Hygiene and Environmental Health Advances 4 (2022) 100026

Contents lists available at ScienceDirect

Hygiene and Environmental Health Advances

journal homepage: www.elsevier.com/locate/heha

Robust time-series analysis of the effects of environmental factors on the CoViD-19 pandemic in the area of Milan (Italy) in the years 2020–21

Carlo Grillenzoni

IUAV: Institute of Architecture, University of Venice, Italy

ARTICLE INFO

Keywords:
ARX models
Gain multipliers
Granger causality
M-type estimators
Residual outliers

ABSTRACT

The effects of environmental factors on the spread of the CoViD-19 pandemic have been widely debated in the scientific literature. The results are important for understanding the outbreak dynamics and for defining health measures of prevention and containment. Using multivariate autoregressive (AR) models and robust statistics of causality, this paper analyzes the effect of 19 time series (10 physical and 9 social) on 3 daily CoViD-19 series (infected, hospitalized, deaths) in the Milan area for about 16 months. Robust M-estimation shows the weak effect of climatic and pollution factors, while authority restrictions, people mobility, smart working and vaccination rate have a significant impact. In particular, the vaccination campaign is important for reducing hospitalizations and deaths.

1. Introduction

Discovering the causes of events is the fundamental purpose of scientific reasoning. The CoViD-19 pandemic is a complex phenomenon that has important consequences on world society. Understanding direct and indirect factors of its spread is fundamental for making the right decisions of control. The major way of infection is the personal contact, but what is the role of the external (socio-environmental) conditions? And how do these conditions interact with personal contacts? The availability of data is important for such analysis, but statistical techniques of data treatment are even more important. The study of causality has been formalized by Clive Granger (2005 Nobel prize for economics), using time-series data and their autoregressive (AR) modelling. In this paper, we follow this approach.

Direct contact between people is the main cause of virus infection, including swap cough and sneeze droplets (see Wang et al. 2021). External environmental factors may intervene in the diffusion of pandemic episodes in two ways: directly, as means of transmissions, such as dust and wind, or indirectly as conditions that weaken the human health system, such as cold and pollution. Moreover, cool and wet weather conditions induce indoor activities and human gatherings, and so personal contacts; however, government restrictions on people's mobility, reduce car traffic and, in turn, also the air pollution.

Among the pollution agents, fine particulate matter (PM) may both carry pathogen agents and cause oxidative stress and inflammation in the lungs, especially those with magnetic (water-soluble metal ions) components (Martinez et al., 2022). Hence, PM could represent a double agent of CoViD-19 propagation, as increases the persistence of the virus in the air and act on the receptor ACE2, which is important for the virus entry into cells (Borro et al., 2020). Thus, many studies have analyzed the presence of the CoViD-19 virus in the air and on PM particles in various environments, detecting a limited presence of virus RNA in outdoor contexts (Setti et al., 2020) and so low risk of contamination (Belosi et al., 2021), unless large gatherings of people are allowed, and social distancing and masking are avoided.

Unlike classical influenza viruses, which almost occur in the winter season (as shown by the regular seasonality of the mortality series), recent events have shown that CoViD virus is less dependent on climatic factors (as the third wave started in summer). In this case, the effect of people's habits should be considered, although they are difficult to detect and measure. Analogously, the effectiveness of health-authority restrictions should be investigated, either with respect to the actual reduction of the people's movement or the mitigation of infections and casualties (García-Cremades et al., 2021). In this case, the timeliness of government decisions and their spatial diffusion play an important role.

This analysis looks like that between air pollution and lung mortality where, in *regional* and spatial data, *spurious* correlation may arise from the common dependence of the variables on population factors. In particular, the number of people affected by CoViD and the level of PM may be inflated by their common dependence on the population level and population density (Hou et al., 2021). Dealing with per-capita CoViD patients and per-capita PM levels is the simplest solution (Copiello and Grillenzoni, 2020); however, to understand the real drives and dynamics of the CoViD pandemic, the analysis of time-series data is necessary.

In general, spurious (indirect) correlation is *not* present in time-series Y_t, X_t where the unidirectional (past-present) ordering of the index t

E-mail address: carlog@iuav.it

https://doi.org/10.1016/j.heha.2022.100026
Received 29 March 2022; Received in revised form 3 September 2022; Accepted 8 September 2022

hinders the interchangeability of data. It may arise at a simultaneous level (e.g. D'Albis et al. 2021, Table A2) if the series have a common trend and are only regressed on their present, as $Y_t = \alpha + \beta\ X_t + e_t$ (see Granger, 2012). However, when AR models are used, and proper lagged variables Y_{t-h}, X_{t-k} are introduced (to avoid residual autocorrelation), then genuine relationships can be detected. In the Granger's causality analysis, the prediction ability of X_{t-k} on Y_t is statistically tested and can be applied to pandemic events and environmental factors.

