Measuring the Causal Effects of the COVID-19 Pandemic on Commodity Prices

Abstract

Commodity exchanges are fundamental to the global economy. Measuring the pandemic's impacts on commodity prices can help policy makers devise strategies that minimize economic losses in times of challenging global events. This paper compares the causal effects of the numbers of new COVID-19 cases on days d_{t-3}, d_{t-2}, and d_{t-1} on prices of commodities (i.e., corn, soybean, and crude oil) on day d_t in the U.S. from December 14, 2020 when COVID-19 vaccines were first distributed to June 9, 2021. Unlike previous attempts, we employed a causal discovery procedure and allowed for the possible presence of unmeasured confounders by using the tsGFCI algorithm for studying the underlying causal structure among the variables. Our findings suggest that the numbers of daily new cases on days d_{t-3}, d_{t-2}, and d_{t-1} all have negative effects on day d_t's commodity prices. However, the number of cases on a given day affects the prices of different commodities on day d_t unevenly. Moreover, in contrast to when the numbers of cases are low, when the numbers of cases are high, cases from more distant days have greater effects on day d_t's prices. This suggests that commodity markets might take more time to adjust to greater shocks.

1. Background & Significance

Many developing countries depend heavily on commodity exchanges—trades of raw materials such as oil and agricultural products. Due to their global nature, commodity markets are extremely sensitive to shocks, including the COVID-19 pandemic. Measuring the pandemic's impacts on commodity prices is of significant importance. Not only does it allow market participants to better predict prices, but it can also help policy makers in developing nations many of which rely greatly on commodity exports devise strategies that minimize economic losses in times of challenging global events.

Although there have been several attempts in measuring the causal effects of the pandemic on commodity prices, to our knowledge, few of them are based on causal discovery techniques. A notable attempt we have examined is by Kamdem et al. (2020) which is based on Granger causality. This method's main drawback is that it does not take latent variables into account. As raw materials underpin many daily activities in the global economy, we believe there are several underlying factors which can affect prices. Therefore, it is crucial to consider the possibility of unmeasured confounders when assessing how the pandemic affects commodity prices. To this end, we used tsGFCI—a hybrid constraint-based and score-based algorithm for causal structure learning from time-series data in settings with unmeasured confounding (Malinsky and Spirtes, 2018). In particular, we compared the causal effects of the numbers of new COVID-19 cases on days d_{t-3}, d_{t-2}, and d_{t-1} on the prices of corn, soybean, and crude oil on day d_t in the U.S. from December 14, 2020 when COVID-19 vaccines were first distributed (U.S. Department of Health & Human Services, 2021) to June 9, 2021. To record the pandemic's severity and vaccination progress, we used the numbers of daily new cases and new vaccinations in Dong et al. (2020), respectively. Commodity prices were taken from Yahoo Finance (2021). Since holiday seasons can affect prices, we also used data on major holidays in the U.S. from Weakley (2020).

2. Methods

2.1 Causal Discovery

We employed TETRAD (Center for Causal Discovery, n.d.), a software for general causal discovery, to learn the underlying causal system. We first used a *Data* box to convert our data into a *time lag* data set with 5 time lags. With this, we assumed the causal structure among the variables are the same for any 5 consecutive days in the time period we consider. This is a reasonable assumption given our causal query and our limited computational power. We then fed the new dataset into the tsGFCI algorithm. Since our data contain both discrete and continuous variables, we used the *Degenerate Gaussian Likelihood Ratio Test* for constraints testing and the *Degenerate Gaussian BIC Score* for BIC Score calculation (Andrews et al., 2019). We assumed faithfulness (see Appendix A), and applied a penalty of 0.2 for the number of parameters in the BIC Score calculation. The cutoff value α for rejecting an edge was 0.05. We did 50 bootstraps and chose Preserved(0) as the ensemble method. The output is a *Partial Ancestral Graph* (PAG) (see Appendix B). In a PAG, a $\circ\!\!\rightarrow$ edge between A and B implies that either A is a cause of B, or a latent variable is a cause of A and B, or both. We discarded the first two time lags due to possible convergence issues (Ebert-Uphoff and Deng, 2012). We also manually processed some edges to prevent

