Supporting Information:

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

Qiao Kang[1]‡, Xing Song[1]‡, Xiaying Xin[1]‡, Bing Chen[1]*, Yuanzhu Chen[2], Xudong Ye[1], Baiyu Zhang[1]

1 Northern Region Persistent Organic Pollution Control (NRPOP) Laboratory, Faculty of Engineering and Applied Science, Memorial University of Newfoundland,
St. John’s, NL, Canada, A1B 3X5

2 School of Computing, Queen's University, Kingston, ON, Canada, K7L 2N8

*Corresponding author. E-mail address: Bing Chen, bchen@mun.ca

‡These authors contributed equally to this work.

This SI file contains:

25 Pages

5 Text Sections

5 Tables

5 Figures

List of Supporting Information:

Text S1. Model Comparison and Selection

A suitable machine learning algorithm was needed to be selected for the feature selection task in the framework. The algorithm should (1) excel in solving complex high-volume continuous data regression problems, (2) be capable of tackling the over-fitting problem, and (3) have a highly interpretable structure for research purposes. At the beginning of the study, we considered many machine learning algorithms, including Supporting Vector Machine (SVM), Random Forest (RF) and LightGBM. However, due to the three demands aforementioned, not all algorithms were competent for the feature selection task in the proposed framework. For instance, SVM's low interpretability is unfavourable for our task, especially when the kernel is non-linear; Random Forest is not a decent regressor and can prone to overfitting problems. From the perspective of the abovementioned point (3), comparing to all the NN-based solutions with limited interpretability, CART-based (Classification And Regression Tree) algorithms seemed to be a more reasonable choice. Since Random Forest was no longer considered due to its lack of resilience to overfitting, we had only two candidates left: XGBoost[1] and LightGBM.[2] The two algorithms are quite similar: both are tree boosting – based algorithms with superb performance. However, we still chose XGBoost as our feature selection method due to it utilized a pre-sorted strategy for node splitting, which could find a slightly more accurate demarcation point in the feature. On the contrary, the histogram-based splitting algorithm used by LightGBM focuses on speed and memory efficiency, which is not a priority in the feature selection task.

We would also like to discuss the two neural network-based time series algorithms, Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM), which specialized in time series analysis. Though they excel in many real-world time series prediction tasks and have been commonly used, we tested the two algorithms and realized that both could hardly fit the proposed framework. First of all, they lacked interpretability during feature selection in the case study. Furthermore, their advantage in real-world extrapolate prediction could not be fully utilized in our framework, which focused on causal inference. Thus, we finally decided not incorporating the two methods in the proposed framework.

Text S2. XGBoost Workflow

A typical workflow of the algorithm is: (1) traverse all features in the dataset and sort the instances by eigenvalues separately; (2) determine the split points for each feature by finding the point with the highest information gain of all possible split points; (3) construct the optimal tree structure by choosing the best split strategy for all the features. Equation (S1) shows the calculation of information gain in XGBoost:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \qquad \text{(S1)}$$

In the equation, G and H are defined as the sum of the first and second derivatives, respectively, of all the samples in a node L or R. λ and γ are constants. The formula could be decomposed as the gain scores on the new left branch, on the new right branch, on the original node, and the additional leaf's regularization. The node split only if the gain is greater than zero.

Text S3. A Brief Introduction to Do-calculus and SCM

To have a deeper insight of causal inference, Structural Causal Model (SCM) and do-calculus, we need to distinguish two conditional distributions that one might want to estimate during data analysis. The two distributions are given below:

Observation p(y|x): The conditional distribution of Y when the observation variable X has the value of x.

Intervene p(y|do(x)): The conditional distribution of Y when the observation variable X is set to x.

Though being similar, the two distributions are totally different. The second distribution involves intervention during the data generation process by purposely setting the value of X to x. Though the first distribution commonly exists in supervised learning problems, only the second distribution can answer causal problems.

The primary goal of do-calculus is to estimate p(y|do(x)) based on observed data outside of a controlled randomized experiment if no access to the measurement can be directly acquired.[3] It is an axiomatic system for replacing probability formulas containing the do operator with ordinary conditional probabilities. Let G be the directed acyclic graph associated with a causal model, and let $P(\dot{})$ stand for the probability distribution induced by the model. For any disjoint subsets of X,Y,Z, and W we have the following rules:[4]

Rule 1: Insertion/deletion of observations

$$P\ (y|\hat{x},z,w) = \ P\ (y|\hat{x},w)\ if\ (Y\ \perp\ Z|X,W\ )_{G_{\overline{X}}} \tag{S2}$$

Rule 2: Action/observation exchange

$$P\ (y|\hat{x},\hat{z},w) = \ P\ (y|\hat{x},z,w)\ if\ (Y\ \perp\ Z|X,W\ )_{G_{\overline{X}\underline{Z}}} \tag{S3}$$

Rule 3: Insertion/deletion of actions

$$P\ (y|\hat{x},\hat{z},w) = \ P\ (y|\hat{x},w)\ if\ (Y\ \perp\ Z|X,W\ )_{G_{\overline{X},\overline{Z(W)}}} \tag{S4}$$

where Z(W) is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$. All the do-calculus-based operations are based on the three rules mentioned above.

Equation (S5) shows the formula to calculate average treatment given a set of identified backdoor variables:

$$ATE = \ \sum_{W} P(Y|T,W)P(W) \tag{S5}$$

Where Y is the outcome, T is the treatment, and W is a set of identified backdoor variables.

Text S4. Framework Benchmark

We deployed SCM on two public datasets: Infant Health and Development Program (IHDP)[5] Dataset and Lanlode Dataset.[6] IHDP dataset (n=747) is from a randomized experiment that began in 1985 targeting infant health, which means that the ground truth of the causal relationship in the dataset is known. The dataset consists of measurements on the child (birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health, index sex, twin status) as well as mother status and behaviours during the pregnancy (consumption status of cigarettes, alcohol and drugs, age, marital status, educational attainment, employment, prenatal care, family residing site). The treatment variable in the dataset is if the infant received both intensive high-quality child care and home visits from a trained provider. LaLonde Dataset (n=445) is another well-known dataset that aims to investigate the effect of an employment training program, National Supported Work Demonstration (NSW), on wage increases (i.e., real income in 1978). Since the applicants were admitted randomly to the program, the ground truth within the dataset is also known as in the IHDP Dataset. The dataset also has the features such as age, years of schooling, indicator variables for race, martial status, high school diploma, real earnings in 1974 and 1975, and whether earnings in 1974 or 1975 being zero. The SCM identified causal relationships in both dataset, which passed three refutation methods (i.e., add random common cause, replace treatment with placebo, remove random subset of data). The estimates causal effects are 3.41 and 1614.16, respectively. A Jupyter notebook with the causal estimation and refutation results can be found in the GitHub repository of the study (https://github.com/kangqiao-ctrl/EnvCausal/tree/main/benchmark/hdpi_lalonde).