There are various studies that investigate the Granger's causality between CoViD spread and environment conditions; they mainly differ in the nature of the covariates X_t. The first group deals with *physical* factors: Delnevo et al. (2020), Zheng et al. (2021), Martinez et al. (2022) consider pollution indicators and classical AR models; while Sarkodie et al. (2020), D'Albis et al. (2021), Sharma et al. (2021) focus on climatic conditions and use dynamic panel data models. The second group deals with *social* factors: Magyar et al. (2021), García et al. (2021), Sato et al. (2021) consider human interaction indicators based on Internet; while Habib et al. (2021), Li et al. (2021), Mastakouri and Schölkopf (2020)) use real mobility and authority restrictions data. Table A in the Appendix summarizes their technical aspects, which support the existence of causality between CoViD outbreaks and various environmental conditions.

However, the results of these studies may be affected by the quality of data (short time series), model specification (small lag orders) and non-stationarity of residuals (presence of outliers). In this article, we avoid these drawbacks by using a comprehensive dataset, large multivariate models and robust estimators. The plan of work is as follows: Section 2 describes the case-study and its dataset; Section 3 presents the statistical models and methods; Section 4 provides the estimation results; Section 5 discusses and compare the main findings.

2. The case study

The metropolitan area of Milan is the most populated and developed part of Italy; it accounts for 3.25 million inhabitants with a density of about 2100 per km^2. For this reason, it was the *epicenter* of the first CoViD-19 outbreak in Europe in Spring 2020. In this study, we consider the daily number of new CoViD cases Y_t (infected, hospitalized and deaths) in the period from Feb. 24, 2020 to July 4, 2021; a total of N=497 days. These series are displayed in Fig. 1a, showing a clear situation of *non-stationarity*, i.e. their path is characterized by changes in level and variability, located at specific periods (Grillenzoni, 1998).

The first attempt to *regularize* such series consists of transforming them into natural logarithms, as $\log(Y_t + 1)$; indeed, Fig. 1b shows a significant stabilization in mean and variance. This provides good statistical properties to parameter estimates, such as efficiency and unbiasedness, which are necessary for hypothesis testing. Fig. 1c, d show the residuals of AR models of lag-order 10, fitted to the series of daily new infected cases. One can see that residuals of the log-transformed series are more heteroskedastic, Normally distributed and with few outliers. However, the null hypotheses of stationarity and Gaussianity are still rejected in Fig. 1d, and this affects the statistical properties of the models. Therefore, in the next sections, the models will be fitted with *robust* estimators (Huber, 1981), which allow unbiasedness to both coefficients and their standard errors.

A distinctive feature of the series of infected cases is the weekly (7 days) periodicity which becomes apparent from $t > 200$ (Oct. 2020), when the second wave of the pandemic started. This cycle is the result of the regularization of the monitoring process and data collection by health authorities. About the significant outlier (greater than 5σ) which is present in Fig. 1d, at $t = 120$, it may be the result of a data counting error; therefore, its corresponding observation could be replaced by the average $(Y_{t-1} + Y_{t+1})/2$.

The main goal of this paper is to identify the social and physical factors that favor or contrast the spread of the CoViD pandemic. Following the indicators of Table A, we consider four main groups of *exogenous* variables X_{jt}, such as:

(1) Spatial factors: new CoViD cases in the contiguous cities of Lodi and Bergamo;
(2) Authority measures: Vaccination rate, Restriction levels, urban Traffic;
(3) Social indicators: Public Internet usage, as a proxy of the social distancing;
(4) Climatic variables: Temperature, Humidity, Rain, Wind, Pressure and Solar irradiance;
(5) Pollution variables: PM_{10}, $PM_{2.5}$, NO_2 and O_3 (as daily averages).

All variables are listed in detail in Table 1; their data are collected from the Internet sites of the national health authority, local municipality and regional environmental agency. A set of these data is plotted in Fig. 2: they show nonstationary patterns as Fig. 1a,b; for this reason, they are also transformed in logarithm.

3. Statistical methods

3.1. Model representation

The typical feature of a time series Y_t is the presence of autocorrelation (ACR); this represents the dependence of the series on its past values Y_{t-k}. This relationship has the meaning of *memory* of the data generating process and is useful for forecasting future values Y_{t+h}. Besides, it must be adequately represented in regression models as $Y_t = \alpha + \beta X_t + e_t$, in order to get unbiased estimates of the parameters $\alpha,\ \beta$ and their standard errors, which measure the dependence of Y_t on the series X_t.