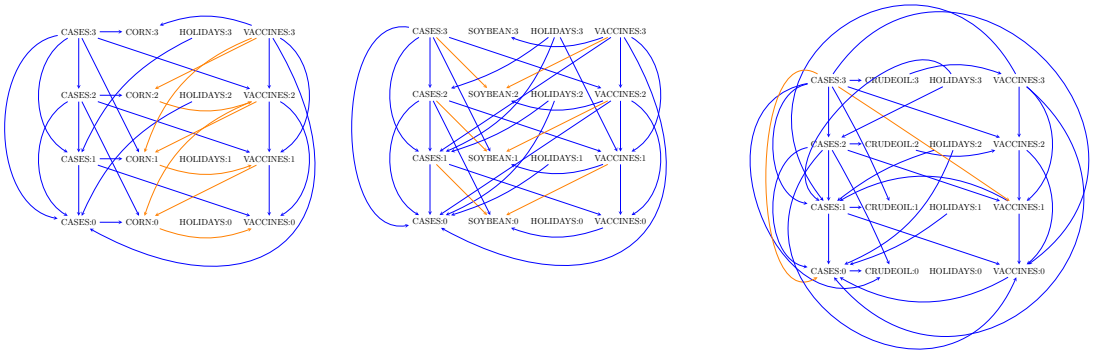


Figure 1: The PAGs for Prices of Corn (left), Soybean (center), and Crude Oil (right).

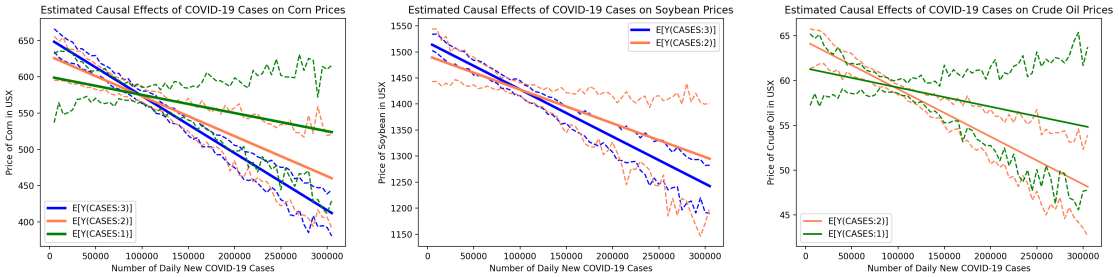
cyclicity and make use of the fact that no factor can cause whether a day is a holiday (see Appendix C). The resulting PAGs for corn, soybean, and crude oil prices are in Figure 1. A $\circ\!\!\rightarrow$ egde is drawn as $\rightarrow$. Note that CASES:3 refers to the number of cases reported on day d_{t-3}, CORN:2 refers to the price of corn on day d_{t-2}, HOLIDAYS:1 refers to whether day d_{t-1} is a holiday, VACCINES:0 refers to the number of new vaccinations on day d_t, etc.

2.2 Causal Identification & Estimation

In this work, we interpreted a $\circ\!\!\rightarrow$ edge from A to B in the PAG as both A is a cause of B ($A \rightarrow B$) and a latent variable is a cause of A and B ($A \leftrightarrow B$) in the Acyclic Directed Mixed Graph (ADMG) to which the observed distribution factorizes (see Appendix A). Then, we used the backdoor criterion, a graphical criterion first appeared in Pearl (1995), to adjust for confounding in non-experimental data. A set of variables Z satisfies the backdoor criterion with respect to treatment A and outcome Y in an ADMG $\mathcal{G}$ if no variable in Z is a descendant of A, and Z blocks all paths between A and Y that starts with an edge into A ($A \leftarrow \ldots Y$, or $A \leftrightarrow \ldots Y$). Then, the backdoor adjustment formula can be stated as follows,

$$\mathbb{E}[Y(a)] = \sum_{Z} p(Z) \times \mathbb{E}[Y \mid A = a, Z].$$

An equivalent definition of the backdoor criterion can be found in Appendix D. Using the backdoor criterion, the empty set is a valid adjustment set to estimate the causal effect of CASES:3 on CORN:0 since there are no paths of the forms CASES:3 $\leftarrow \ldots$ CORN:0 or CASES:3 $\leftrightarrow \ldots$ CORN:0 in the corresponding ADMG. However, the causal effect of CASES:3 on CRUDEOIL:0 is unidentifiable because there are no variables that are not descendants of CASES:3 that can block the CASES:3 $\leftrightarrow$ CASES:0 $\rightarrow$ CRUDEOIL:0 path.



