Environmental Science & Technology

pubs.acs.org/est

Article

Machine Learning-Aided Causal Inference Framework for Environmental Data Analysis: A COVID-19 Case Study

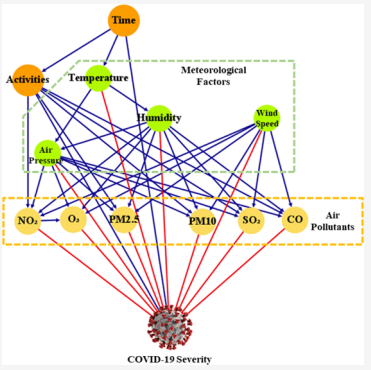
Qiao Kang,[§] Xing Song,[§] Xiaying Xin,[§] Bing Chen,[*] Yuanzhu Chen, Xudong Ye, and Baiyu Zhang

Cite This: *Environ. Sci. Technol.* 2021, 55, 13400−13410

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Links between environmental conditions (e.g., meteorological factors and air quality) and COVID-19 severity have been reported worldwide. However, the existing frameworks of data analysis are insufficient or inefficient to investigate the potential causality behind the associations involving multidimensional factors and complicated interrelationships. Thus, a causal inference framework equipped with the structural causal model aided by machine learning methods was proposed and applied to examine the potential causal relationships between COVID-19 severity and 10 environmental factors (NO_2, O_3, PM2.5, PM10, SO_2, CO, average air temperature, atmospheric pressure, relative humidity, and wind speed) in 166 Chinese cities. The cities were grouped into three clusters based on the socio-economic features. Time-series data from these cities in each cluster were analyzed in different pandemic phases. The robustness check refuted most potential causal relationships' estimations (89 out of 90). Only one potential relationship about air temperature passed the final test with a causal effect of 0.041 under a specific cluster-phase condition. The results indicate that the environmental factors are unlikely to cause noticeable aggravation of the COVID-19 pandemic. This study also demonstrated the high value and potential of the proposed method in investigating causal problems with observational data in environmental or other fields.



KEYWORDS: *structural causal model, causal inference, COVID-19, machine learning, air pollutant, meteorological factor*

INTRODUCTION

After 12 months of the first COVID-19 case report in Wuhan, China,[1] a new SARS-CoV-2 variant was identified by the United Kingdom authorities on December 19, 2020.[2] Two months later, the new variant with potentially higher transmissibility and fatality[3] has been found in 10 Canadian provinces[4] as well as in the United States and other 91 countries.[5] As of June 30, 2021, multiple SARS-CoV-2 variants are circulating globally,[6] and the COVID-19 pandemic has claimed 3.93 million lives.[7] The urgency of suppressing the COVID-19 pandemic has never been greater.[8,9] Although SARS-CoV-2 can only be viable in aerosol for a limited period (3−16 h),[10,11] COVID-19 was still reported to be capable of transmitting through the dissemination of suspended infectious aerosols[12−14] in addition to unprotected contact with infectious individuals[15−17] and fomite (contaminated surface).[10,18,19] Thus, as an effort to tackle the pandemic, the scientific community is examining factors associated with the pandemic, including environmental conditions such as meteorological factors and air pollution. As a result, correlations of air pollution and meteorological factors with COVID-19 severity have been reported worldwide.[20−33] These reported links lead to the speculation that some causal mechanisms may exist behind the associations. For instance, low wind speed may promote the suspension of infectious particles;[34,35] exposure to air pollution may compromise people's immune systems and further induce a higher infection rate.[36−38]

Though no consensus has been reached,[39−41] researchers deployed various approaches on several types of observational data, searching for clues to the causal links' existence. However, some issues are emerging, while the research is becoming increasingly in-depth. The first issue is the confusion between correlation and causation.[42] Due to ambiguous hypotheses and similarities between the two concepts, misidentifying the correlations as causalities is common. Another issue is the inappropriate use of conventional methods without the support of prior knowledge, which was constantly being overlooked in the existing studies. These methods include time-series analysis such as the Granger causal test[43−45] and machine learning models.[46,47] Besides, some essential confounders, such as social-economical factors and inbound traffic flows from the pandemic epicentre,[48−51] were commonly omitted in the existing studies. Many spurious correlations could emerge due to such omission.[52] Finally, among all the studies that attempted to estimate the causal effects quantitatively, few incorporated methods to refute the relationships or falsify the assumptions.

Received: April 5, 2021
Revised: August 22, 2021
Accepted: August 23, 2021
Published: September 24, 2021

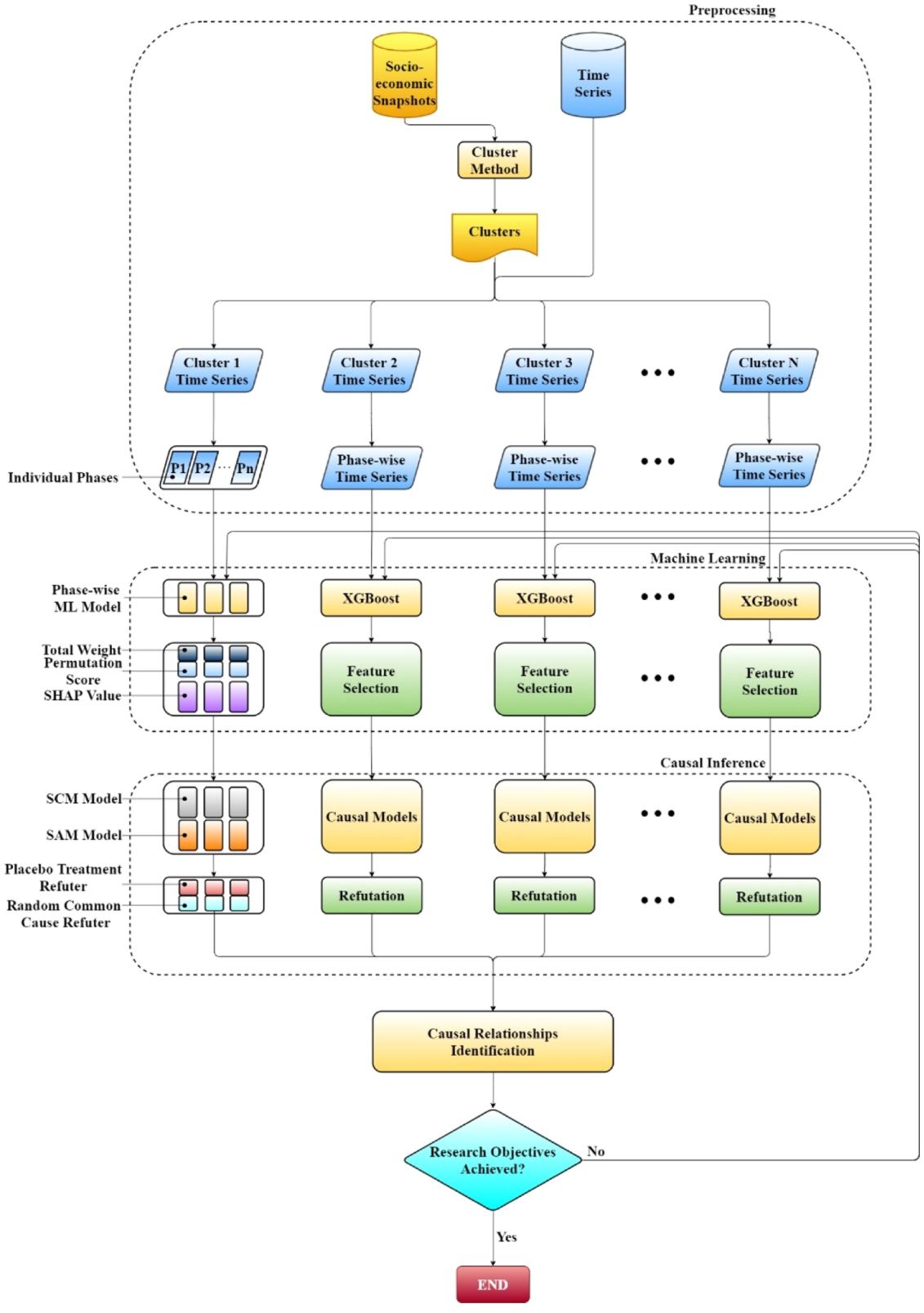


Figure 1. Schematic diagram of the causal framework. Note: "cluster 1" shows individual methods/parts in the component by making all the components transparent.

The step is quite essential, especially when the ground truth of the causal links is unknown.[53] The above issues are not isolated but are commonly seen in environmental studies when a causal question has to be answered based on observational data without the aid of randomized experiments,[54] such as in policy impact evaluation and climate change attribution.[55,56] Thus, environmental studies can greatly benefit from a new framework for causal inference.