Structural Agnostic Model (SAM) was applied on another well-known dataset, Sachs Dataset,[7] which consists of simultaneous measurements of multiple phosphorylated proteins and phospholipid components in thousands of individual primary human immune system cells. The dataset was generated with molecular interventions which perturbed the cells. We applied SAM to the dataset to test capability in recovering the causal network. Two important metrics, Precision and Recall were calculated based on Equation S6 and S7:

$$Precision = \frac{TP}{TP+FP} \tag{S6}$$

$$Recall = \frac{TP}{TP+FN} \tag{S7}$$

Where TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives.

For the Sachs dataset, SAM acquired an Area Under Precision-Recall Curve (AUPR) of 0.311. Since the corresponding baseline is 0.168 for this case, AUPR of such value is considered decent. A Jupyter notebook of the SAM benchmark on Sachs Dataset can be found in the GitHub repository of the study (https://github.com/kangqiao-ctrl/EnvCausal/tree/main/benchmark/sachs).

Text S5. Additional Experiments

During the investigation, some additional machine learning experiments other than the one presented in the main text were conducted during the study. The brief descriptions of the experiments were given below:

Removing the elapsed days feature. We removed the elapsed days in this experiment and trained the XGBoost models on the rest of the data. Feature importance was given in Figure S6. Under the setting, the air temperature became one of the top contributors among all clusters due to its high collinearity with elapsed time. We have reasons to believe that introducing elapsed days can weaken spurious correlations originated from any other highly time-correlated features such as air temperature.

Seven-day moving average. Instead of the original three-day moving average strategy, we applied a seven-day moving average on the time series dataset for machine learning and SCM analysis. No significant changes were observed. The results can be found in the GitHub repository of the study (https://github.com/kangqiao-ctrl/EnvCausal/tree/main/additional_experiments/7-day-moving-average).

Targeting cases per capita. In this experiment, we set the daily new cases per capita as the machine learning regression target instead of the absolute case number. No significant difference was observed in the result. It might because that the cities have already been clustered based on their socio-economic status include population. The results can be found in the GitHub repository of the study (https://github.com/kangqiao-ctrl/EnvCausal/tree/main/additional_experiments/morbidity_target).

Table S1. Features and Data Sources

Feature and Unit (if applicable)	Mean	Standard Deviation	Min	Max	Source
Snapshot Dataset					
Population (Thousand People)	5,624.67	4,029.80	720.96	31,243.20	B
City area (km^2)	11,733.64	9,080.77	1,459.00	82,402.00	Y
Population density (People per km^2)	652.19	694.77	24.31	6,729.49	E
GDP (Billion USD)	66.02	81.53	5.13	552.18	B
Primary sector (Billion USD)	3.54	2.51	0.17	21.87	B
Secondary sector (Billion USD)	25.80	26.94	1.89	151.89	B
Tertiary sector (Billion USD)	36.68	57.05	2.85	427.53	B
Primary sector percentage (%)	8.42	5.09	0.09	23.08	B
Secondary sector percentage (%)	41.33	7.58	16.16	60.00	B
Tertiary sector percentage (%)	50.25	8.08	33.54	83.52	B
GDP per capita (Thousand USD)	10.52	5.29	4.01	29.05	E
Elderly population percentage (%)	19.50	4.50	4.92	32.20	B, Y
Hospital beds per thousand people	6.22	1.22	3.82	9.67	B
Registered medical doctors per thousand people	2.81	0.76	1.32	5.76	B
Registered nurses per thousand people	3.19	1.01	1.27	6.71	B, Y
Travellers from Wuhan (Thousand People)	23.98	85.02	0.00	691.87	Baidu LBS
Wuhan travellers per thousand population	6.01	23.303	0.00	187.25	E
Average degree of activeness	5.36	0.64	2.98	7.08	Baidu LBS

Timeseries Dataset					
PM2.5 (μg/m^3)	46.67	31.34	3.67	349.00	CNEMC
PM10 (μg/m^3)	70.27	38.73	6.33	378.00	NEMC
SO_2 (μg/m^3)	10.33	7.415	1.67	92.00	NEMC
CO (mg/m^3)	0.81	0.35	0.20	4.50	NEMC
NO_2(μg/m^3)	25.26	11.17	2.67	87.00	NEMC
O_3 (μg/m^3)	83.82	22.06	5.00	166.67	NEMC
Relative humidity (%)	71.23	18.20	8.00	100.00	BIN
Atmospheric pressure (hpa)	991.77	50.30	644.33	1,035.33	BIN
Wind speed (m/s)	2.23	1.31	0.10	11.47	BIN
Average air temperature	8.98	6.34	-22.00	27.68	BIN
Degree of activeness	3.59	1.34	0.31	8.81	Baidu LBS
New confirmed cases	6.17	34.26	0.00	1,021.00	CCDC
Morbidity rate	0.02	0.11	0.00	3.21	E

Note: “B” “Y” “E” in the Source columns indicate “Bulletin” “Yearbook” “Engineered Feature” respectively; “CCDC” is the abbreviation of the Chinese Center for Disease Control and Prevention; Percentage of Elderly and Registered Nurses per Thousand People were collected from multiple sources, including each city’s 2019 Statistical Bulletin, the 2018 Statistical Yearbook, or directly acquired from City-level Civil Affairs Bureau, depending on the availability of the data; The air pollution data was provided by the China National Environmental Monitoring Center, while the meteorological data was collected from an Application Programming Interface (API) provided by BINSTD, a data trading company; Travellers from Wuhan and Degree of Activeness in each city are calculated based on Baidu Location-based Service(LBS) data.