ACR issues can be addressed by introducing lagged terms Y_{t-k} into the models, thus obtaining the auto-regressive (AR) representation

$$AR(p): \quad Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \ldots + \alpha_p Y_{t-p} + e_t, \qquad e_t \sim IN\left(0,\ \sigma_e^2\right),$$
$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + e_t, e_t \sim \mathrm{IN}(0, \sigma_e^2), \qquad t = p+1, ..., N, \quad (1)$$

where p is the order of memory and e_t are independent Normal (IN) residuals.

In the presence of an explanatory variable X_t, the model of Eq. (1) can be further enriched with the lagged terms X_{t-k}, obtaining the so-called ARX model:

$$Y_t = \alpha_0 + \left(\alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p}\right) + \ldots$$
$$+\left(\delta_0 X_{t-d} + \cdots + \delta_q X_{t-d-q}\right) + u_t,$$
$$u_t \sim \mathrm{IN}\left(0, \sigma_u^2\right), \qquad t = d + q + 1, ..., N, \quad (2)$$

where $u_t \sim \mathrm{IN}(0,\ \sigma_u^2) u_t$ are residuals, q is the order of the exogenous part and d is the delay factor. The variance $\sigma_u^2 < \sigma_e^2$ is smaller than that of e_t in Eq. (1) and the relative difference $D = 1 - \sigma_u^2/\sigma_e^2$ provides the basis for the Granger causality. Instead, a measure of the causal impact $X_t \to Y_t$ is provided by the sum of the regression coefficients $g = \sum_{j=0}^{q} \delta_j$, which may be positive or negative.

3.2. Parameter estimation

Given a sample of N observations Y_t, X_t, the set of parameters of the ARX model (2) $\boldsymbol{\beta}' = [\alpha_0, \alpha_1 \ldots \alpha_p, \delta_0, \delta_1 \ldots \delta_q]$ can be estimated with ordinary least squares (OLS), by using the vector of regressors $\boldsymbol{x}_t = [1,\ Y_{t-1} \ldots Y_{t-p},\ X_{t-d} \ldots X_{t-d-q}]'$ as

$$\hat{\boldsymbol{\beta}}_N = \left(\sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1} \sum_{t=1}^{N} \boldsymbol{x}_t Y_t,$$
$$\boldsymbol{S}_N = \left(\sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1} \hat{\sigma}_u^2, \quad (3)$$