Thanks to the growing research on causal inference in the statistics and artificial intelligence field during the past few

https://doi.org/10.1021/acs.est.1c02204
Environ. Sci. Technol. 2021, 55, 13400−13410

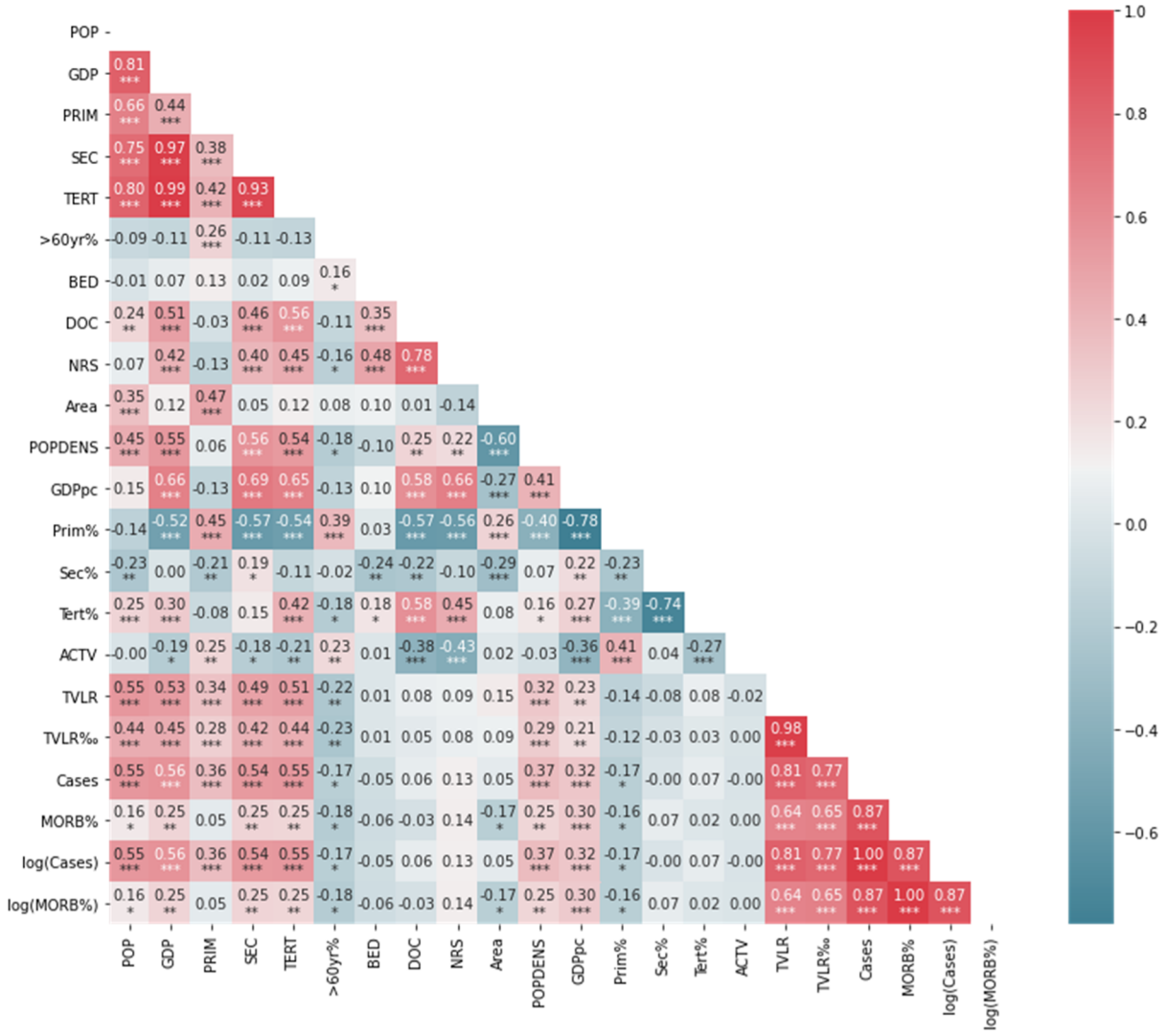


Figure 2. Spearman correlation heatmap of the snapshot data set with statistical significance. Note: * indicates—$P \leq 0.05$, ** indicates—$P \leq 0.01$, and *** indicates—$P \leq 0.001$. POP: population; PRIM/SEC/TERT: primary, secondary, and tertiary sector of GDP; >60 yr %: elderly population percentage; BED/DOC/NRS: hospital beds/registered medical doctors/registered nurses per 1000 population; TVLR: inbound travelers from Wuhan; TVLR‰: Wuhan travelers per 1000 population; and ACT: average degree of activeness.

years,[57−59] some novel and effective methods were born and thrived from the rich discussions, enabling us to develop a new causal framework with the desired features. To build such a framework which can conduct causal reasoning from observational data, among the most discussed methods, we selected the structural causal model (SCM), one of the most established causal inference methods[48] as the causal engine. The method has the following characteristics: (1) it uses prior knowledge regarding the data-generating process as an input. (2) Intervention (i.e., purposely modify the condition to observe the response of the result) is a supported action in SCM in the form of do-calculus. The two features enabled SCM to perform causal reasoning from observational data. On the other hand, since the framework needs to be resilient to some common characteristics in environmental data sets such as frequent outliers, non-normal distribution, and limited sample size,[60] some functional components were also embedded to ensure the applicability and adaptivity of the framework. These components include (a) a backup prior-knowledge extractor in case the prior knowledge is limited or not accessible; (b) a feature selection component, which can significantly reduce the computational time while acquiring data insights into the causal reasoning; and (c) a refutation module that can test the proposed causal relation's robustness, which is especially helpful when the relationship is unconfirmed.

This study aims to propose a causal inference framework and to investigate the potential causal relationships between COVID-19 severity and environmental factors, including six air pollution indicators and four meteorological factors, in 166 Chinese cities. The social-economic diversity among these cities makes China an ideal study area for investigating the causal relationships under multiple socio-economical scenarios.[61] This study attempts to provide evidence for causal inference about environmental factors and COVID-19 severity to support the decision-making process for global and regional pandemic countermeasures in the current phase of the COVID-19 pandemic and to establish an applicable and robust causal recovery framework for the environmental science community.

MATERIALS AND METHODS

Framework Design, Study Area, and Data Sources. The workflow of the causal inference framework in this study is

https://doi.org/10.1021/acs.est.1c02204
Environ. Sci. Technol. 2021, 55, 13400−13410

illustrated in Figure 1. This framework is suitable for environmental causal reasoning problems under different socio-economic conditions. During the data processing phase in the framework, socio-economic data will be used to generate clusters of different administrative units (i.e., countries, provinces, cities, and so forth) with similar socio-economic conditions. Time-series data from each administrative unit will be assigned to the corresponding clusters. When the trends in the target time series are obvious, need-based time segmentation can be further applied. In that case, a time-series segment (e.g., P_1−P_n in Figure 1) will become the smallest unit for further analysis. Each segmentation will be analyzed by the machine learning module, followed by the causal inference module. Models for each unit will be trained by a selected machine learning algorithm and then interpreted by multiple metrics for feature selection. The interpretation can also support causal relationship identification in later procedures. Data will be input along with a directed acyclic graph (DAG) in the causal inference module. If no graph can be provided due to limited knowledge, a backup method can be called to generate a quasicausal relationship graph as the DAG input. After quantitative estimation, the potential causal relationships will undergo two refutation processes as a robustness check.

This study investigated the causal relationship between environmental factors and COVID-19 severity. One hundred sixty-six key air quality monitoring cities recognized by the Ministry of Environmental Protection of China were selected as the study area due to their representativeness in their corresponding regions as well as their complete COVID-19 case and environmental monitoring data. Compared to the original 168 key monitoring cities, we excluded Wuhan (the epicenter) and Dongying (had zero cases during the first wave pandemic) for the study. The socio-economic data were from each city's Statistical Bulletin/Yearbook or directly acquired from city-level Civil Affairs Bureaus. Most environmental data were acquired from the China National Environmental Monitoring Center. The COVID-19-related data were obtained from the Chinese Center for Disease Control and Prevention. The numbers of inbound travelers from Wuhan and the degree of activeness in each city were calculated based on Baidu location-based service (LBS) data.