Table S2. Cities in different clusters

Cluster	City names
Cluster l (Megacities)	Beijing, Shanghai, Chongqing, Suzhou, Chengdu, Guangzhou, Shenzhen
Cluster 2 (Major Cities)	Shenyang, Dalian, Fuzhou, Xiamen, Nanning, Haikou, Guiyang, Kunming, Lhasa, Lanzhou, Xining, Yinchuan, Ürümqi, Tianjin, Shijiazhuang, Taiyuan, Jinan, Qingdao, Zhengzhou, Hohhot, Baotou, Nanjing, Wuxi, Changzhou, Hangzhou, Ningbo, Wenzhou, Shaoxing, Jiaxing, Jinhua, Hefei, Xi'an, Tongchuan, Nanchang, Changsha, Zhuhai, Foshan, Huizhou, Dongguan, Zhongshan
Cluster 3 (Common Cities)	Chaoyang, Jinzhou, Huludao, Changchun, Harbin, Tangshan, Qinhuangdao, Handan, Xingtai, Baoding, Zhangjiakou, Chengde, Cangzhou, Langfang, Hengshui, Datong, Shuozhou, Xinzhou, Yangquan, Changzhi, Jincheng, Lüliang, Jinzhong, Linfen, Yuncheng, Zibo, Zaozhuang, Weifang, Jining, Tai'an, Rizhao, Linyi, Dezhou, Liaocheng, Binzhou, Heze, Kaifeng, Pingdingshan, Anyang, Hebi, Xinxiang, Jiaozuo, Puyang, Xuchang, Luohe, Nanyang, Shangqiu, Xinyang, Zhoukou, Zhumadian, Luoyang, Sanmenxia, Xuzhou, Nantong, Lianyungang, Huai'an, Yancheng, Yangzhou, Zhenjiang, Taizhou, Suqian, Huzhou, Quzhou, Taizhou, Lishui, Zhoushan, Wuhu, Bengbu, Huainan, Ma'anshan, Huaibei, Tongling, Anqing, Huangshan, Fuyang, Suzhou, Chuzhou, Lu'an, Xuancheng, Chizhou, Bozhou, Xianyang, Baoji, Weinan, Zigong, Luzhou, Deyang, Mianyang, Suining, Neijiang, Leshan, Meishan, Yibin, Ya'an, Ziyang, Nanchong, Guan'gan, Dazhou, Xianning, Xiaogan, Huanggang, Huangshi, Ezhou, Xiangyang, Yichang, Jingmen, Jingzhou, Suizhou, Pingxiang, Xinyu, Yichun, Jiujiang, Zhuzhou, Xiangtan, Yueyang, Changde, Yiyang, Jiangmen, Zhaoqing

Table S3. Average feature values in different city clusters

Feature and Unit (if applicable)	Cluster 1	Cluster 2	Cluster 3
Population (Thousands)	19,019.47	6,266.86	4,620.88
City Area (km^2)	19,653.14	10,440.70	11,702.39
Population Density (People per km^2)	2,386.85	868.45	477.46
GDP (Billion USD)	380.27	97.44	36.97
Primary Sector (Billion USD)	5.82	2.92	3.61
Secondary Sector (Billion USD)	117.50	39.16	15.92
Tertiary Sector (Billion USD)	256.87	55.37	17.44
Primary Sector Percentage (%)	18.25	34.68	10.47
Secondary Sector Percentage (%)	32.57	39.55	42.45
Tertiary Sector Percentage (%)	65.58	57.00	47.08
GDP Per Capita (Thousand Yuan)	147.6	105.19	57.13
Elderly Population Percentage (%)	17.46	16.54	20.62
Hospital Beds per Thousand People	6.37	6.73	6.04
Registered Medical Doctors per Thousand People	3.49	3.57	2.51
Registered Nurses per Thousand People	4.30	4.32	2.74
Wuhan Travellers (Thousand People)	31.69	7.49	29.07
Wuhan Travellers per Thousand Citizens	15.93	10.72	7.93
Average Degree of Activeness	4.87	4.84	5.57

Table S4. Final XGBoost model hyperparameters and R^2

	Cluster 1			Cluster 2			Cluster 3		
	Spreading	Post-peak	Overall	Spreading	Post-peak	Overall	Spreading	Post-peak	Overall
max_depth	4	4	2	4	3	3	4	5	5
min_child_weight	9	8	8	9	8	8	8	6	3
n_estimators	50	50	50	50	50	50	50	50	50
learning_rate	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
R^2	0.965	0.965	0.796	0.891	0.806	0.768	0.877	0.842	0.837

*The ranges of the hyperparamters used in GridSearchCV are given as below: max_depth [2,10]; min_child_weight [2,10]; n_estimators [25,50,75,100,150,200,250,300]; learning_rate [0.01,0.05,0.1,0.2,0.3]

Table S5. Weighted adjacency matrix generated by SAM

Cluster 1 - Overall													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.818	0.004	0.596	0.240	0.035	0.010	0.005	0.250	0.539	0.230	0.312	0.415
PM10	0.260	0.000	0.411	0.008	0.547	0.452	0.335	0.358	0.255	0.026	0.150	0.031	0.011
SO_2	0.082	0.247	0.000	0.295	0.368	0.440	0.488	0.296	0.282	0.118	0.001	0.366	0.010
CO	0.634	0.012	0.547	0.000	0.730	0.316	0.677	0.007	0.466	0.020	0.323	0.300	0.543
NO_2	0.009	0.052	0.600	0.259	0.000	0.053	0.001	0.009	0.684	0.053	0.236	0.165	0.037
O_3	0.135	0.010	0.008	0.377	0.025	0.000	0.393	0.244	0.233	0.519	0.445	0.307	0.143
HMD	0.161	0.340	0.535	0.345	0.003	0.318	0.000	0.184	0.409	0.392	0.432	0.287	0.260
PRES	0.151	0.312	0.513	0.326	0.686	0.786	0.174	0.000	0.319	0.544	0.616	0.177	0.002
WSPD	0.004	0.002	0.313	0.178	0.383	0.001	0.267	0.010	0.000	0.001	0.012	0.517	0.001
TEMP	0.465	0.475	0.551	0.258	0.278	0.338	0.251	0.425	0.487	0.000	0.597	0.265	0.455
ACTV	0.014	0.016	0.046	0.015	0.455	0.024	0.002	0.003	0.600	0.083	0.000	0.330	0.515
CASES	0.031	0.144	0.003	0.143	0.268	0.002	0.234	0.167	0.326	0.073	0.014	0.000	0.127
DAYS	0.196	0.502	0.125	0.371	0.855	0.132	0.014	0.004	0.252	0.348	0.975	0.888	0.000
Cluster 1 – Spreading phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.696	0.043	0.553	0.036	0.057	0.004	0.006	0.158	0.110	0.045	0.045	0.043
PM10	0.051	0.000	0.179	0.045	0.037	0.500	0.021	0.028	0.095	0.074	0.005	0.002	0.012
SO_2	0.001	0.200	0.000	0.147	0.008	0.010	0.709	0.237	0.129	0.169	0.305	0.471	0.002
CO	0.117	0.001	0.415	0.000	0.372	0.002	0.007	0.001	0.092	0.015	0.004	0.006	0.010
NO_2	0.075	0.008	0.004	0.221	0.000	0.005	0.005	0.023	0.572	0.002	0.445	0.004	0.003
O_3	0.014	0.039	0.001	0.029	0.076	0.000	0.698	0.158	0.101	0.008	0.269	0.001	0.002
HMD	0.001	0.017	0.048	0.003	0.035	0.226	0.000	0.004	0.232	0.030	0.005	0.009	0.011
PRES	0.224	0.171	0.557	0.010	0.578	0.584	0.021	0.000	0.769	0.404	0.021	0.005	0.001
WSPD	0.002	0.013	0.218	0.008	0.235	0.005	0.074	0.002	0.000	0.012	0.003	0.031	0.004
TEMP	0.159	0.002	0.617	0.010	0.002	0.473	0.382	0.471	0.137	0.000	0.608	0.507	0.031
ACTV	0.017	0.146	0.027	0.043	0.235	0.409	0.628	0.010	0.130	0.003	0.000	0.014	0.031

