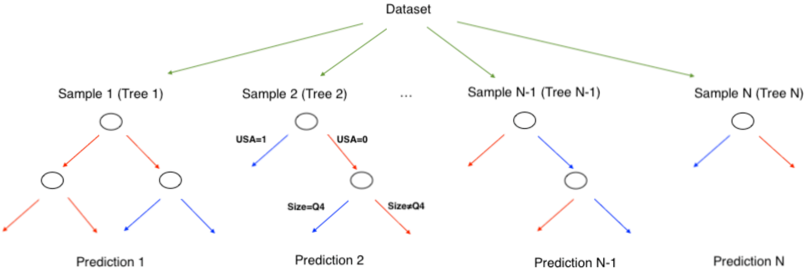
$$l(\beta) = \sum_{i=1}^{n} [y_i x_i \beta - log(1 + e^{x_i \beta})] - \lambda \sum_{j=1}^{p} \beta_j^2$$

- **Logit-LASSO** adds a L_1 penalty to $l(\beta)$, that forces some parameters to be exactly zero:

$$l(\beta) = \sum_{i=1}^{n} [y_i x_i \beta - log(1 + e^{x_i \beta})] - \lambda \sum_{j=1}^{p} \mid \beta_j \mid$$

Back

- **RF** is composed by Random Trees. The final outcome of the RF is the average of the N predictions.



App. 2: Pros and Cons of different models

Back

- Logit as a benchmark model. Predicted performance is expected to be bad under large data sets or without a theoretical grounded model. Moreover, it is a high computational *cost* model. Possible problem of overfitting.
- Logit-Ridge is *faster* than Logit (for any fixed value of *lambda*). Good predictive performance when many variables of the model are relevant. But still possible overfitting problems when just few variables are relevant.
- Logit-LASSO has the benefit of *reducing the number of predictors* in the final model. Powerful when only a bunch of predictors have a *lot of predictor power*.
- RF is more robust to outliers. Moreover it takes into account all possible interactions, without specifying them. Every tree is independent of each other so RF *avoids overfitting*. However, RF has a high computational *cost*.

Back

Pred	Actual
0.83	1
0.72	0
0.44	1
0.31	0
0.61	1
0.55	0
0.91	1

		Actual	
		1	0
Predict	1	3(TP)	2(FP)
	0	1(FN)	1(TN)

TPR = 3/(3+1) = 3/4 FPR = 2/(2+1) = 2/3