Measures of Variables and Data Processing. Two data sets, the "snapshot" data set and the time-series data set, were composed and investigated. The "snapshot" data set is a cross-sectional data set consisting of all the 166 cities' socio-economic profiles before the 2020 Spring Festival. The features include the following:

- The inhabited populations (1000 people)
- Population density (people per km^2)
- Area of the cities (km^2)
- Total gross domestic product (GDP in billion USD)
- GDP by sectors (primary, secondary, and tertiary in billion USD) and the corresponding percentages
- GDP per capita (1000 USD)
- Elderly population percentage (over 60 years old)
  - This feature was added since senior citizens are vulnerable to COVID-19 due to their fragile immune systems.[62]
- Numbers of hospital beds, medical doctors, and nurses per 1000 population
  - The public healthcare development indexes were added, considering that the COVID-19 patients need timely and intensive care.
- Transient population flow from Wuhan (1000 people)
  - The 15-day accumulative inbound travelers from Wuhan to all the 166 selected cities before the pandemic outbreak were estimated according to the intracity migration index (IMI), a data set developed based on Baidu's LBS. The data set has also served the same purpose in previous related studies.[63]
- Average degree of activeness before the outbreak
  - Based on the IMI data mentioned above, the degree of activeness in each city from January 10 to January 23, 2020, was used to calculate the average value.

The snapshot data set was used as the basis for the following clustering process. A correlation heatmap of the snapshot data set is given in Figure 2.

The time-series data set comprises 13 time series for each selected city during the first wave of the pandemic. The 76-day lockdown period of Wuhan was selected as the time span for all the time series, which started from January 22 to April 8, 2020. The time series include the following:

- Six air pollutants' concentrations (PM2.5, PM10, SO_2, NO_2, and O_3 in $\mu g/m^3$ and CO in mg/m^3)
- Average air temperature (TEMP in °C)
- Relative humidity (HMD in percentage)
- Atmospheric pressure (PRES in hpa)
- Wind speed (WSPD in m/s)
- Daily degree of activeness (ACTV)

There are also two other features in the data set:

- Daily newly confirmed COVID-19 cases (CASES)
  - In contrast to the moving average method, confirmed case time series were processed with a 3-day moving sum strategy for a more intuitive analysis process.
- Elapsed days (DAYS)
  - Counted from the first day with a confirmed COVID-19 case in each city.

We applied 3-day moving average to the above time series to reduce the random noises in the data set while focusing on the potential short-term effects.[64,65] In the time-series data set, each feature's mean values in the corresponding city were used to impute a small portion (~0.23%) of missing values. A statistical description of two data sets and each feature's corresponding data source have been listed in Table S1.

Models and Data Analysis. In this study, socio-economical factors should be considered essential since they are decisive for the human activity patterns and many other pandemic critical factors in different cities.[66] Thus, the selected 166 cities were clustered based on the mentioned snapshot data set. Such clusters can provide insights into the pandemic severity under different conditions and avoid the possible "Simpson's paradox".[67] Principal component analysis (PCA)[68] was selected as the dimensionality reduction technique to ensure a better clustering performance and resilience to the "curse of dimensionality". Since PCA is sensitive to the variance within the data set, the snapshot data set was standardized before the procedure to minimize the impact of different feature scales and variance.[69−72] After a series of experiments, we noticed that over 62% of the variance could be explained with only three PCAs,

https://doi.org/10.1021/acs.est.1c02204
Environ. Sci. Technol. 2021, 55, 13400−13410

which is sufficient for further analysis. Thus, the original 18-dimensional data set was compressed to 3-dimensional by selecting three principal components. The contribution of each feature to individual principal components is given in Figure S1, along with the explained variance ratios. For city clustering, the time-proven k-means method was selected due to its effectiveness and efficiency.[73] The "elbow method"[74] was used to determine the numbers of the cluster. Three was selected as the cluster number based on the result of the elbow method, which is given in Figure S2. Due to the assumption that each factor may have different behaviors during different pandemic phases,[75] the cluster-wise time series were further divided into two segments by specific demarcation points. The splitting dates were February 3, 2020, for cluster 1 and February 6, 2020, for clusters 2 and 3, corresponding to each cluster's pandemic spreading and postpeak phases. The corresponding trends can also be observed in Figure S5a. Hereafter, nine sub-data sets were generated from the time-series data set based on the pandemic development perspective (overall, spreading phase, and postpeak phase) and three city clusters. Each sub-data set will be further analyzed by both causal inference models and machine learning algorithms.

The XGBoost algorithm was selected for feature selection and knowledge extraction. A detailed discussion on the selection process can be found in the Supporting Information as Text S1. Models were trained with the aid of k-fold cross-validation. The cross-validation method splits an existing data set into k number of folds, where each fold will be used as a testing set against the rest of the data. In this way, the impact of overfitting or sampling bias can be minimized. In the study, we set $k = 5$ based on the number of instances (700−8500) in nine sub-data sets. The hyperparameter ranges for GridSearchCV, the final hyperparameter values, and r^2 of each trained model are given in Table S4. After the training process, two feature importance evaluation metrics, total gain and permutation score, were used to interpret each trained model. The total gain in an XGBoost model is the product of a feature's gain score and the frequency of the feature being used for node splitting when constructing the model. The permutation score is another useful metric defined as the decrease of the model performance when a single feature is randomly shuffled.[76] One common shortfall of the two metrics is their weakness in identifying if features' contribution is positive or negative. Thus, SHAP was introduced as another method for interpretation as it can indicate the feature contribution's direction (i.e., positive or negative), which enabled the researchers to select features of interest for further analysis.[77]

For causal inference, constructing a graphical causal model in the form of a DAG is the first step of the SCM.[53,78] Each DAG node represents a variable, and an arrow indicates a causal link, either an assumed or confirmed one. The graph allows users to explicitly introduce prior knowledge and untested assumptions about the data-generating process. Figure 3 shows the graphic causal model for this study. Proven causal relationships are given as blue arrows. Causal relationships among elapsed days, degree of activeness, and COVID-19 cases were considered, as well as those between meteorological factors and the air pollutants. The transformation from NO_2 to O_3[79] and the interactions among air temperature, relative humidity, wind speed, and atmospheric pressure[80] were also taken into account. Unproven causal links included potential causal relationships between the COVID-19 cases and different environmental factors. After creating the DAG, the average treatment effects (ATEs)[81,82] of the potential causal relationships could be estimated. The default-incorporated algorithm is a linear estimator. To capture nonlinear causal effects, we selected the DMLOrthoForest[83] method to provide nonlinear estimations. The linear estimator has been preserved as a complement to the machine learning-based estimator. A more detailed introduction to the do-calculus and SCM can be found in the Supporting Information. Note that causal estimation was conducted on a normalized copy of the data set for enabling the comparison between different features.

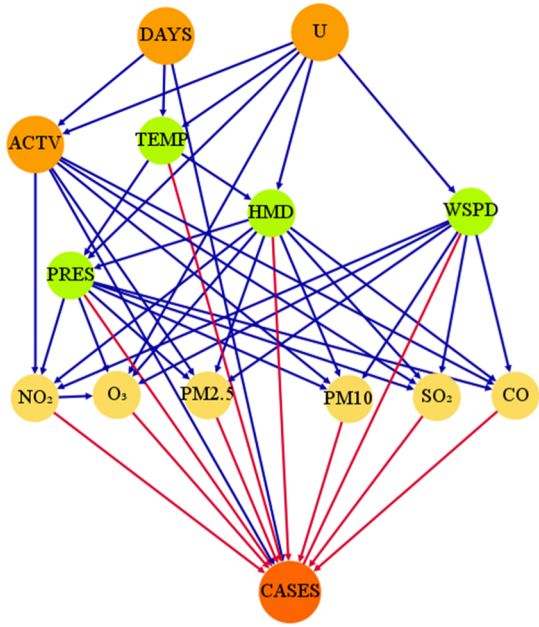


Figure 3. Causal relationships among environmental factors and COVID-19 cases. All proven causal links are given as blue arrows, and unproven causal relationships are marked by red arrows. Note: ACTV—daily degree of activeness; DAYS—elapsed days; HMD—relative humidity; PRES—atmospheric pressure; TEMP—average air temperature; U—unobserved confounders; and WSPD—wind speed.

Structural agnostic modeling (SAM) was deployed as the backup knowledge extractor. The neural-network-based algorithm has been proven robust in recovering nonlinear causal relationships between continuous variables with a superb performance.[84] In this case study, though the DAG was constructed, SAM was used to generate a weighted adjacency matrix from the data set. Each weight in the matrix represents the corresponding causal relationship's strength. The matrix can be used to support the final decision-making about the causal relationship from another perspective.

In order to test the robustness of an assumed causal relationship, two refutation methods, adding random common cause (RCC) and placebo treatment (PT), were selected to test the robustness of each causal relationship. RCC adds an independent random variable as a common cause to the data set, and PT replaces the chosen treatment variable's value with some independent random values. For a robust relationship, its estimated effect is expected to remain stable under the RCC refutation test. On the contrary, effects estimated under the PT test should be zero instead of the original value.[53] Based on the two refutation methods, a four-level robust check criterion was set in the case study to ensure the robustness of a causal estimation. First, an estimate must pass both refutation tests, PT and RCC, to be considered. Being more specific, the estimates under the RCC test should be within 10% variance of the original value, which is the first level. Then, another three

https://doi.org/10.1021/acs.est.1c02204
Environ. Sci. Technol. 2021, 55, 13400−13410

tolerance thresholds (i.e., the maximum allowed variation of an estimate) will be set to evaluate the considered estimations under the RCC test. In this study, the four levels were 10 (the initial threshold), 5, and 1% and 5‰, indicating an increasingly strict criterion. A potential causal relationship should pass the 5‰ threshold to be considered robust enough.