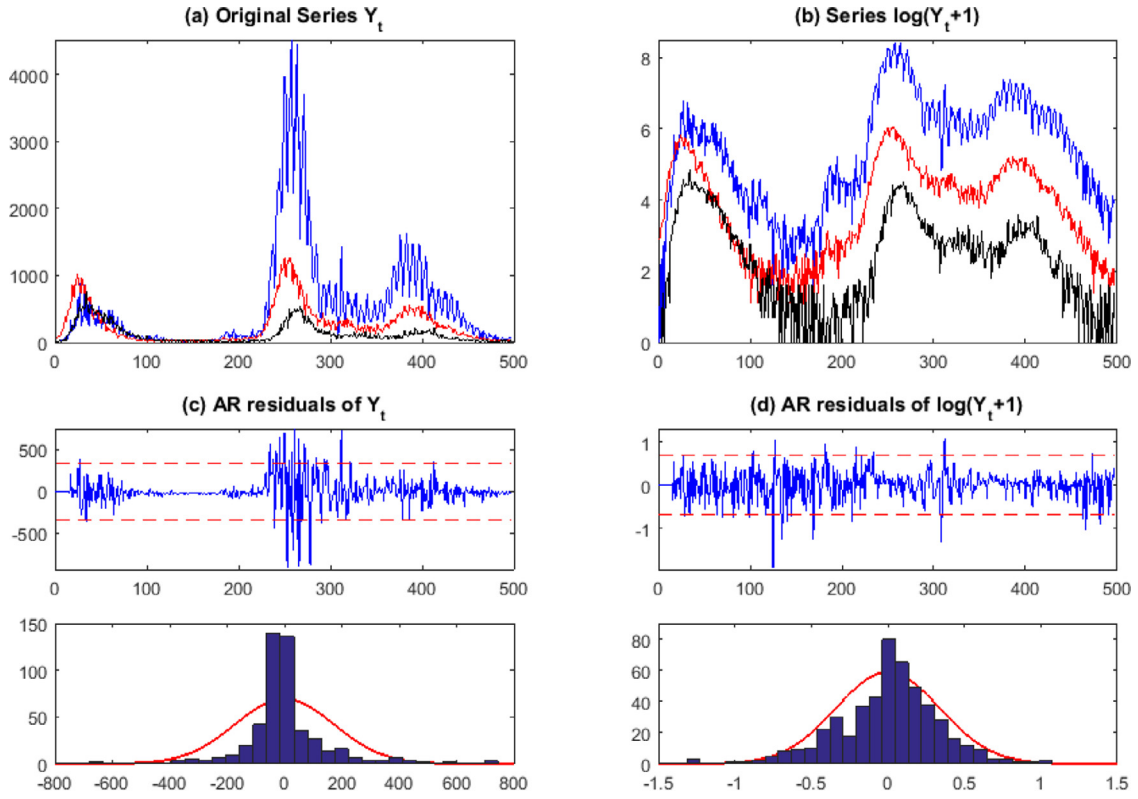

Fig. 1. Daily new CoViD-19 cases in the Milan area from Feb. 24, 2020 to July 4, 2021 ($N = 497$): Infected (blue), Hospitalized (red), Deaths (black). Panels: (a) Original series (with rescaled hospitalized and deaths); (b) Logarithm of the series: $\log(Y_t + 1)$. (c, d) Residuals of AR models with bands $\pm 2\sigma$ and histograms for comparison with the Normal density.

where $\boldsymbol{S}_N$ is the dispersion matrix of $\hat{\boldsymbol{\beta}}_N$. In real data, the process u_t of Eq. (2) is usually *heteroskedastic*; i.e. its variance σ_u^2 is time-varying as σ_t^2. This yields *inefficient* estimates of the parameters $\alpha_i,\ \delta_j$, and *biased* estimates of their standard errors. It follows that statistical inference on the model (2) is biased and so are the tests for causality. To solve this problem, one may adopt heteroskedastic consistent (HC) estimates of the matrix $\boldsymbol{S}_N$, by using $\hat{\sigma}_t^2 = \hat{u}_t^2$ and the *sandwich* matrix (see White 1980)

$$\boldsymbol{S}_{\mathrm{HC}} = \Big(\sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t'\Big)^{-1} \Big(\sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t'\, \hat{u}_t^2\Big) \Big(\sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t'\Big)^{-1} . \tag{4}$$

with the diagonal entries S_{ii} of the matrix (4), one may build asymptotically Normal statistics for evaluating the significance of estimates: $\hat{Z}_i = \hat{\beta}_i / \sqrt{S_{ii}}$.

When the non-stationarity also involves *outliers* (as in Fig. 1c, d), *robust* statistical methods are necessary to avoid the, more severe, problem of bias of the estimates $\hat{\boldsymbol{\beta}}_N$ (see Huber 1981 p.155). The simplest robust approach is the M-type (maximum likelihood-type), which minimizes the sum of a non-negative bounded function $\rho(.)$ of u_t, namely

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} = arg\ \min \sum_{t=1}^{N} \rho\Big(Y_t - \boldsymbol{\beta}'\, \boldsymbol{x}_t\Big), \quad 0 < \rho(u) < u^2. \tag{5}$$

The loss function $\rho(.)$ involves a *tuning* coefficient that must be selected according to the rate of outlier contamination. It is *not* uniformly differentiable and, therefore, Eq. (5) must be minimized iteratively; however, it is smooth enough to allow for convergence even with over-parametrized models. Accordingly, the standard errors of $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ may be computed as in Eq. (3) by using a robust estimate of σ_u^2 (see Huber 1981, p.175). The mean absolute deviation (MAD) $\hat{\sigma}_u^* = \mathrm{median}(|\hat{u}_t|)/0.6745$ may underestimates σ_u, but provides a useful statistic for *cleaning* the residuals as $\hat{u}_t^* = \mathrm{sign}(\hat{u}_t) 2\hat{\sigma}_u^*$, when $|\hat{u}_t| > 2\hat{\sigma}_u^*$ (see Grillenzoni 1997). With the errors $\hat{u}_t^*$, one may compute the robust sum of squares and the variance for the classical *t, F*-statistics; they could also be used in the dispersion matrix (4).

3.3. Causality statistics

The causality between two time-series X_t, Y_t was firstly defined by Granger (1969) and relies on their ARX representations (1)-(2). It states that $X_t \to Y_t$, if the variance reduction $(\sigma_e^2 - \sigma_u^2)$ is significant. Assuming that models are rightly specified (hence $e_t\ , u_t$ are uncorrelated and Normal), the test for causality can be based on the classical *F*-statistic

$$\hat{F}_X = \frac{(\hat{\sigma}_e^2 - \hat{\sigma}_u^2)\ /\ (q+1)}{\hat{\sigma}_u^2\ /\ (N-d-m)} \quad \sim \quad F(q+1;\ N-d-m),$$
$$\hat{F}_X = \frac{(\hat{\sigma}_e^2 - \hat{\sigma}_u^2)\ /\ (q+1)}{\hat{\sigma}_u^2\ /\ (N-d-m)} \quad \sim \quad \mathrm{F}(q+1;\ N-d-m), \tag{6}$$

where $m = (2 + p + q)$. Equivalently, given the relationship between t, F statistics, there exists causality $X_t \to Y_t$ if *at least one* of the estimates $(\hat{\delta}_0,\ \hat{\delta}_1, \dots,\ \hat{\delta}_q)$ in Eq. (2) is statistically significant.