(a) Corn (b) Soybean (c) Crude Oil

Figure 2: Estimated Causal Effects of COVID-19 Cases on Commodity Prices Using Linear Models

3. Results & Discussion

Assuming the models are linear, after applying the backdoor formula to the corresponding adjustment sets (see Appendix E), we obtained estimates as in Figure 2. We calculated the causal effects of each value for the number of new cases in the range (5000, 310000) on commodity prices (the minimum and maximum number of new cases in our data set is 5751 and 303487, respectively), the next value differed from the previous by 5000. For each value, we estimated the 95% confidence interval with 50 bootstraps.

Our results suggest that the numbers of new cases reported on days d_{t-3}, d_{t-2}, and d_{t-1} all have negative impacts on all commodity prices on day d_t. However, the number of new cases on a given day affects commodity prices differently. For instance, since the relationship between the number of new cases and price is assumed to be linear, our results indicate that an increase of 50000 cases on day d_{t-2} leads to a drop of approximately 25USX, 30USX, and 2USX in prices of corn, soybean, and crude oil on day d_t, respectively. Interestingly, across all commodities considered, when the numbers of daily new cases are low (below 100000), the cases reported on more recent days have stronger effects on day d_t's prices. For instance, considering corn, when the numbers of cases are at 50000, cases on days d_{t-1}, d_{t-2}, and d_{t-3} cause prices of corn on day d_t to be at around 590USX, 610USX, and 630USX, respectively. Yet, when the numbers of cases are above 100000, the cases reported on later days have weaker effects on day d_t's prices. This suggests that commodity markets might take more time to respond to greater shocks. Our estimations also show that the more recent the days are, or the higher the numbers of cases are; the wider the confidence intervals tend to become, indicating greater uncertainty in estimating the causal effects. When examining non-linear models, the estimates also show similar conclusions (see Appendix F).

This analysis has several limitations. First, the data set contains many dates for which commodity prices are not recorded. The small sample size might bias our estimates. Second, due to computational limitations, we could only examine five time lags in our causal discovery procedure. As mentioned, commodity markets are extremely complicated due to several underlying physical and economic processes. Therefore, it is possible that factors from more distant days can affect prices on day d_t. For future work, one might consider adding more variables, or examine more time lags and compare the effects of the numbers of cases at different time steps on day d_t's commodity prices.

References

Bryan Andrews, Joseph Ramsey, and Gregory F. Cooper. Learning High-dimensional Directed Acyclic Graphs with Mixed Data-types. *Proceedings of Machine Learning Research*, 2019.

Center for Causal Discovery. The TETRAD Project. Causal models and statistical data. Available at `http://www.phil.cmu.edu/projects/tetrad/` (2021/12/05), n.d.

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, 2020. ISSN 1473-3099.

Imme Ebert-Uphoff and Yi Deng. Causal Discovery for Climate Research Using Graphical Models. *Journal of Climate*, 25(17):5648–5665, 2012.

Jules Sadefo Kamdem, Rose Bandolo Essomba, and James Njong Berinyuy. Deep learning models for forecasting and analyzing the implications of COVID-19 spread on some commodities markets volatilities. *Chaos, Solitons & Fractals*, 140:110215, 2020.

Daniel Malinsky and Peter Spirtes. Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery*, volume 92 of *Proceedings of Machine Learning Research*, pages 23–47, 2018.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 12 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.669. URL `https://doi.org/10.1093/biomet/82.4.669`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Thomas S. Richardson and James M. Robins. Single World Intervention Graphs (SWIGs) : A Unification of the Counterfactual and Graphical Approaches to Causality. 2013.