The causal effect estimation and refutation were achieved based on the DoWhy package, a Python package specialized in providing a causal inference interface. The RCC and PT algorithms used in the framework can be found within the package.[53] The DMLOrthoForest algorithm and the SAM algorithm implemented in this study can be found in EconML[83] and Causal Discovery Toolbox,[85] respectively. A framework benchmark that applied SCM and SAM on three public data sets with known ground truth is given in the Supporting Information as a robustness check for the proposed framework.

RESULTS

Three city clusters are presented in Figure 4 with the geographical locations of all the selected cities (n = 166). The

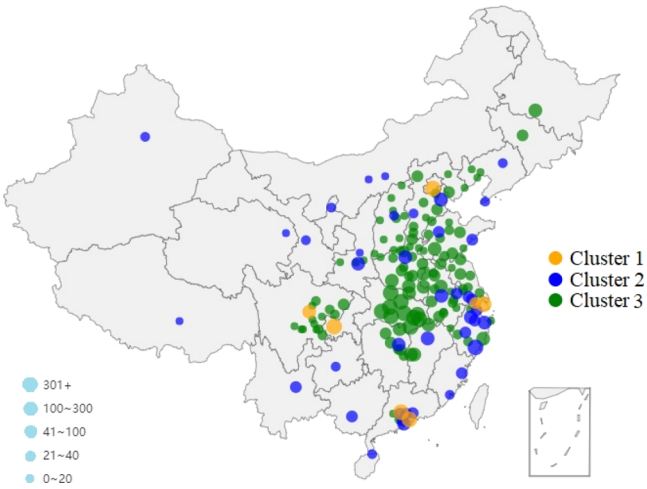


Figure 4. Selected cities' locations with the circle size indicating the total confirmed COVID-19 cases.

distribution of all cities in different clusters in the principal component space is given in the Supporting Information as Figure S3. A full city list of all three clusters is available in Table S2. In summary, cluster 1 (n = 7) comprised megacities with advantages in many socio-economic aspects. Cities in cluster 2 (n = 40) are mostly provincial capitals and other major cities, and the majority of cluster 3 (n = 119) are ordinary urbanized cities. The average feature values in different city clusters are given in Table S3. No significant difference could be found in the elderly population percentages (16.54−20.62%) and healthcare development indexes (beds per 1000: ∼6.38, doctors per 1000: ∼3.19, and nurses per 1000: ∼3.79) among the three clusters. The degrees of average activeness in the three clusters were also at the same level (∼5.09) though a relatively higher activeness degree (5.57) can be observed in cluster 3.

Figure 5 shows the feature importance and the SHAP values. Features with positive contributions in each trained machine learning model were highlighted. Both the total gain and the permutation score were normalized to a 0−1 range for easier comparison and visualization. Note that as two baseline features, elapsed days and degree of activeness were designed to be never highlighted. For elapsed days, it is noticeable that the sign of its contribution varied in different sub-data sets. It was the dominating factor among all three phases in cluster 1 with normalized feature importance above 0.60, the top contributor for the postpeak phase in cluster 2 with a feature importance of 0.35, and had lower feature importance varied from 0.05 to 0.15 for other periods in clusters 2 and 3. Similar to elapsed days, the degree of activeness also dominated in one cluster, cluster 3, with its feature importance maintained around 0.29. It reached the top in cluster 2 from the overall perspective with a feature importance of 0.23, had lower yet considerable feature importance (∼0.15) in the other two cluster 2 phases, and became less significant in cluster 1.

From the interpretation results, positive contributions of air pollutants were generally low. Most air pollutants' feature importance was below 0.05, except for a few specific pollutants under particular settings. The exceptions included SO_2 in the cluster 1 spreading phase (0.10), CO from cluster 3 overall perspective (0.06), and postpeak phase (0.14), as well as PM2.5 (0.07) from the overall perspective in cluster 3. Meteorological factors had a higher potential contribution than air pollutants, which can be observed in Figure 5. All meteorological factors had been highlighted at least once. Atmospheric pressure was highlighted among all sub-data sets nine times. Its feature importance was within a range of 0.08−0.20 in cluster 2 and cluster 3, whereas it was less significant in cluster 1 with its feature importance varying from 0.03 to 0.09. The air temperature was the second-most highlighted meteorological feature, which had been highlighted six times. When highlighted, its feature importance varies from 0.02 to 0.22. Relative humidity showed most of its observable potential contribution in cluster 3, with a feature importance ranging from 0.03 to 0.10. Wind speed had some minor contribution less than 0.03 in cluster 3.

Table 1 shows the 25 potential causal relationships that passed the initial refutation with positive ATE, as well as their behaviors when facing lower RCC tolerances. Nine out of the twenty-five relationships were about air pollution indicators. The majority of the connections were nonlinear with two linear exceptions: relative humidity (cluster 1 postpeak phase) and PM10 (cluster 3 spreading phase). The effect of reducing the RCC threshold was prominent. Decreasing the tolerance from 10 to 5% eliminated nine candidate relationships. Tolerance's dropping from 5 to 1% removed another 11 candidates. NO_2, one of the most reported air pollutants with a correlation between COVID-19 severity, did not pass the initial causal screening. As for SO_2 in cluster 1, although it had positive contributions in cluster 1 machine learning models and passed the initial round of the refutation test, it did not survive the first tolerance drop. When the tolerance dropped to 5‰, only one potential causal relationship survived: air temperature in cluster 2 spreading phase with a causal effect of 0.041.

DISCUSSION

It is noticeable that the majority of the pollutants' contributions in the machine learning models were negative. This characteristic can be easily observed from NO_2, which had significant negative contributions (0.1−0.2) and was considerably correlated with the degree of activeness (ρ = 0.47) in cluster 2. Meanwhile, the negative contributions were reflected in the estimated effects from the SCM as well: all of NO_2's nonlinear ATE values were negative. From the results, NO_2 is more likely to be an indicator of human activity in selected Chinese cities rather than a causal factor to COVID-19 cases. O_3, another air

https://doi.org/10.1021/acs.est.1c02204
Environ. Sci. Technol. 2021, 55, 13400−13410

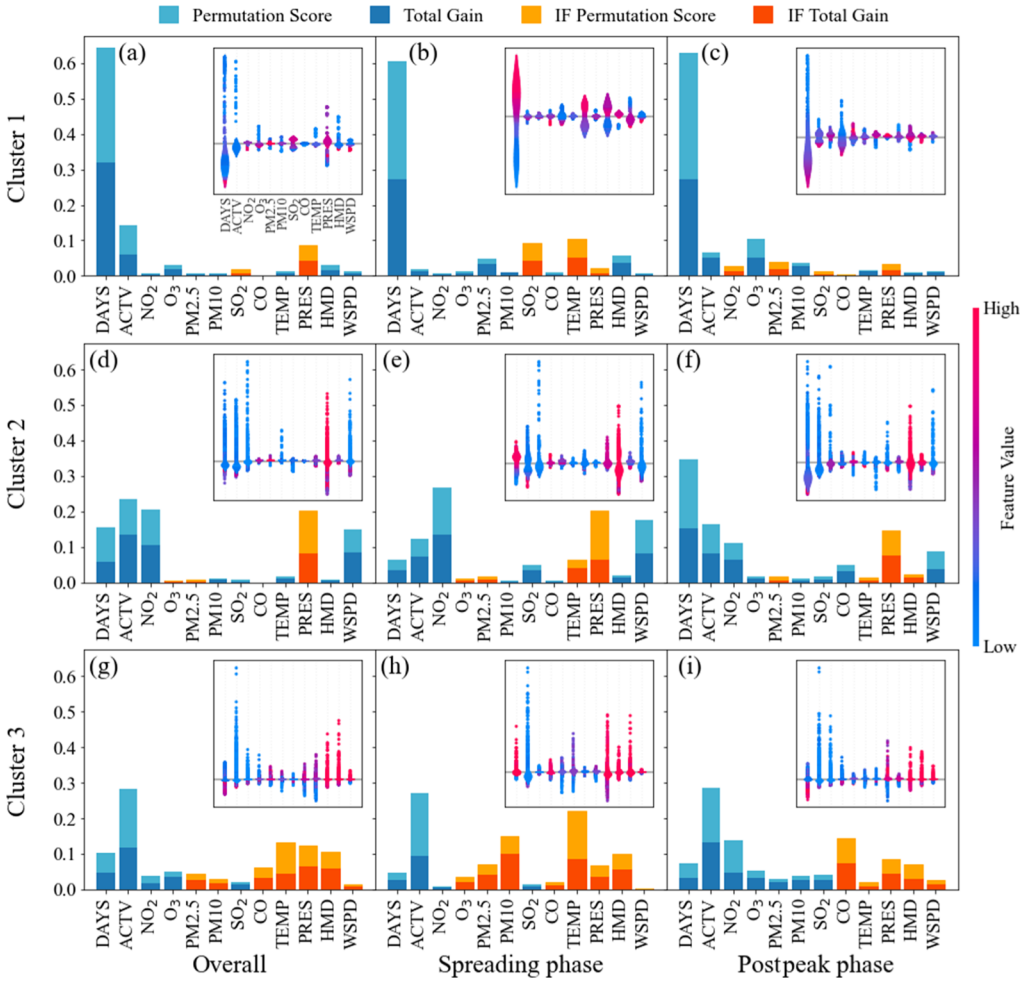


Figure 5. Feature importance and SHAP value of features in machine learning models. Features with positive SHAP values (subplots) are highlighted with orange and red in the feature importance plot. In the SHAP value, subplots blue and red indicate lower and higher feature values, respectively. Instance points above the gray zero axis indicate positive SHAP values and vice versa. SHAP value features are in the same order as in subplot (a).