CASES	0.004	0.006	0.019	0.053	0.093	0.003	0.691	0.065	0.381	0.095	0.062	0.000	0.174
DAYS	0.002	0.042	0.019	0.280	0.852	0.476	0.631	0.011	0.079	0.588	0.747	0.700	0.000
Cluster 1 – Post-peak phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.443	0.006	0.436	0.028	0.005	0.023	0.010	0.009	0.128	0.291	0.113	0.009
PM10	0.476	0.000	0.901	0.005	0.025	0.152	0.300	0.838	0.090	0.028	0.003	0.012	0.615
SO_2	0.138	0.016	0.000	0.056	0.034	0.347	0.267	0.070	0.632	0.016	0.171	0.009	0.052
CO	0.507	0.008	0.076	0.000	0.509	0.097	0.735	0.042	0.486	0.062	0.002	0.267	0.006
NO_2	0.005	0.004	0.503	0.217	0.000	0.059	0.004	0.508	0.881	0.031	0.477	0.394	0.051
O_3	0.003	0.006	0.167	0.469	0.300	0.000	0.517	0.112	0.001	0.214	0.029	0.571	0.045
HMD	0.037	0.127	0.250	0.080	0.009	0.138	0.000	0.238	0.629	0.022	0.001	0.318	0.012
PRES	0.245	0.020	0.525	0.148	0.280	0.426	0.222	0.000	0.444	0.379	0.896	0.730	0.002
WSPD	0.024	0.001	0.180	0.032	0.014	0.001	0.223	0.015	0.000	0.005	0.005	0.059	0.001
TEMP	0.585	0.125	0.561	0.621	0.075	0.446	0.741	0.270	0.314	0.000	0.579	0.033	0.071
ACTV	0.001	0.005	0.056	0.011	0.047	0.032	0.011	0.004	0.054	0.017	0.000	0.139	0.147
CASES	0.027	0.006	0.013	0.010	0.004	0.031	0.006	0.005	0.009	0.010	0.129	0.000	0.037
DAYS	0.008	0.176	0.171	0.507	0.517	0.148	0.438	0.029	0.895	0.379	0.681	0.735	0.000
Cluster 2 - Overall													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.718	0.199	0.785	0.263	0.026	0.477	0.402	0.471	0.192	0.097	0.155	0.178
PM10	0.445	0.000	0.152	0.402	0.320	0.334	0.459	0.207	0.008	0.005	0.003	0.008	0.332
SO_2	0.280	0.618	0.000	0.477	0.564	0.007	0.428	0.431	0.323	0.479	0.258	0.034	0.263
CO	0.310	0.277	0.426	0.000	0.355	0.393	0.224	0.303	0.084	0.357	0.001	0.273	0.455
NO_2	0.087	0.263	0.319	0.420	0.000	0.510	0.049	0.152	0.249	0.443	0.379	0.205	0.404
O_3	0.003	0.151	0.310	0.152	0.163	0.000	0.298	0.183	0.066	0.310	0.279	0.149	0.162
HMD	0.239	0.265	0.307	0.319	0.288	0.486	0.000	0.448	0.288	0.352	0.003	0.005	0.480
PRES	0.319	0.132	0.188	0.299	0.398	0.268	0.352	0.000	0.287	0.208	0.501	0.415	0.539
WSPD	0.001	0.167	0.270	0.538	0.162	0.200	0.294	0.446	0.000	0.017	0.256	0.466	0.316
TEMP	0.276	0.167	0.207	0.415	0.075	0.397	0.409	0.306	0.167	0.000	0.370	0.340	0.436
ACTV	0.293	0.397	0.179	0.026	0.399	0.416	0.016	0.092	0.241	0.309	0.000	0.546	0.405

CASES	0.308	0.450	0.001	0.352	0.029	0.144	0.201	0.390	0.152	0.087	0.127	0.000	0.596
DAYS	0.361	0.472	0.071	0.457	0.481	0.369	0.006	0.007	0.071	0.271	0.736	0.166	0.000
Cluster 2 – Spreading phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.812	0.008	0.514	0.354	0.017	0.573	0.279	0.254	0.523	0.005	0.068	0.221
PM10	0.082	0.000	0.250	0.409	0.409	0.010	0.051	0.018	0.032	0.224	0.022	0.088	0.004
SO_2	0.149	0.324	0.000	0.233	0.077	0.126	0.116	0.261	0.340	0.290	0.297	0.014	0.006
CO	0.133	0.007	0.457	0.000	0.359	0.381	0.048	0.040	0.314	0.203	0.030	0.018	0.403
NO_2	0.011	0.088	0.132	0.481	0.000	0.109	0.005	0.082	0.235	0.064	0.278	0.351	0.307
O_3	0.002	0.001	0.113	0.196	0.008	0.000	0.183	0.126	0.403	0.247	0.012	0.007	0.079
HMD	0.078	0.137	0.374	0.005	0.001	0.692	0.000	0.450	0.322	0.118	0.089	0.003	0.464
PRES	0.003	0.313	0.361	0.244	0.193	0.117	0.127	0.000	0.085	0.165	0.009	0.238	0.024
WSPD	0.068	0.002	0.124	0.112	0.005	0.186	0.002	0.184	0.000	0.003	0.058	0.294	0.012
TEMP	0.003	0.011	0.442	0.307	0.177	0.243	0.308	0.168	0.486	0.000	0.259	0.116	0.260
ACTV	0.009	0.121	0.254	0.002	0.176	0.131	0.002	0.148	0.256	0.336	0.000	0.331	0.386
CASES	0.207	0.161	0.004	0.008	0.029	0.190	0.421	0.405	0.423	0.087	0.004	0.000	0.545
DAYS	0.003	0.004	0.001	0.089	0.056	0.450	0.086	0.020	0.010	0.274	0.483	0.247	0.000
Cluster 2 – Post-peak phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.514	0.194	0.411	0.408	0.003	0.340	0.208	0.062	0.260	0.195	0.001	0.002
PM10	0.360	0.000	0.112	0.022	0.236	0.166	0.187	0.071	0.063	0.025	0.034	0.100	0.231
SO_2	0.069	0.456	0.000	0.158	0.239	0.002	0.272	0.173	0.173	0.431	0.225	0.351	0.227
CO	0.279	0.306	0.333	0.000	0.437	0.147	0.003	0.315	0.332	0.248	0.190	0.214	0.254
NO_2	0.057	0.313	0.384	0.250	0.000	0.007	0.062	0.250	0.257	0.107	0.212	0.235	0.195
O_3	0.135	0.079	0.046	0.228	0.065	0.000	0.189	0.279	0.171	0.245	0.010	0.047	0.247
HMD	0.232	0.271	0.254	0.303	0.212	0.426	0.000	0.411	0.259	0.306	0.078	0.147	0.003
PRES	0.276	0.105	0.273	0.244	0.185	0.196	0.278	0.000	0.317	0.169	0.467	0.422	0.092
WSPD	0.002	0.108	0.271	0.111	0.001	0.002	0.020	0.303	0.000	0.037	0.284	0.585	0.035
TEMP	0.128	0.235	0.229	0.363	0.005	0.555	0.299	0.447	0.126	0.000	0.088	0.159	0.380
ACTV	0.264	0.198	0.368	0.120	0.190	0.148	0.105	0.096	0.167	0.238	0.000	0.120	0.241