Sewall Right. Correlation and Causation. *Journal of Agricultural Research*, pages 557–580, 1921.

U.S. Department of Health & Human Services. Covid-19 Vaccine Distribution: The Process. Available at `https://www.hhs.gov/coronavirus/covid-19-vaccines/distribution/index.html` (2021/12/04), 2021.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence*, 1990.

Donnete Weakley. US Holiday Dates (2014-2021). Available at `https://www.kaggle.com/donnetew/us-holiday-dates-2004-2021?select=US+Holiday+Dates+%282004-2021%29.csv` (2021/12/05), 2020.

Yahoo Finance. Commodity Futures Price History. Available at https://www.kaggle.com/mattiuzc/commodity-futures-price-history (2021/12/05), 2021.

Appendix A. The Faithfulness Assumption

A graph $\mathcal{G} = (V, E)$ is a directed acyclic graph (DAG) if there is at most one edge between each pair of vertices, E only contains directed edges, and $\mathcal{G}$ has no directed cycles. A structural equation model (SEM) $\mathcal{M}$ defined on a set of variables $V = \{V_1, \ldots, V_k\}$ is a system of non-parametric structural equations that models how values for each variable in V are generated (Right, 1921). A SEM $\mathcal{M}$ on V induces a distribution $p(V)$ which factorizes with respect to its corresponding DAG $\mathcal{G}$ as follows,

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)),$$

where $\text{pa}_{\mathcal{G}}(V_i)$ denotes the parents of V_i in $\mathcal{G}$. Then, conditional independences in $p(V)$ can be derived from $\mathcal{G}$ via a graphical criterion called *d-separation*. That is, $(X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep in } \mathcal{G}} \implies (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$. In this paper, we assume $p(V)$ is faithful, i.e., $(X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep in } \mathcal{G}} \iff (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$.

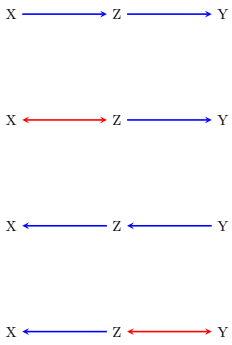
A graph $\mathcal{G} = (V, E)$ is an acyclic directed mixed graph (ADMG) if there are at most two edges between each pair of vertices, E only contains directed or bidirected edges, and G has no directed cycles. Similar to d-separation, *m-separation* allows us to read off conditional independences from an ADMG. There are three triplets that define m-separation: forks, chains, and colliders. Conditional independences implied by each triplet are described in Figure 3. Notice that forks and chains are open, but become blocked upon conditioning. On the other hand, colliders are blocked, but become open upon conditioning. A path from X to Y is blocked if there exists a blocked fork or chain, or a collider that is not active along the path. X and Y are m-separated given Z in an ADMG $\mathcal{G}$ if all paths from X to Y in $\mathcal{G}$ are blocked by Z.

Given a DAG with unobserved variables $\mathcal{G}(V \cup U)$, a unique latent projection ADMG $\mathcal{G}(V)$ can be obtained using the latent projection operator. Then, the latent projection ADMG $\mathcal{G}(V)$ preserves all causal information in the DAG $\mathcal{G}(V \cup U)$. Moreover, the latent projection ADMG $\mathcal{G}(V)$ also preserves all statistical information on variables in V. Specifically, given a distribution $p(V \cup U)$ that factorizes with respect to a DAG $\mathcal{G}(V \cup U)$, the marginal distribution $p(V)$ satisfies the global Markov property with respect to the latent projection ADMG $\mathcal{G}(V)$, i.e., $(X \perp\!\!\!\perp Y \mid Z)_{\text{m-sep in } \mathcal{G}} \implies (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$, and $(X \perp\!\!\!\perp Y \mid Z)_{\text{m-sep in } \mathcal{G}(V)} \iff (X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep in } \mathcal{G}(V \cup U)}$ for all disjoint subsets X, Y, Z of V (Verma and Pearl, 1990). As we restrict our study to only faithful distributions, it follows that $(X \perp\!\!\!\perp Y \mid Z)_{\text{m-sep in } \mathcal{G}(V)} \iff (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V \cup U)}$.