pollutant that may compromise the human respiratory system,[86,87] also had some negative causal effects (four out of nine), indicating that it was unlikely to worsen the pandemic, especially considering that none of its relationships passed the final refutation. We assume that these negative contributions were due to the connections between air pollution and human activity: the primary sources of NO_2 and O_3 in China are anthropogenic activities, mostly from industrial and mobile sources.[88,89] Implementing and lifting the lockdown policies might further influence the fluctuation of the pollutants' concentrations.[90−92] Note that the genuine causal relationships among lockdown implementation, human activities, and air pollution do not guarantee observable correlations. A visualization of NO_2 and O_3's trends is given in Figure S5d,e.

Among all the relationships about PM2.5, PM10, SO_2, and CO, PM10 in the cluster 3 spreading phase was the only relationship that passed the 1% threshold refutation test with positive ATE values (0.079), though it did not pass the final refutation. As for PM2.5, both its relationships in Table 1 could not pass the 1% threshold refutation. The only CO-related causality failed the 5% threshold test in the cluster 3 postpeak phase; SO_2 in cluster 3 spreading phase failed the 1% refutation. The results indicate that the robustness of these causal relationships was not sufficient. The same deductions can be applied to the meteorological relationships, where most of them failed the second level of refutation (5%), implying their insignificance in the proposed causal problem. Though the majority of the relationships were refuted at the end, air temperature in the cluster 2 spreading phase passed the final refutation with a causal effect of 0.041. Technically, the values indicate that 1 °C air temperature increase in cluster 2 during the spreading phase will lead to approximately 0.183 new confirmed cases. However, its final RCC refutation variance was 0.00498, indicating that the temperature-case causal estimation almost failed the final refutation test (5‰ threshold). Based on all the results, though a specific causal relationship's existence cannot be completely ruled out, the discussed factors' causal effects on the COVID-19 severity are likely to be limited since the estimates reported in the study were by no means conclusions but traces of evidence of the causal links. Thus, instead of drawing conclusions, it is more reasonable to deduce that the environmental factors were unlikely to exacerbate the COVID-19 pandemic in these Chinese cities from a short-term perspective.

■ ENVIRONMENTAL IMPLICATIONS

This study demonstrated a new causal reasoning method based on observational data with the aid of causal inference models and machine learning techniques. To investigate a causal problem with observational data, we also considered the prior knowledge

https://doi.org/10.1021/acs.est.1c02204
Environ. Sci. Technol. 2021, 55, 13400−13410

Table 1. Refutation Results of Potential Impactful Environmental Factors[a]

					threshold		
city cluster	pandemic phase	feature	relationship type	ATE	5%	1%	5‰
cluster 1	overall	PRES	nonlinear	0.045	F		
	spreading	SO_2	nonlinear	0.476	F		
	postpeak	HMD	linear	0.051	P	P	F
cluster 2	overall	PM2.5	nonlinear	0.323	P	F	
		O_3	nonlinear	0.012	P	F	
		PRES	nonlinear	0.055	P	F	
	spreading	PRES	nonlinear	0.118	P	F	
		TEMP	nonlinear	0.041	P	P	P
	postpeak	PM2.5	nonlinear	0.290	P	F	
		HMD	nonlinear	0.046	F		
		PRES	nonlinear	0.066	P	F	
		TEMP	nonlinear	0.030	P	F	
cluster 3	overall	O_3	nonlinear	0.016	P	P	F
		HMD	nonlinear	0.018	P	F	
		PRES	nonlinear	0.008	F		
		WSPD	nonlinear	0.009	P		
	spreading	PM10	linear	0.079	P	P	F
		SO_2	nonlinear	0.057	P	F	
		HMD	nonlinear	0.046	F		
		TEMP	nonlinear	0.073	F		
		WSPD	nonlinear	0.005	P	F	
	postpeak	PM10	nonlinear	0.035	F		
		CO	nonlinear	0.092	F		
		PRES	nonlinear	0.016	F		
		WSPD	nonlinear	0.022	P	F	

[a]Nonlinear: estimated by DMLOrthoForest; linear: estimated by the linear estimator; ATE; threshold: the maximum allowed variation of an estimate used to evaluate the RCC refutation results; P: pass; F: fail. PRES: atmospheric pressure; HMD: relative humidity; TEMP: air temperature; and WSPD: wind speed. All the estimates in the table passed the RCC test with a 10% threshold.

an indispensable part of the system. In the case study, the information about socio-economic and temporal factors was brought into the equation by explicitly identifying interrelations between variables in the DAG and slicing the data through multiple data-processing techniques such as city clustering and phase-wise analysis. Through the observational data from 166 Chinese cities, we examined most of the reported potential causal relationships between environmental factors and COVID-19 severity from a short-term perspective with the proposed causal inference framework. Based on the results, we refuted most of the estimations of the links (89 out of 90) under nine different cluster-phase settings. However, there was still one estimation of air temperature that passed the final robustness check under a specific cluster-phase condition, indicating the existence of a possible short-term causality. Nevertheless, the results showed that the impact caused by the environmental factors on the severity of COVID-19 was limited across all three clusters. Commonly discussed factors such as rational policy-making, sufficient public awareness, and effective quarantine and isolation are still crucial for containing the ongoing COVID-19 pandemic.

It is worth noting that the proposed framework might be less efficient when having a data set with the following characteristics: (1) low dimensionality, (2) discrete variables, and (3) a lack of accessible prior knowledge regarding the problem. On the other hand, this showed the value of incorporating the machine learning-based feature selection module and the backup causal network recovery tool in the framework since they can provide alternative interpretation and feature screening processes and further enhance the framework's versatility.

Instead of providing a one-size-fits-all solution, the framework aimed to provide a novel approach aided by contemporary computational methods for causal analysis with environmental data. With the volume of environmental data surging, many environmental areas such as risk assessment,[93] sustainable development,[94,95] and environmental econometric[96] have been advanced significantly through the growing and expanding use of data science tools. Investigating causal questions based on observational data will become more valuable and highly desired in the environmental field. However, no matter how much the causal inference tools evolved, the results will always require careful examination and interpretation based on both scientific knowledge and practical meanings. We hope that the proposed method could elicit further discussion and research on environmental causal inference.

DATA AVAILABILITY

All the analyzed and processed data in the study can be found along with results of the case study in the GitHub repository: https://github.com/kangqiao-ctrl/EnvCausal.

CODE AVAILABILITY

All the scripts used in the study for data processing and analyzing are available in the form of .py or .ipynb files in the following GitHub repository: https://github.com/kangqiao-ctrl/EnvCausal.

https://doi.org/10.1021/acs.est.1c02204
Environ. Sci. Technol. 2021, 55, 13400−13410

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.1c02204.

Additional background information about the model selection process, XGBoost algorithm, do-calculus and SCM, method benchmark, auxiliary experiments, data set metadata, model hyperparameters, and SAM results,visualization of PCA, K-means, feature importance from an auxiliary experiment, and data of six time series (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Bing Chen − *Northern Region Persistent Organic Pollution Control (NRPOP) Laboratory, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's A1B 3X5 Newfoundland and Labrador, Canada;* orcid.org/0000-0003-1041-4525; Email: bchen@mun.ca

Authors

Qiao Kang − *Northern Region Persistent Organic Pollution Control (NRPOP) Laboratory, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's A1B 3X5 Newfoundland and Labrador, Canada;* orcid.org/0000-0002-4599-3929

Xing Song − *Northern Region Persistent Organic Pollution Control (NRPOP) Laboratory, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's A1B 3X5 Newfoundland and Labrador, Canada;* orcid.org/0000-0003-0553-1196

Xiaying Xin − *Northern Region Persistent Organic Pollution Control (NRPOP) Laboratory, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's A1B 3X5 Newfoundland and Labrador, Canada*

Yuanzhu Chen − *School of Computing, Queen's University, Kingston K7L 2N8 Ontario, Canada*

Xudong Ye − *Northern Region Persistent Organic Pollution Control (NRPOP) Laboratory, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's A1B 3X5 Newfoundland and Labrador, Canada*

Baiyu Zhang − *Northern Region Persistent Organic Pollution Control (NRPOP) Laboratory, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's A1B 3X5 Newfoundland and Labrador, Canada*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.1c02204

Author Contributions

§These authors contributed equally to this work.