The approach proposed by Granger (1969) is predictive and based on the capability of the series X_t to improve the forecasts of Y_t . However, it does not consider the actual *impact* (positive, null or negative) of the input on the output of the system (2). Following Box and Jenkins (1976), this feature can be measured by the steady-state *gain G*, which is the long-term change in Y_{t+k} yielded by a unit step-change in X_t (that is $Y_\infty = G\,X$). It can be estimated as:

$$\hat{g}_X = \big(\hat{\delta}_0 + \hat{\delta}_1 + \cdots + \hat{\delta}_q\big), \tag{7a}$$

$$\hat{G}_X = \hat{g}_X / \big(1 - \hat{\alpha}_1 - \cdots - \hat{\alpha}_p\big), \tag{7b}$$

Table 1
List of dependent (Y) and independent (X) variables used in the paper.

Group	Variable	Description	Source
Y Dependent	YI	n. of Infected cases in Milan	https://CoViD19.infn.it/
"	YH	n. of Hospitalized cases in Milan	"
"	YD	n. of Death cases if Milan	"
X1 Spatial	YI Lodi	n. of Infected cases in Lodi	https://CoViD19.infn.it/
"	YH Lodi	n. of Hospitalized cases in Lodi	"
"	YD Lodi	n. of Death cases in Lodi	"
"	YI Bergamo	n. of Infected cases in Bergamo	"
"	YH Bergamo	n. of Hospitalized cases in Bergamo	"
"	YD Bergamo	n. of Death cases in Bergamo	"
X2 Authorit	Vaccinations	quote of vaccinated population (%)	https://raw.githubusercontent.com/
"	Restrictions	Authority alert levels (1-5)	https://ourworldindata.org/
"	Traffic vehicles	n. vehicles entering downtown	https://dati.comune.milano.it/
X3 Internet	HS download	downloads in public hotspots (Gb)	https://dati.comune.milano.it/
"	HS upload	uploads in public hotspots (Gb)	"
"	HS users	n. of users of public hotspots	"
"	HS logins	n. of accesses to public hotspots	"
X4 Meteo	Sunlight	n. of hours of sunlight	https://it.climate-data.org/
"	Temperature	daily average temperature (C)	https://www.ilmeteo.it/
"	Humidity	average relative humidity (pct)	"
"	Wind	average wind speed (km/h)	"
"	Pressure	average atmospheric pressure (mbar)	"
"	Rain	total precipitation (mm)	"
X5 Pollution	PM_{10}	average particulate matter ($\mu g/m^3$)	https://www.arpalombardia.it/
"	$PM_{2.5}$	average fine particles ($\mu g/m^3$)	"
"	NO_2	average nitrogen dioxide ($\mu g/m^3$)	"
"	O_3	average ozone level ($\mu g/m^3$)	"