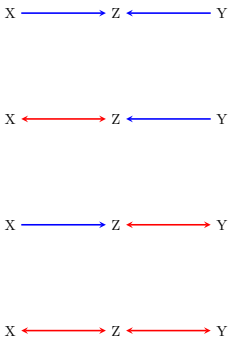
Appendix B. Partial Ancestral Graphs

A partial ancestral graph (PAG) is a graph that represents a Markov equivalence class of ADMGs (i.e., a set of ADMGs that imply the same conditional independences via m-separation). There are four types of edges in a PAG: $\rightarrow$, $\leftrightarrow$, $\circ\!\!\rightarrow$, and $\circ\!\!-\!\!\circ$. A $\circ\!\!\rightarrow$ edge between A and B implies that either A is a cause of B, or a latent variable is a cause of A and B, or both. A $\circ\!\!-\!\!\circ$ edge between A and B implies that exactly one of the following holds: (a) A is a cause of B, (b) B is a cause of A, (c) a latent variable is a cause of both A and B, (d) both a and c, or (e) both b and c (Center for Causal Discovery, n.d.).

(a) Fork: $X \not\perp\!\!\!\perp Y$, $X \perp\!\!\!\perp Y \mid Z$.



(b) Chain: $X \not\perp\!\!\!\perp Y$, $X \perp\!\!\!\perp Y \mid Z$.



(c) Collider: $X \perp\!\!\!\perp Y$, $X \not\perp\!\!\!\perp Y \mid Z$ or D where D contains at least one descendant of Z.

Figure 3: Conditional Independences Implied by the Three Triplets That Define m-separation.

Appendix C. Manual Processing of Edges in the Output PAG

Due to computational limit, we could only perform 50 bootstraps. Therefore, we set other parameters to be less restrictive. That is, it was harder for the algorithm to reject the existence of an edge. As a result, the output graph was relatively dense. Then, we manually processed the output as follows. Since no factor can cause whether a day is a holiday or not, we deleted any incoming edges $\rightarrow$ to all HOLIDAYS:i variables. For the same reason, any $\circ\!\!\rightarrow$ edge from any HOLIDAYS:i variable was interpreted as a $\rightarrow$ edge. We also deleted the CRUDEOIL:0 $\rightarrow$ VACCINES:0 edge since it would create the CRUDEOIL:0 $\rightarrow$ VACCINES:0 $\rightarrow$ CASES:0 $\rightarrow$ CRUDEOIL:0 cycle. Out of the three edges in this cycle, the CRUDEOIL:0 $\rightarrow$ VACCINES:0 edge is the least possible. Similarly, we deleted the CRUDEOIL:1 $\rightarrow$ VACCINES:1 edge.

Appendix D. Another Definition of the Backdoor Criterion

Richardson and Robins (2013) proposed an equivalent definition of the backdoor criterion through the *Single World Intervention Graphs (SWIGs)*. Given an ADMG $\mathcal{G}$, the SWIG $\mathcal{G}(a)$ corresponding to a hypothesized intervention of setting $A = a$ can be constructed

by first splitting A into a "random" vertex A and a "fixed" vertex a. In $\mathcal{G}(a)$, A inherits all incoming edges ($\circ \leftrightarrow A$, or $\circ \rightarrow A$), and a inherits all outgoing edges ($A \rightarrow \circ$) in $\mathcal{G}$. Moreover, all descendants V_i of a in $\mathcal{G}(a)$ are converted to potential outcomes $V_i(a)$. Using the SWIGs, the backdoor criterion can be stated as follows. A set of variables Z satisfies the backdoor criterion with respect to treatment A and outcome Y in an ADMG $\mathcal{G}$ if A and $Y(a)$ are m-separated given Z, and Z does not contain any potential outcomes in the SWIG $\mathcal{G}(a)$.