Author Contributions

Q.K., X.S., X.X., B.C., and B.Z. conceived the study. Q.K. and Y.C. wrote the program script, X.S. and X.Y. collected and processed the data, Q.K., X.S., and X.X. wrote the first version of the manuscript, and B.C., B.Z., and Y.C. contributed to subsequent revisions. All authors contributed to analysis and result interpretation.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Special thanks for the support from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Foundation for Innovation (CFI), and the Compute Canada and the Chinese Scholarship Council (CSC). We are particularly grateful to the editors and the anonymous reviewers for their insightful comments and suggestions.

■ REFERENCES

(1) Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; Niu, P.; Zhan, F.; Ma, X.; Wang, D.; Xu, W.; Wu, G.; Gao, G. F.; Tan, W. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **2020**, *382*, 727−733.

(2) Kirby, T. New Variant of SARS-CoV-2 in UK Causes Surge of COVID-19. *Lancet Respir. Med.* **2021**, *9*, e20−e21.

(3) NERVTAG. *NERVTAG Paper on COVID-19 Variant of Concern B.1.1.7; New and Emerging Respiratory Virus*; Threats Advisory Group, 2021; p 9.

(4) Thompson, N. More contagious U.K. COVID-19 variant now found in all 10 provinces. https://globalnews.ca/news/7640125/coronavirus-canada-update-feb-13/.(accessed 02 03, 2021).

(5) O'Toole, Á.; Hill, V.; Pybus, O.; Watts, A.; Bogoch, I. Tracking the International Spread of SARS-CoV-2 Lineages B.1.1.7 and B.1.351/501Y-V2. https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592.(accessed 02 03, 2021).

(6) Walensky, R. P.; Walke, H. T.; Fauci, A. S. SARS-CoV-2 Variants of Concern in the United States—Challenges and Opportunities. *JAMA* **2021**, *325*, 1037−1038.

(7) Dong, E.; Du, H.; Gardner, L. An Interactive Web-Based Dashboard to Track COVID-19 in Real Time. *Lancet Infect. Dis.* **2020**, *20*, 533−534.

(8) Mofijur, M.; Fattah, I. M. R.; Alam, M. A.; Islam, A. B. M. S.; Ong, H. C.; Rahman, S. M. A.; Najafi, G.; Ahmed, S. F.; Uddin, Md. A.; Mahlia, T. M. I. Impact of COVID-19 on the Social, Economic, Environmental and Energy Domains: Lessons Learnt from a Global Pandemic. *Sustainable Prod. Consumption* **2021**, *26*, 343−359.

(9) Nordhagen, S.; Igbeka, U.; Rowlands, H.; Shine, R. S.; Heneghan, E.; Tench, J. COVID-19 and Small Enterprises in the Food Supply Chain: Early Impacts and Implications for Longer-Term Food System Resilience in Low- and Middle-Income Countries. *World Dev.* **2021**, *141*, 105405.

(10) van Doremalen, N.; Bushmaker, T.; Morris, D. H.; Holbrook, M. G.; Gamble, A.; Williamson, B. N.; Tamin, A.; Harcourt, J. L.; Thornburg, N. J.; Gerber, S. I.; Lloyd-Smith, J. O.; de Wit, E.; Munster, V. J. Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *N. Engl. J. Med.* **2020**, *382*, 1564−1567.

(11) Fears, A. C.; Klimstra, W. B.; Duprex, P.; Hartman, A.; Weaver, S. C.; Plante, K. S.; Mirchandani, D.; Plante, J. A.; Aguilar, P. V.; Fernández, D.; Nalca, A.; Totura, A.; Dyer, D.; Kearney, B.; Lackemeyer, M.; Bohannon, J. K.; Johnson, R.; Garry, R. F.; Reed, D. S.; Roy, C. J. Persistence of Severe Acute Respiratory Syndrome Coronavirus 2 in Aerosol Suspensions. *Emerg. Infect. Dis.* **2020**, *26*, 2168−2171.

(12) World Health Organization. *Advice on the Use of Masks in the Context of COVID-19: Interim Guidance*; World Health Organization: Geneva, June, 2020; Vol. *5*.

(13) Bourouiba, L. Turbulent Gas Clouds and Respiratory Pathogen Emissions: Potential Implications for Reducing Transmission of COVID-19. *JAMA* **2020**, *323*, 1837−1838.

(14) Mittal, R.; Ni, R.; Seo, J.-H. The Flow Physics of COVID-19. *J. Fluid Mech.* **2020**, *894*. https://doi.org/10.1017/jfm.2020.330. DOI: 10.1017/jfm.2020.330

(15) Chan, J. F.-W.; Yuan, S.; Kok, K.-H.; To, K. K.-W.; Chu, H.; Yang, J.; Xing, F.; Liu, J.; Yip, C. C.-Y.; Poon, R. W.-S.; Tsoi, H.-W.; Lo, S. K.-F.; Chan, K.-H.; Poon, V. K.-M.; Chan, W.-M.; Ip, J. D.; Cai, J.-P.; Cheng, V. C.-C.; Chen, H.; Hui, C. K.-M.; Yuen, K.-Y. A Familial Cluster of Pneumonia Associated with the 2019 Novel Coronavirus