CASES	0.201	0.362	0.004	0.004	0.168	0.022	0.054	0.127	0.024	0.070	0.094	0.000	0.467
DAYS	0.002	0.070	0.008	0.107	0.376	0.220	0.115	0.023	0.023	0.250	0.435	0.214	0.000

Cluster 3 - Overall

	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.539	0.017	0.426	0.193	0.199	0.192	0.204	0.228	0.211	0.002	0.226	0.136
PM10	0.423	0.000	0.195	0.029	0.138	0.194	0.268	0.060	0.204	0.222	0.150	0.065	0.005
SO_2	0.229	0.259	0.000	0.231	0.249	0.092	0.312	0.243	0.190	0.208	0.209	0.006	0.024
CO	0.145	0.253	0.129	0.000	0.221	0.210	0.249	0.308	0.089	0.181	0.207	0.119	0.207
NO_2	0.091	0.212	0.215	0.212	0.000	0.209	0.100	0.126	0.207	0.272	0.144	0.069	0.179
O_3	0.206	0.238	0.295	0.140	0.171	0.000	0.206	0.189	0.227	0.227	0.277	0.121	0.131
HMD	0.202	0.205	0.238	0.135	0.243	0.361	0.000	0.190	0.238	0.210	0.207	0.148	0.210
PRES	0.005	0.290	0.106	0.202	0.150	0.227	0.185	0.000	0.181	0.252	0.111	0.165	0.092
WSPD	0.129	0.186	0.161	0.269	0.132	0.314	0.163	0.138	0.000	0.204	0.221	0.123	0.218
TEMP	0.206	0.228	0.223	0.166	0.160	0.221	0.176	0.211	0.182	0.000	0.166	0.071	0.200
ACTV	0.063	0.140	0.143	0.178	0.176	0.239	0.140	0.119	0.186	0.165	0.000	0.106	0.244
CASES	0.011	0.000	0.008	0.217	0.364	0.038	0.173	0.329	0.053	0.217	0.634	0.000	0.278
DAYS	0.197	0.204	0.317	0.179	0.330	0.226	0.012	0.111	0.131	0.193	0.157	0.099	0.000

Cluster 3 – Spreading phase

	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.389	0.050	0.686	0.341	0.479	0.662	0.278	0.021	0.223	0.173	0.090	0.143
PM10	0.833	0.000	0.475	0.755	0.517	0.080	0.405	0.002	0.082	0.221	0.251	0.034	0.154
SO_2	0.277	0.380	0.000	0.334	0.734	0.057	0.452	0.264	0.130	0.291	0.209	0.001	0.010
CO	0.037	0.243	0.207	0.000	0.208	0.058	0.592	0.041	0.270	0.162	0.001	0.161	0.225
NO_2	0.058	0.214	0.436	0.391	0.000	0.288	0.253	0.002	0.034	0.008	0.048	0.016	0.022
O_3	0.132	0.004	0.145	0.334	0.143	0.000	0.282	0.158	0.284	0.167	0.284	0.126	0.189
HMD	0.154	0.000	0.321	0.098	0.196	0.280	0.000	0.202	0.254	0.256	0.234	0.155	0.151
PRES	0.069	0.005	0.264	0.329	0.278	0.372	0.464	0.000	0.353	0.127	0.304	0.002	0.108
WSPD	0.328	0.297	0.423	0.069	0.188	0.276	0.437	0.211	0.000	0.004	0.083	0.083	0.206
TEMP	0.076	0.350	0.607	0.077	0.234	0.354	0.418	0.512	0.172	0.000	0.321	0.125	0.196
ACTV	0.172	0.258	0.140	0.026	0.234	0.413	0.280	0.004	0.145	0.131	0.000	0.052	0.150

CASES	0.000	0.001	0.005	0.272	0.504	0.092	0.144	0.484	0.130	0.582	0.877	0.000	0.957
DAYS	0.016	0.063	0.217	0.331	0.027	0.412	0.422	0.100	0.142	0.310	0.711	0.073	0.000
Cluster 3 – Post-peak phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.389	0.050	0.686	0.341	0.479	0.662	0.278	0.021	0.223	0.173	0.090	0.143
PM10	0.833	0.000	0.475	0.755	0.517	0.080	0.405	0.002	0.082	0.221	0.251	0.034	0.154
SO_2	0.277	0.380	0.000	0.334	0.734	0.057	0.452	0.264	0.130	0.291	0.209	0.001	0.010
CO	0.037	0.243	0.207	0.000	0.208	0.058	0.592	0.041	0.270	0.162	0.001	0.161	0.225
NO_2	0.058	0.214	0.436	0.391	0.000	0.288	0.253	0.002	0.034	0.008	0.048	0.016	0.022
O_3	0.132	0.004	0.145	0.334	0.143	0.000	0.282	0.158	0.284	0.167	0.284	0.126	0.189
HMD	0.154	0.000	0.321	0.098	0.196	0.280	0.000	0.202	0.254	0.256	0.234	0.155	0.151
PRES	0.069	0.005	0.264	0.329	0.278	0.372	0.464	0.000	0.353	0.127	0.304	0.002	0.108
WSPD	0.328	0.297	0.423	0.069	0.188	0.276	0.437	0.211	0.000	0.004	0.083	0.083	0.206
TEMP	0.076	0.350	0.607	0.077	0.234	0.354	0.418	0.512	0.172	0.000	0.321	0.125	0.196
ACTV	0.172	0.258	0.140	0.026	0.234	0.413	0.280	0.004	0.145	0.131	0.000	0.052	0.150
CASES	0.000	0.001	0.005	0.272	0.504	0.092	0.144	0.484	0.130	0.582	0.877	0.000	0.957
DAYS	0.016	0.063	0.217	0.331	0.027	0.412	0.422	0.100	0.142	0.310	0.711	0.073	0.000