Using the above definition, $Z = \{\text{CASES:3}\}$ is a valid adjustment set to estimate the causal effects of $A = \text{CASES:2}$ on $Y = \text{CORN:0}$ since in the corresponding SWIG, the only incoming edge into CASES:2 is a directed edge from CASES:3, and CASES:3 is neither a collider nor a descendant of a collider. However, the causal effects of CASES:3 on CRUDEOIL:0 is unidentifiable because in the associated SWIG, there are no factual variables that block the CASES:3 $\leftrightarrow$ CASES:0(cases:3) $\rightarrow$ CRUDEOIL:0(cases:3) path.

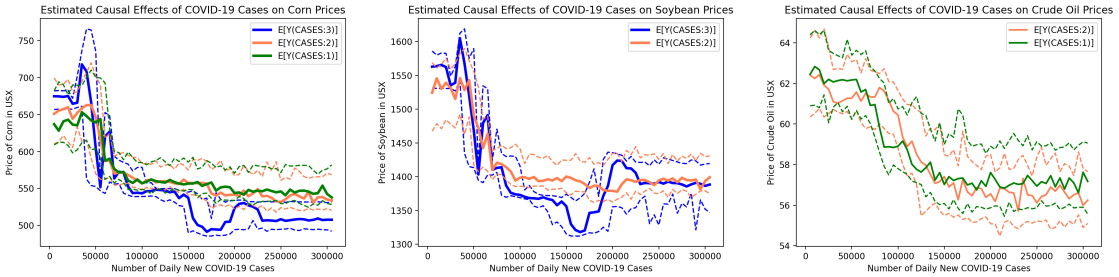
Appendix E. Backdoor Adjustment Sets

A	***Y***	***Z***
CASES:3	CORN:0	{}
CASES:2	CORN:0	{CASES:3}
CASES:1	CORN:0	{CASES:2, CASES:3}
CASES:3	SOYBEAN:0	{}
CASES:2	SOYBEAN:0	{CASES:3, HOLIDAYS:3}
CASES:1	SOYBEAN:0	—
CASES:3	CRUDEOIL:0	—
CASES:2	CRUDEOIL:0	{CASES:3, HOLIDAYS:3}
CASES:1	CRUDEOIL:0	{CASES:2, CASES:3, HOLIDAYS:2, VACCINES:1, VACCINES:3}

Table 1: Backdoor Adjustment Sets Z for Estimating the Causal Effects of A on Y.

Appendix F. Causal Effect Estimations Using Non-linear Models

Our estimates in Section 3 assumed all models are linear. In this section, we assess the sensitivity of the previous estimates by examining non-linear models. Specifically, we used the Random Forest Regressor estimator from *scikit-learn* (Pedregosa et al., 2011). In the period of 178 days that we considered, there were 58 days in which commodity prices were not recorded. This issue of missing data was even exacerbated when we created the time lag data set. As a result, after dropping the null rows to fit the non-linear models, there were only 40 rows left in our data set. This small sample size might bias the estimates. However, given that the vaccines were only distributed within a year when this work was written, we believe it is still worthwhile to examine non-linear models in our estimations. As before, we calculated the causal effects corresponding to each value of the number of daily new cases in the range (5000, 310000) with steps of 5000. For each value, we also estimated the 95% confidence interval with 50 bootstraps. Using this procedure, we obtained estimates as in Figure 4.



(a) Corn (b) Soybean (c) Crude Oil

Figure 4: Estimated Causal Effects of COVID-19 Cases on Commodity Prices Using Random Forest Regressor Estimator

The figures suggest that overall, the numbers of cases on days d_{t-3}, d_{t-2}, and d_{t-1} still have negative effects on day d_t's commodity prices. However, in the two instances where the causal effects of the numbers of cases on day d_{t-3} on commodity prices are identifiable (i.e., corn and soybean), the results suggest that the causal effects reach their highest points when the numbers of cases are around 175000 before becoming weaker as the numbers of cases increase. Interestingly, for all commodities considered, the predicted prices obtained using non-linear models are generally higher than the results from linear models, suggesting weaker estimated effects of the pandemic on commodity prices. Yet, as before, in contrast to when the numbers of cases are low, when the numbers of cases are high, cases recorded on more distant days have stronger effects on day d_t's prices, indicating that commodity markets might take a longer time to respond to more severe shocks.