Indicating Person-to-Person Transmission: A Study of a Family Cluster. *Lancet* **2020**, *395*, 514−523.
(16) Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; Cheng, Z.; Yu, T.; Xia, J.; Wei, Y.; Wu, W.; Xie, X.; Yin, W.; Li, H.; Liu, M.; Xiao, Y.; Gao, H.; Guo, L.; Xie, J.; Wang, G.; Jiang, R.; Gao, Z.; Jin, Q.; Wang, J.; Cao, B. Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *Lancet* **2020**, *395*, 497−506.
(17) Liu, J.; Liao, X.; Qian, S.; Yuan, J.; Wang, F.; Liu, Y.; Wang, Z.; Wang, F.-S.; Liu, L.; Zhang, Z. Community Transmission of Severe Acute Respiratory Syndrome Coronavirus 2, Shenzhen, China, 2020. *Emerg. Infect. Dis.* **2020**, *26*, 1320−1323.
(18) Chia, P. Y.; Coleman, K. K.; Tan, Y. K.; Ong, S. W. X.; Gum, M.; Lau, S. K.; Lim, X. F.; Lim, A. S.; Sutjipto, S.; Lee, P. H.; Son, T. T.; Young, B. E.; Milton, D. K.; Gray, G. C.; Schuster, S.; Barkham, T.; De, P. P.; Vasoo, S.; Chan, M.; Ang, B. S. P.; Tan, B. H.; Leo, Y.-S.; Ng, O.-T.; Wong, M. S. Y.; Marimuthu, K. Detection of Air and Surface Contamination by SARS-CoV-2 in Hospital Rooms of Infected Patients. *Nat. Commun.* **2020**, *11*, 2800.
(19) Guo, Z.-D.; Wang, Z.-Y.; Zhang, S.-F.; Li, X.; Li, L.; Li, C.; Cui, Y.; Fu, R.-B.; Dong, Y.-Z.; Chi, X.-Y.; Zhang, M.-Y.; Liu, K.; Cao, C.; Liu, B.; Zhang, K.; Gao, Y.-W.; Lu, B.; Chen, W. Aerosol and Surface Distribution of Severe Acute Respiratory Syndrome Coronavirus 2 in Hospital Wards, Wuhan, China, 2020. *Emerg. Infect. Dis.* **2020**, *26*, 1586−1591.
(20) Accarino, G.; Lorenzetti, S.; Aloisio, G. Assessing Correlations between Short-Term Exposure to Atmospheric Pollutants and COVID-19 Spread in All Italian Territorial Areas. *Environ. Pollut.* **2021**, *268*, 115714.
(21) Adams, M. D. Air Pollution in Ontario, Canada during the COVID-19 State of Emergency. *Sci. Total Environ.* **2020**, *742*, 140516.
(22) Andree, B. P. J. *Incidence of COVID-19 and Connections with Air Pollution Exposure: Evidence from the Netherlands*; Preprint; Epidemiology, 2020.
(23) Carleton, T.; Cornetet, J.; Huybers, P.; Meng, K. C.; Proctor, J. Global Evidence for Ultraviolet Radiation Decreasing COVID-19 Growth Rates. *Proc. Nord. Aroma Symp.* **2021**, *118*, No. e2012370118.
(24) Kulkarni, H.; Khandait, H.; Narlawar, U. W.; Rathod, P.; Mamtani, M. Independent Association of Meteorological Characteristics with Initial Spread of Covid-19 in India. *Sci. Total Environ.* **2021**, *764*, 142801.
(25) Ma, Y.; Zhao, Y.; Liu, J.; He, X.; Wang, B.; Fu, S.; Yan, J.; Niu, J.; Zhou, J.; Luo, B. Effects of Temperature Variation and Humidity on the Death of COVID-19 in Wuhan, China. *Sci. Total Environ.* **2020**, *724*, 138226.
(26) Rahman, S.; Azad, A. K.; Hasanuzzaman; Salam, R.; Islam, A. R. T.; Rahman, Md. M.; Hoque, M. M. How Air Quality and COVID-19 Transmission Change under Different Lockdown Scenarios? A Case from Dhaka City, Bangladesh. *Sci. Total Environ.* **2021**, *762*, 143161.
(27) Zhang, X.; Tang, M.; Guo, F.; Wei, F.; Yu, Z.; Gao, K.; Jin, M.; Wang, J.; Chen, K. Associations between Air Pollution and COVID-19 Epidemic during Quarantine Period in China. *Environ. Pollut.* **2021**, *268*, 115897.
(28) Bashir, M. F.; Ma, B. J.; Bilal; Komal, B.; Bashir, M. A.; Farooq, T. H.; Iqbal, N.; Bashir, M. Correlation between Environmental Pollution Indicators and COVID-19 Pandemic: A Brief Study in Californian Context. *Environ. Res.* **2020**, *187*, 109652.
(29) Coccia, M. The Effects of Atmospheric Stability with Low Wind Speed and of Air Pollution on the Accelerated Transmission Dynamics of COVID-19. *Int. J. Environ. Stud.* **2021**, *78*, 1−27.
(30) Coccia, M. Effects of the Spread of COVID-19 on Public Health of Polluted Cities: Results of the First Wave for Explaining the Dejà vu in the Second Wave of COVID-19 Pandemic and Epidemics of Future Vital Agents. *Environ. Sci. Pollut. Res.* **2021**, *28*, 19147−19154.
(31) Haque, S. E.; Rahman, M. Association between Temperature, Humidity, and COVID-19 Outbreaks in Bangladesh. *Environ. Sci. Pol.* **2020**, *114*, 253−255.
(32) Rosario, D. K. A.; Mutz, Y. S.; Bernardes, P. C.; Conte-Junior, C. A. Relationship between COVID-19 and Weather: Case Study in a Tropical Country. *Int. J. Hyg Environ. Health* **2020**, *229*, 113587.
(33) Sarkodie, S. A.; Owusu, P. A. Impact of Meteorological Factors on COVID-19 Pandemic: Evidence from Top 20 Countries with Confirmed Cases. *Environ. Res.* **2020**, *191*, 110101.
(34) Coccia, M. How Do Low Wind Speeds and High Levels of Air Pollution Support the Spread of COVID-19? *Atmos. Pollut. Res.* **2021**, *12*, 437−445.
(35) Coccia, M. Factors Determining the Diffusion of COVID-19 and Suggested Strategy to Prevent Future Accelerated Viral Infectivity Similar to COVID. *Sci. Total Environ.* **2020**, *729*, 138474.
(36) Tian, X.; An, C.; Chen, Z.; Tian, Z. Assessing the Impact of COVID-19 Pandemic on Urban Transportation and Air Quality in Canada. *Sci. Total Environ.* **2021**, *765*, 144270.
(37) Kutter, J. S.; de Meulder, D.; Bestebroer, T. M.; Lexmond, P.; Mulders, A.; Richard, M.; Fouchier, R. A. M.; Herfst, S. SARS-CoV and SARS-CoV-2 Are Transmitted through the Air between Ferrets over More than One Meter Distance. *Nat. Commun.* **2021**, *12*, 1653.
(38) Srivastava, A. COVID-19 and Air Pollution and Meteorology-an Intricate Relationship: A Review. *Chemosphere* **2021**, *263*, 128297.
(39) Islam, N.; Bukhari, Q.; Jameel, Y.; Shabnam, S.; Erzurumluoglu, A. M.; Siddique, M. A.; Massaro, J. M.; D'Agostino, R. B. COVID-19 and Climatic Factors: A Global Analysis. *Environ. Res.* **2021**, *193*, 110355.
(40) Qu, G.; Li, X.; Hu, L.; Jiang, G. An Imperative Need for Research on the Role of Environmental Factors in Transmission of Novel Coronavirus (COVID-19). *Environ. Sci. Technol.* **2020**, *54*, 3730−3732.
(41) Sunyer, J.; Dadvand, P.; Foraster, M.; Gilliland, F.; Nawrot, T. Environment and the COVID-19 Pandemic. *Environ. Res.* **2021**, *195*, 110819.
(42) Holland, P. W. Statistics and Causal Inference. *J. Am. Stat. Assoc.* **1986**, *81*, 945−960.
(43) Damette, O.; Goutte, S. *Weather Pollution and Covid-19 Spread: A Time Series and Wavelet Reassessment*; Olivier Damette: May, 2020; *27*, pp 1−22.
(44) Delnevo, G.; Mirri, S.; Roccetti, M. Particulate Matter and COVID-19 Disease Diffusion in Emilia-Romagna (Italy). Already a Cold Case? *Computation* **2020**, *8*, 59.
(45) Mele, M.; Magazzino, C. Pollution, Economic Growth, and COVID-19 Deaths in India: A Machine Learning Evidence. *Environ. Sci. Pollut. Res.* **2020**, *28*, 2669−2677.
(46) Magazzino, C.; Mele, M.; Schneider, N. The Relationship between Air Pollution and COVID-19-Related Deaths: An Application to Three French Cities. *Appl. Energy* **2020**, *279*, 115835.
(47) Mele, M.; Magazzino, C.; Schneider, N.; Strezov, V. NO2 Levels as a Contributing Factor to COVID-19 Deaths: The First Empirical Estimate of Threshold Values. *Environ. Res.* **2021**, *194*, 110663.
(48) Pearl, J. *Causality: Models, Reasoning, and Inference*; Cambridge University Press: Cambridge, U.K., New York, 2000.
(49) Bates, S.; Sesia, M.; Sabatti, C.; Candès, E. Causal Inference in Genetic Trio Studies. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 24117−24126.
(50) Varian, H. R. Causal Inference in Economics and Marketing. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 7310−7315.
(51) Coccia, M. How (Un)Sustainable Environments Are Related to the Diffusion of COVID-19: The Relation between Coronavirus Disease 2019, Air Pollution, Wind Resource and Energy. *Sustainability* **2020**, *12*, 9709.
(52) Imbens, G. W.; Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*; Cambridge University Press: New York, USA, 2015.
(53) Sharma, A.; Kiciman, E. DoWhy: An End-to-End Library for Causal Inference. **2020**, arXiv:2011.04216 [cs, econ, stat].
(54) Rubin, D. B. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *J. Educ. Psychol.* **1974**, *66*, 688−701.
(55) Forster, P. M.; Forster, H. I.; Evans, M. J.; Gidden, M. J.; Jones, C. D.; Keller, C. A.; Lamboll, R. D.; Quéré, C. L.; Rogelj, J.; Rosen, D.;