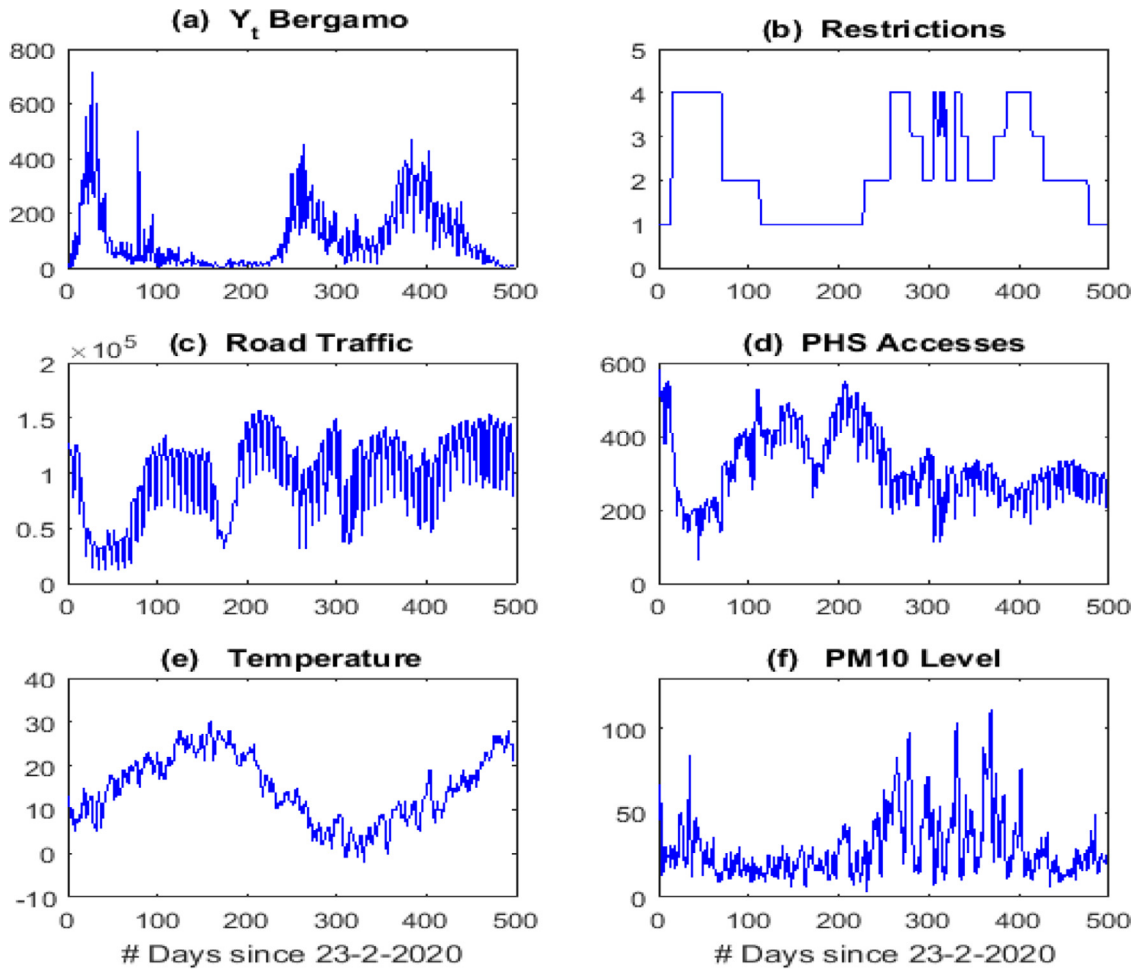

Fig. 2. Graphs of main explanatory variables: (a) Number of daily new infected cases in the nearby area of Bergamo; (b) Levels of authority restrictions (0-5) in the period; (c) Daily number of vehicles entering the central area of Milan; (d) Daily number of accesses to public WiFi hotspots (HS); (e) Daily average temperature; (f) Daily average level of PM_{10}.

where g is the short-term gain, which measures the instantaneous impact.

The importance of these coefficients in the analysis of causality is stressed in Grillenzoni (1997, 2021), where an inferential framework is defined for $\hat{G}_X$. Since the crucial role is played by the numerator (7a), to test for the significance of the impact of $X_t \to Y_t$, one may use the asymptotically Normal statistic

$$\hat{Z}_g = (\hat{\delta}_0 + \hat{\delta}_1 + \cdots + \hat{\delta}_q) / \left(\sum_{i=1}^{q+1} \sum_{j=1}^{q+1} S_{2ij} \right)^{1/2}, \quad (8)$$

where S_{2ij} are the entries in the lower-right block $\boldsymbol{S}_{22}$ of the dispersion matrix (4).

3.4. Models identification

The first step in model building is the identification of the lag orders (p, q, d) of Eq. (2). The log-transformation may reduce non-stationarity, but the dynamic of $\log(Y_t)$ still remains complex, such that the polynomials of the models have an *irregular* (subset) structure, with many missing coefficients $\alpha_i,\ \delta_j$. Further, the AR components α_i depend on both the auto-correlation of Y_t, and the cross-correlation between $X_t,\ Y_t$, which is complicated by the delay factor $d > 0$. In these cases, the identification of (p, q, d) cannot be pursued through the analysis of the correlation functions (see Box and Jenkins 1976) or the information criteria (as that of Akaike), and the model building must be entirely *parametric*, such as:

Algorithm 1:

(1) Select large orders (p, q), on the basis of the available data and a-priori knowledge;
(2) Estimate the model (2) with p+q+2 coefficients, with the robust methods (4)-(5);
(3) Remove non-significant coefficients $\hat{\alpha}_i,\ \hat{\delta}_i$ when their statistics $|\hat{Z}_i| < 2$.

In the presence of multiple factors $X_{jt},\ j = 1, 2, \ldots m$, which are mutually dependent, the causality analysis requires the application of multiple-inputs $\text{ARX}_m(p,q)$ models, as

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{m} \sum_{i=0}^{q} \delta_{ji} X_{jt-d-i} + u_t\ , \quad (9)$$

The identification of the system (9) is challenging even with parametric methods; for example, with $p = q = m = 10$ the model involves 121 coefficients. When there are many regressors, the *backward* stepwise procedure (of selecting significant X_{jt-i}) is biased upward, while the *forward* approach provides estimates which depend on the order of inclusion of X_{jt-i}. As a solution, we propose an extension of the above Algorithm:

Algorithm 2:

(1) Assemble the m subset bivariate models, identified in Steps 1–3 above;
(2) Estimate the resulting model (9) with the robust methods (4)-(5);
(3) Remove non-significant coefficients $\hat{\alpha}_i,\ \hat{\delta}_{ji}$ with statistics $|\hat{Z}_i| < 2$.