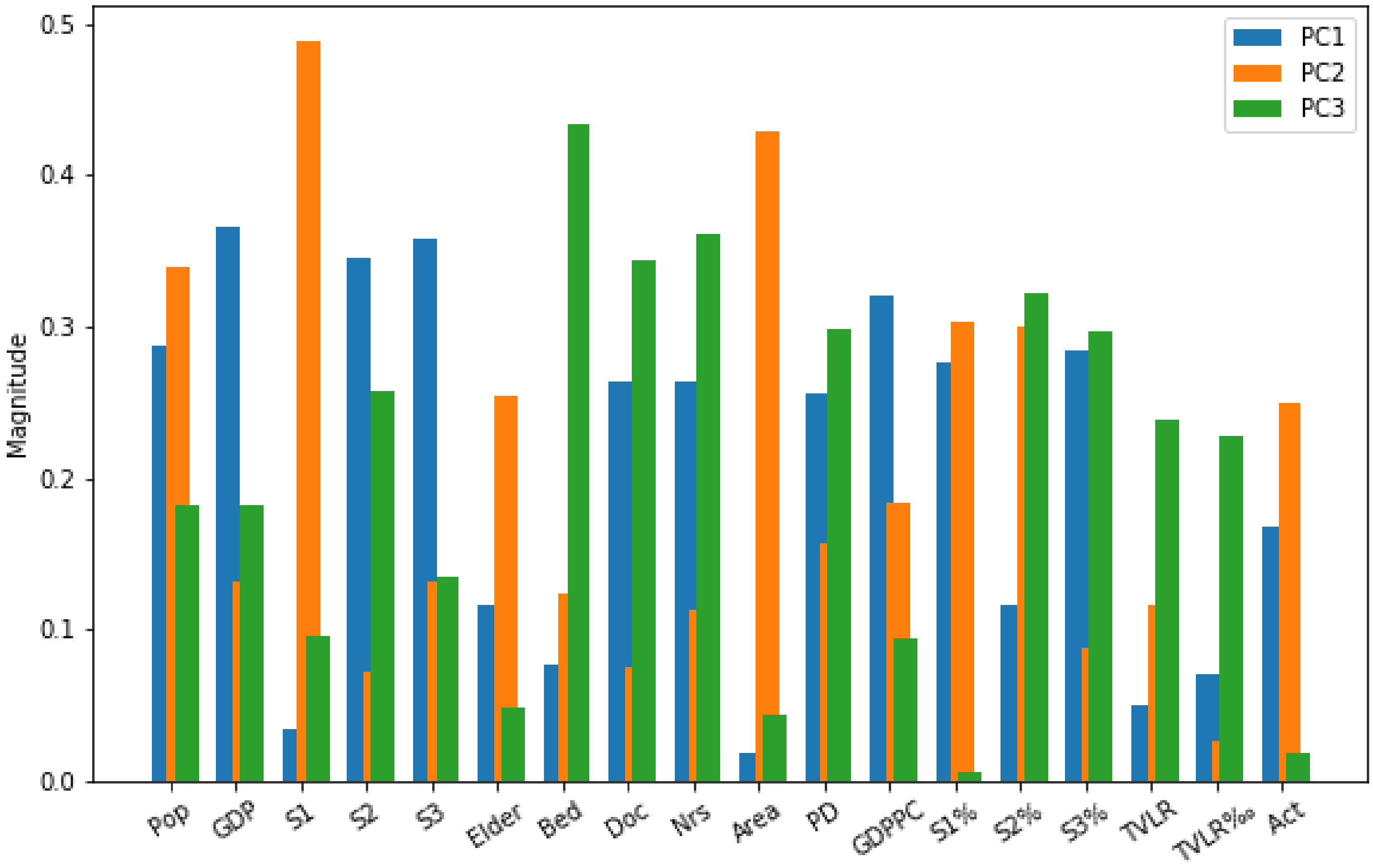

Figure S1. Contribution of each feature to different principal components. Pop: population; S1/S2/S3: primary, secondary, tertiary sector of GDP; Elder: elderly population percentage (over 60-year-old); Bed/Doc/Nrs: hospital beds/ registered medical doctors/ registered nurses per thousand people; TVLR: travellers from Wuhan; TVLR‰: Wuhan travellers per thousand; Act: the average degree of activeness before the 2020 Spring Festival. Explained variance by PC1-3: 31.1%, 18.2%, 14.0%.