Schleussner, C.-F.; Richardson, T. B.; Smith, C. J.; Turnock, S. T. Current and Future Global Climate Impacts Resulting from COVID-19. *Nat. Clim. Chang.* **2020**, *10*, 913−919.
(56) Liu, J.-Y.; Woodward, R. T.; Zhang, Y.-J. Has Carbon Emissions Trading Reduced PM2.5 in China? *Environ. Sci. Technol.* **2021**, *55*, 6631−6643.
(57) Glymour, C.; Zhang, K.; Spirtes, P. Review of Causal Discovery Methods Based on Graphical Models. *Front. Genet.* **2019**, *10*, 524.
(58) Prosperi, M.; Guo, Y.; Sperrin, M.; Koopman, J. S.; Min, J. S.; He, X.; Rich, S.; Wang, M.; Buchan, I. E.; Bian, J. Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare. *Nat. Mach. Intell.* **2020**, *2*, 369−375.
(59) Butcher, B.; Huang, V. S.; Robinson, C.; Reffin, J.; Sgaier, S. K.; Charles, G.; Quadrianto, N. Causal Datasheet for Datasets: An Evaluation Guide for Real-World Data Analysis and Data Collection Design Using Bayesian Networks. *Front. Artif. Intell.* **2021**, *4*, 18.
(60) Ye, X.; Chen, B.; Jing, L.; Zhang, B.; Liu, Y. Multi-Agent Hybrid Particle Swarm Optimization (MAHPSO) for Wastewater Treatment Network Planning. *J. Environ. Manage.* **2019**, *234*, 525−536.
(61) Huang, Q.; Cheng, S. Y.; Li, Y. P.; Li, J. B.; Chen, D. S.; Wang, H. Y. An Integrated MM5-CAMx Modeling Approach for Assessing PM10 Contribution from Different Sources in Beijing, China. *J. Environ. Inf.* **2010**, *15* (2), 47−61.
(62) Rothan, H. A.; Byrareddy, S. N. The Epidemiology and Pathogenesis of Coronavirus Disease (COVID-19) Outbreak. *J. Autoimmun.* **2020**, *109*, 102433.
(63) Bao, R.; Zhang, A. Does Lockdown Reduce Air Pollution? Evidence from 44 Cities in Northern China. *Sci. Total Environ.* **2020**, *731*, 139052.
(64) Jing, L.; Chen, B.; Zhang, B.; Ye, X. Modeling Marine Oily Wastewater Treatment by a Probabilistic Agent-Based Approach. *Mar. Pollut. Bull.* **2018**, *127*, 217−224.
(65) Li, P.; Wu, H. J.; Chen, B. RSW-MCFP A Resource-Oriented Solid Waste Management System for a Mixed Rural-Urban Area through Monte Carlo Simulation-Based Fuzzy Programming. *Math. Probl Eng.* **2013**, *2013*, 780354.
(66) Coccia, M. The Relation between Length of Lockdown, Numbers of Infected People and Deaths of Covid-19, and Economic Growth of Countries: Lessons Learned to Cope with Future Pandemics Similar to Covid-19 and to Constrain the Deterioration of Economic System. *Sci. Total Environ.* **2021**, *775*, 145801.
(67) Blyth, C. R. On Simpson's Paradox and the Sure-Thing Principle. *J. Am. Stat. Assoc.* **1972**, *67*, 364−366.
(68) Pearson, K. LIII On Lines and Planes of Closest Fit to Systems of Points in Space. *London Edinburgh Philos. Mag. J. Sci.* **1901**, *2*, 559−572.
(69) Bellman, R. E. *Adaptive Control Processes: A Guided Tour*; Princeton University Press, 2015.
(70) Song, X.; Lye, L. M.; Chen, B.; Zhang, B. Differentiation of Weathered Chemically Dispersed Oil from Weathered Crude Oil. *Environ. Monit. Assess.* **2019**, *191*, 270.
(71) Xin, X.; Huang, G.; An, C.; Feng, R. Interactive Toxicity of Triclosan and Nano-TiO2 to Green Alga Eremosphaera Viridis in Lake Erie: A New Perspective Based on Fourier Transform Infrared Spectromicroscopy and Synchrotron-Based X-Ray Fluorescence Imaging. *Environ. Sci. Technol.* **2019**, *53*, 9884−9894.
(72) Xin, X.; Huang, G.; An, C.; Raina-Fulton, R.; Weger, H. Insights into Long-Term Toxicity of Triclosan to Freshwater Green Algae in Lake Erie. *Environ. Sci. Technol.* **2019**, *53*, 2189−2198.
(73) Steinhaus, H. Sur La Division Des Corps Maťeriels En Parties. *Bull. Acad. Polon. Sci.* **1957**, *IV*, 801−804.
(74) Thorndike, R. L. Who Belongs in the Family? *Psychometrika* **1953**, *18*, 267−276.
(75) Coccia, M. The Impact of First and Second Wave of the COVID-19 Pandemic in Society: Comparative Analysis to Support Control Measures to Cope with Negative Effects of Future Infectious Diseases. *Environ. Res.* **2021**, *197*, 111099.
(76) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(77) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56−67.
(78) Pearl, J.; Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*; Basic Books: New York, 2018.
(79) Fahey, D. W.; Hübler, G.; Parrish, D. D.; Williams, E. J.; Norton, R. B.; Ridley, B. A.; Singh, H. B.; Liu, S. C.; Fehsenfeld, F. C. Reactive Nitrogen Species in the Troposphere: Measurements of NO, NO2, HNO3, Particulate Nitrate, Peroxyacetyl Nitrate (PAN), O3, and Total Reactive Odd Nitrogen (NO y) at Niwot Ridge, Colorado. *J. Geophys. Res.: Atmos.* **1986**, *91*, 9781−9793.
(80) Pearce, J. L.; Beringer, J.; Nicholls, N.; Hyndman, R. J.; Tapper, N. J. Quantifying the Influence of Local Meteorology on Air Quality Using Generalized Additive Models. *Atmos. Environ.* **2011**, *45*, 1328−1336.
(81) Heckman, J. J. Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* **1978**, *46*, 931−959.
(82) Heckman, J. J. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*; NBER, 1976; Vol. *5*, pp 475−492.
(83) Microsoft Research. *EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation*, 2020.
(84) Kalainathan, D.; Goudet, O.; Guyon, I.; Lopez-Paz, D.; Sebag, M. Structural Agnostic Modeling: Adversarial Learning of Causal Graphs. **2020**, arXiv:1803.04929 [stat].
(85) Kalainathan, D.; Goudet, O. Causal Discovery Toolbox: Uncover Causal Relationships in Python. **2019**, arXiv:1903.02278 [stat].
(86) Gao, W.; Tie, X.; Xu, J.; Huang, R.; Mao, X.; Zhou, G.; Chang, L. Long-Term Trend of O3 in a Mega City (Shanghai), China: Characteristics, Causes, and Interactions with Precursors. *Sci. Total Environ.* **2017**, *603*−*604*, 425−433.
(87) van der, A. R. J.; Eskes, H. J.; Boersma, K. F.; van Noije, T. P. C.; Van Roozendael, M.; De Smedt, I.; Peters, D. H. M. U.; Meijer, E. W. Trends, Seasonal Variability and Dominant NOx Source Derived from a Ten Year Record of NO2 Measured from Space. *J. Geophys. Res.: Atmos.* **2008**, *113*(). https://doi.org/10.1029/2007JD009021. DOI: 10.1029/2007jd009021
(88) LIU, H.; ZHANG, M.; HAN, X. A Review of Surface Ozone Source Apportionment in China. *Atmos. Oceanic Sci. Lett.* **2020**, *13*, 470−484.
(89) Xue, L. K.; Wang, T.; Gao, J.; Ding, A. J.; Zhou, X. H.; Blake, D. R.; Wang, X. F.; Saunders, S. M.; Fan, S. J.; Zuo, H. C.; Zhang, Q. Z.; Wang, W. X. Ground-Level Ozone in Four Chinese Cities: Precursors, Regional Transport and Heterogeneous Processes. *Atmos. Chem. Phys.* **2014**, *14*, 13175−13188.
(90) Diao, Y.; Kodera, S.; Anzai, D.; Gomez-Tames, J.; Rashed, E. A.; Hirata, A. Influence of Population Density, Temperature, and Absolute Humidity on Spread and Decay Durations of COVID-19: A Comparative Study of Scenarios in China, England, Germany, and Japan. *One Health* **2021**, *12*, 100203.
(91) Shen, L.; Zhao, T.; Wang, H.; Liu, J.; Bai, Y.; Kong, S.; Zheng, H.; Zhu, Y.; Shu, Z. Importance of Meteorology in Air Pollution Events during the City Lockdown for COVID-19 in Hubei Province, Central China. *Sci. Total Environ.* **2021**, *754*, 142227.
(92) Xu, K.; Cui, K.; Young, L.-H.; Hsieh, Y.-K.; Wang, Y.-F.; Zhang, J.; Wan, S. Impact of the COVID-19 Event on Air Quality in Central China. *Aerosol Air Qual. Res.* **2020**, *20*, 915−929.
(93) Coccia, M. An Index to Quantify Environmental Risk of Exposure to Future Epidemics of the COVID-19 and Similar Viral Agents: Theory and Practice. *Environ. Res.* **2020**, *191*, 110155.
(94) Caldevilla-Domínguez, D.; Barrientos-Báez, A.; Padilla-Castillo, G. Twitter as a Tool for Citizen Education and Sustainable Cities after COVID-19. *Sustainability* **2021**, *13*, 3514.
(95) Doyle, A.; Hynes, W.; Purcell, S. M. Building Resilient, Smart Communities in a Post-COVID Era: Insights From Ireland. *Int. J. E Plann. Res.* **2021**, *10*, 18−26.
(96) Zhou, C.; Yang, G.; Ma, S.; Liu, Y.; Zhao, Z. The Impact of the COVID-19 Pandemic on Waste-to-Energy and Waste-to-Material Industry in China. *Renew. Sustain. Energy Rev.* **2021**, *139*, 110693.