This approach has the capability of deleting the regressors X_{jt-i} which are related to Y_t just because depend on other X_{ht-k}; hence, it enables to analyze the multivariate causality in a transparent way. Furthermore, by keeping the size of the regression matrix low (as only significant bivariate terms X_{jt-i} are included in the multivariate model), the Algorithm may avoid the numerical problems of near-singularity of the matrices in Eqs. (3), (4), which are typical of multicollinearity (see Grillenzoni 2021).

4. Statistical results

Given the weekly periodicity of the confirmed CoViD infections, we first fit an AR(p=14) model to $\log(YI_t + 1)$, obtaining only 7 significant coefficients $\hat{\alpha}_i$ at lags: 1,2,5,6,7,9,11, with $\hat{\sigma}_e^2$= 0.345. The selection of the order q is complicated by the fact that the effect of social and physical factors on CoViD outbreaks is very complex and with long run. The *mean delay* between infection and symptoms onset is about 1 week (like that from symptoms to hospitalization), while that from symptoms to death is about 3 weeks; see Table 2, based on the review by Sutherland et al. (2021) and Martinez et al. (2022). Similarly, the delay of external factors on infections should be 1-2 weeks, which is the time that people react/adapt to climate conditions and authority constraints. For these reasons, in the following we set p =14 and q =21-35, depending on the CoViD data (infected, hospitalized, deaths), see Table 2.

4.1. Bivariate models

Table 3 reports the causality statistics of ARX(14,21) models of the CoViD infected series in Fig. 1 and the exogenous variables in Table 1. All series are expressed in logarithm and the models have been estimated with the robust method (5) with *bisquare* function $\rho(u)$ and tuning coefficient 4.5; this allows 95% relative efficiency with respect to OLS.

One may note that the role of climatic factors is weak or absent (temperature, humidity, pressure, rain); the role of pollutants is significant in terms of predictive statistics D, F in Eq. (6), but is weak with respect to their impacts g, G in Eq. (7). In particular, the ozone has a significant *negative* effect on Y_t, which may be consistent with the disinfectant capability of O_3 in indoor environments (e.g. Tizaoui et al. 2022). Major causality effects are provided by social factors, especially the neighboring areas of Bergamo and Lodi (but with different signs), together with the rate of vaccinated people, the authority restrictions and the people's mobility, both as vehicle traffic and access to public hotspots (HS) and Internet usage.

The results of Table 3 are substantially confirmed on the *differenced* series $y_t = \log(Y_t) - \log(Y_{t-1})$, on which the analysis may also be carried out. The reason is that the trends in the time series of Fig. 1 are not marked (in particular, on the data expressed in logarithm), and the statistical properties of estimators in Eqs. (3)–(5) still hold on the series with trends (see Granger 2012 and Grillenzoni 1998). In general, it is preferable to work directly on original data to have interpretable signs of $\hat{\beta}_j$ and gains g, G.

4.2. Multivariate models

Results of Table 3 do not consider the mutual dependence of the regressors X_{jt} and of their lagged terms; the application of the multivariate system (9) is then necessary. Following the identification algorithm of Section 3.4 (i.e. assembling the bivariate subset models of Table 3), yields a system (9) with *only* 74 parameters, on which the backward selection procedure works without bias (see Grillenzoni 2021). Table 4 reports the estimates obtained with the estimator (5), with bisquare loss and coefficient 4.5. While in the bivariate models of Table 3, OLS and M-estimates provide similar results, in the multivariate system (9) the robust method identifies a model with −40% of coefficients than OLS. The most significant covariates in Table 4 are the social variables as neighboring areas and mobility trends, both for vehicles and persons (these are the practical effects of authority restrictions). As regards the pollution factors, only $PM_{2.5}$ and NO_2 are present in the model, but PM has a *negative* impact, which means that PM may contrast CoViD infection; a feature which is also present in Sharma et al. (2021). Instead, HS variables have the right signs, as WiFi upload activity is related to people smart working (based on public HS services), whereas the number of logins is related to outdoor people.

As regards the evaluation of the model fitting in Table 4, the robust index $R^2 = 1 - \hat{\sigma}^2_{u^*} / \hat{\sigma}^2_{Y^*} = 0.819$ is obtained from the series $\hat{u}_t^*, \hat{Y}_t^*$ cleaned by outliers with the 2σ rule and their MAD statistics (see Section 3.2). Further, the overall contribution of the exogenous variables X_{jt-i} (with respect to the simple AR model of YI_t) is provided by the statistic $\hat{F}_X = 8.62 > 2.07 = \text{F}_{0.01}$ (16,453), which is 99% significant.