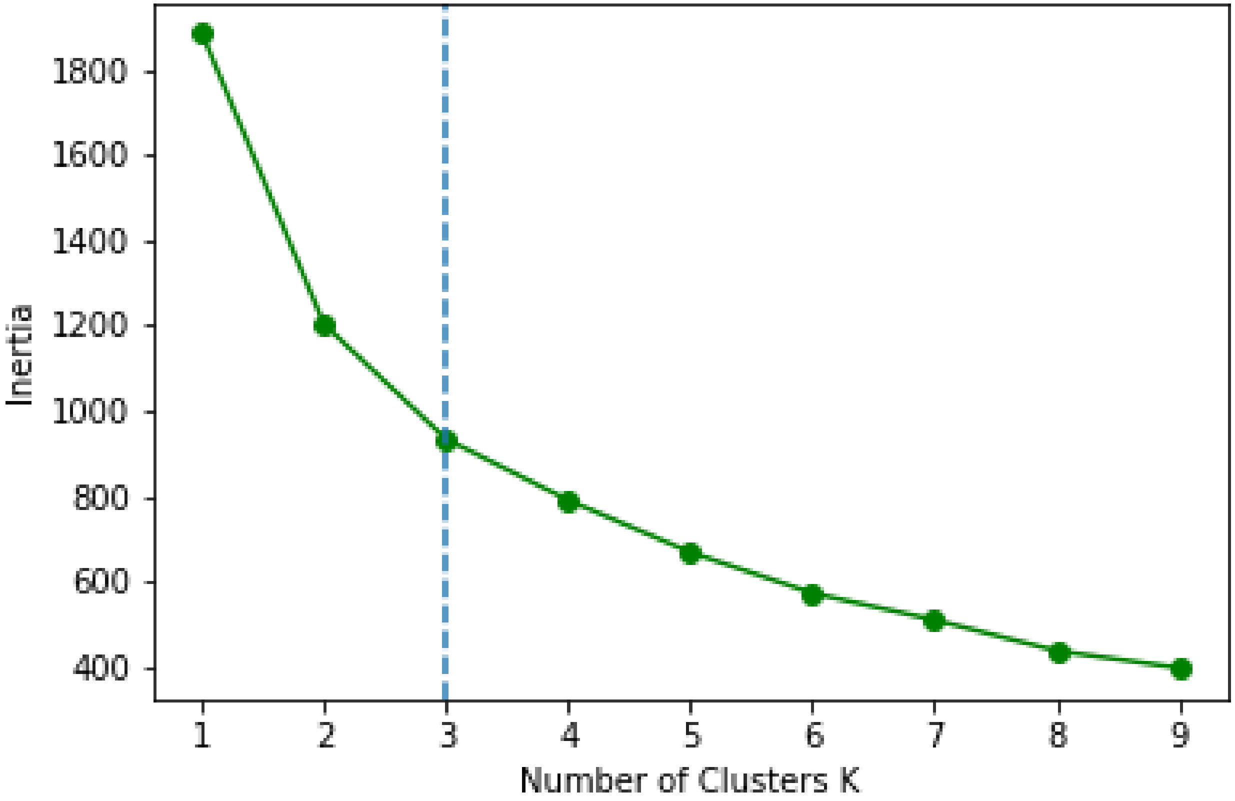

Figure S2. Explained variance and number of clusters. The "elbow" is indicated by the blue dashed line. The number of clusters chosen should therefore be 3.

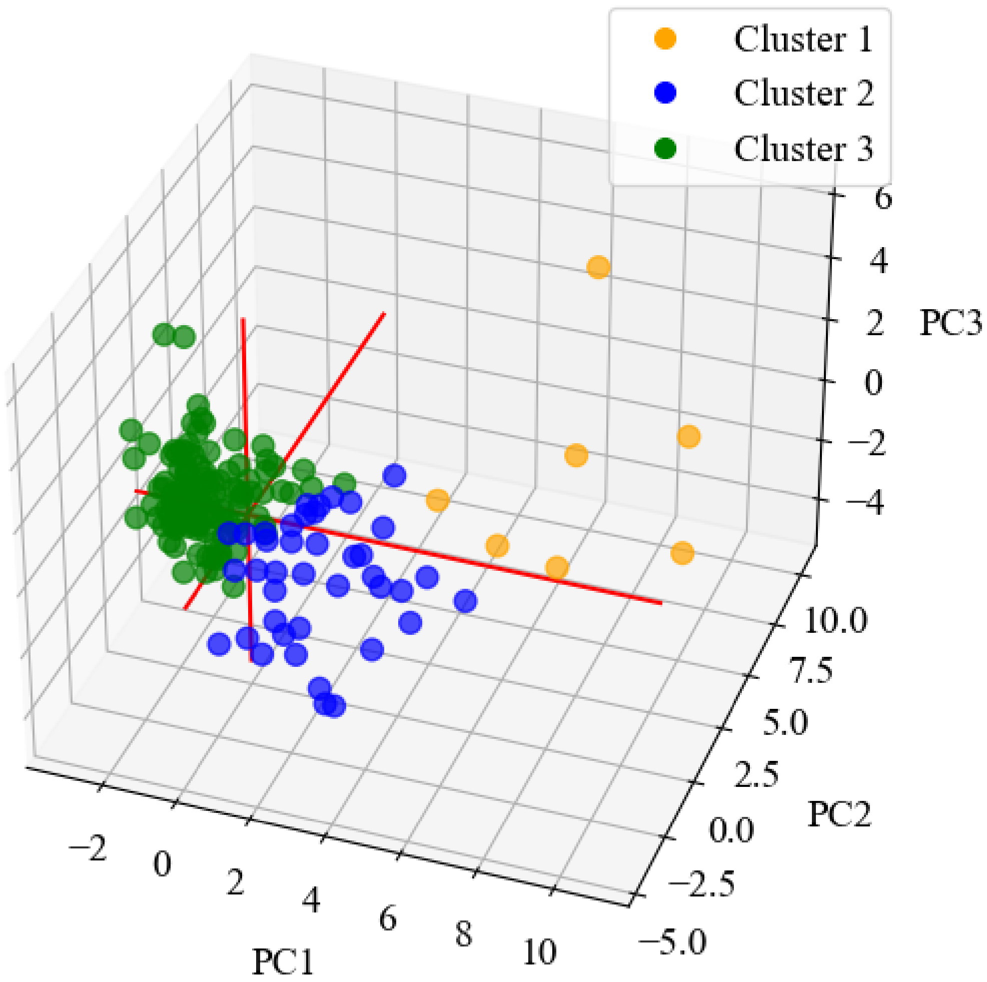

Figure S3. Clustered cities in the principal component space.

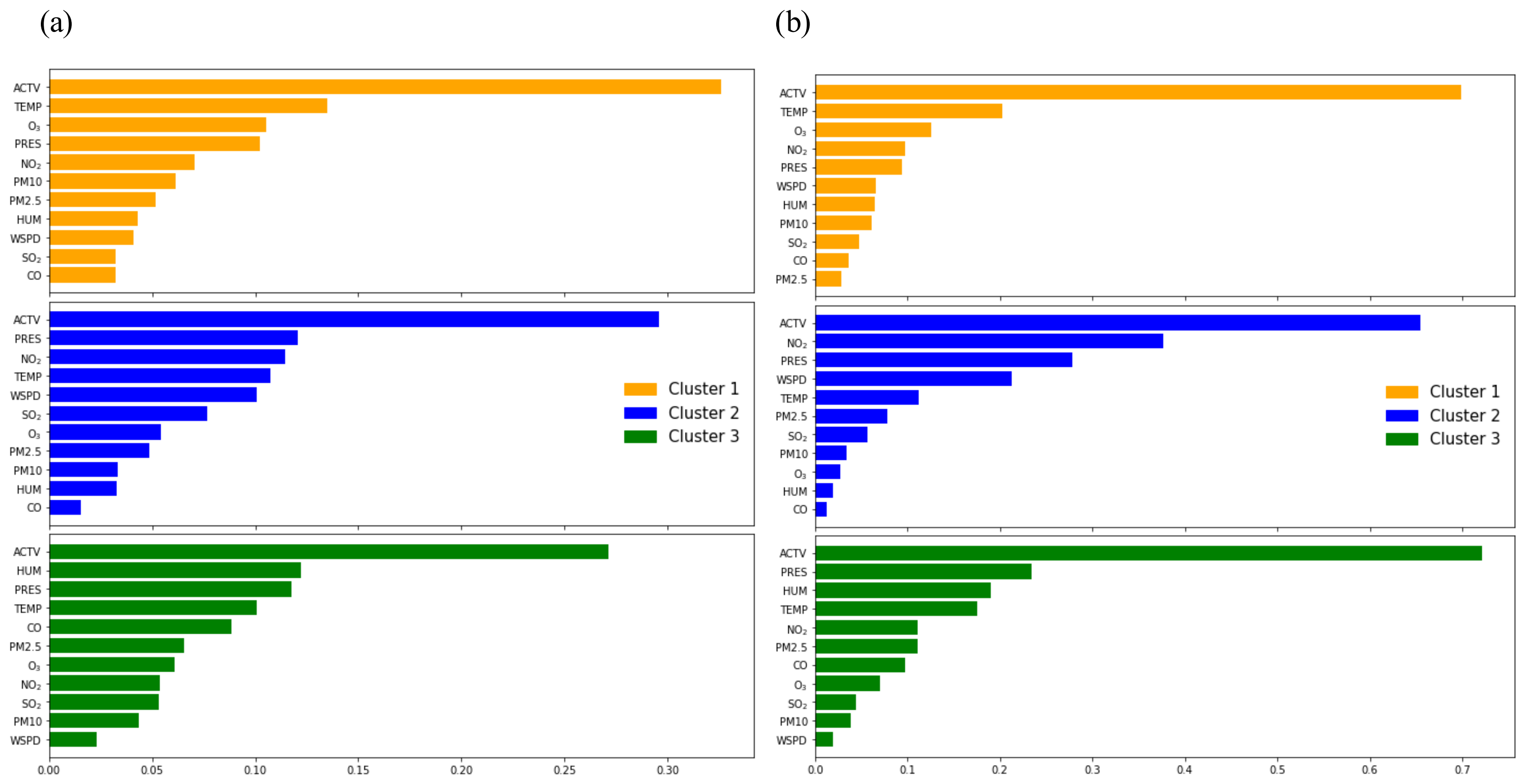

Figure S4. Feature importance and ranking in different clusters with no “elapsed days” feature (*a) Normalized Total Gain, (b)Permutation Importance*

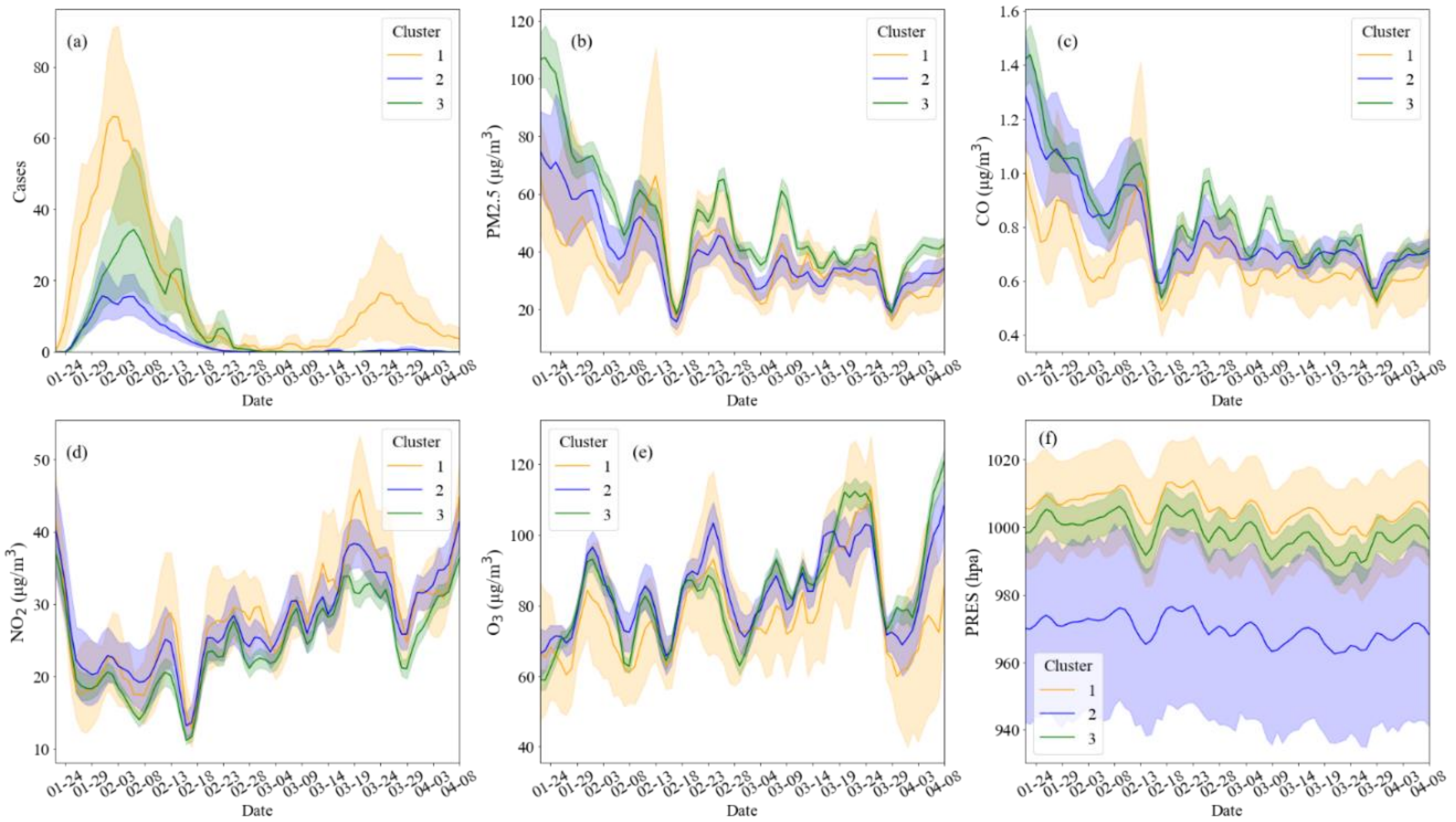

Figure S5. COVID-19 cases (a) and five selected features: (b) PM2.5, (c) CO, (d) NO$_2$, (e) O$_3$, (f) atmospheric pressure. Colored bands indicate 95% confidence intervals.

References

(1) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: San Francisco California USA, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.
(2) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in neural information processing systems* 9.
(3) Pearl, J. *Causality: Models, Reasoning, and Inference*; Cambridge University Press: Cambridge, U.K. ; New York, 2000.
(4) Huang, Y.; Valtorta, M. Pearl's Calculus of Intervention Is Complete. *arXiv:1206.6831 [cs]* **2012**.
(5) Louizos, C.; Shalit, U.; Mooij, J.; Sontag, D.; Zemel, R.; Welling, M. Causal Effect Inference with Deep Latent-Variable Models. *arXiv:1705.08821 [cs, stat]* **2017**.
(6) Dehejia, R. H.; Wahba, S. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association* **1999**, *94* (448), 1053–1062. https://doi.org/10.1080/01621459.1999.10473858.
(7) Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; Nolan, G. P. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science* **2005**, *308* (5721), 523–529. https://doi.org/10.1126/science.1105809.