Table 2
Summary of the delays (in number of days) between CoViD pandemic events.

Events	Medians' Range	Mean	Order q	Data Y_t
External Causes to Infection	1-7	5	.	.
Infection to Symptoms onset	1-14	5.5	21	Infected
Symptoms to Hospitalization	1-6.7	5	.	.
Symptoms to ICU	6-10.5	7	28	Hospitalized
Symptoms to Death	14-56	17.5	.	.
ICU to Death	7-12.5	9	35	Deaths

Table 3
Causality statistics of ARX(14,21) models (2) of the infected (I) CoViD series YI_t on social and environmental factors X_{jt} described in Table 1.

X_j	D_σ	m_X	$\hat{F}_X$	$F_{0.01}$	p-val	$\hat{g}_X$	$\hat{Z}_g$	p-val	$\hat{G}_X$
YI Lodi	0.185	6	20.901	3.057	0.000	0.102	3.314	0.000	0.879
YI Bergamo	0.133	6	17.778	3.360	0.000	-0.017	-0.696	0.243	-0.833
Vaccinations	0.046	6	3.723	2.841	0.001	-0.371	-0.959	0.169	-12.633
Restrictions	0.037	3	18.183	6.689	0.000	-0.216	-2.859	0.002	-38.420
Traffic vehic	0.176	5	19.712	3.057	0.000	0.071	2.129	0.017	2.200
HS downl.	0.008	2	2.016	4.651	0.134	-0.020	-0.881	0.189	-0.554
HS upload	0.013	2	3.003	4.651	0.051	-0.080	-1.378	0.084	-2.131
HS n. users	0.117	7	10.242	2.841	0.000	0.200	2.855	0.002	11.279
HS n. logins	0.093	6	9.443	3.057	0.000	0.166	2.012	0.022	8.538
Sunlight	0.019	1	8.981	6.689	0.004	-0.254	-3.307	0.000	-3.466
Temperature	0	0	0	0	1	0	0	1	0
Humidity	0	0	0	0	1	0	0	1	0
Wind	0.042	5	4.108	3.057	0.001	-0.097	-1.341	0.090	-2.535
Pressure	0	0	0	0	1	0	0	1	0
Rain	0	0	0	0	1	0	0	1	0
$PM_{2.5}$	0.021	3	5.109	4.651	0.006	0.057	1.421	0.078	1.174
PM_{10}	0.025	3	3.992	3.824	0.008	0.078	1.766	0.039	1.694
NO_2	0.115	5	9.974	2.841	0.000	0.207	2.958	0.002	3.414
O_3	0.018	1	8.383	6.689	0.004	-0.049	-2.377	0.009	-0.931

Legend: $D_\sigma = 1 - \hat{\sigma}_u^2/\hat{\sigma}_e^2$ is the % reduction of the residual variance due to X_t; m_X is the number of 95% significant regressors X_{jt-i}; the causality indicators $\hat{F}_X$, $\hat{g}_X$, $\hat{Z}_g$ are explained in Eqs. (6)–(8). The p-values are one-sided even for Z-statistics and the characters in bold indicate very significant results.

Table 4
Robust estimates of the model (9) on YI_t = CoViD infected data, with m=19 covariates and orders p=14, q=21. Only 95% significant coefficients $\hat{\beta}_l$ are reported (i.e. $|\hat{Z}_l| > 1.96$).

X_{jt-i}	lag i	$\hat{\beta}_l$	$\hat{Z}_l$	p-val.	$\hat{g}_j$	$\hat{Z}_g$	p-val.
Const.	0	-0.641	-1.302	0.096			
YI_{t-1}	1	0.240	6.689	0.000			
YI_{t-2}	2	0.247	6.727	0.000			
YI_{t-3}	3	0.170	4.589	0.000			
YI_{t-6}	6	0.095	2.704	0.003			
YI_{t-7}	7	0.271	7.200	0.000			
YI_{t-11}	11	-0.101	-3.333	0.000			
YI Lodi	0	0.123	5.919	0.000	.	.	
YI Lodi	17	-0.041	-2.222	0.013	0.083	2.980	0.001
YI Berg.	0	0.124	5.602	0.000	.	.	.
YI Berg.	7	-0.060	-2.519	0.006	.	.	.
YI Berg.	9	-0.058	-2.675	0.004	.	.	.
YI Berg.	10	-0.051	-2.204	0.014	-0.046	-1.009	0.156
Traffic	1	0.133	2.823	0.002	.	.	.
Traffic	4	-0.206	-3.979	0.000	.	.	.
Traffic	5	-0.275	-3.955	0.000	.	.	.
Traffic	12	0.308	5.661	0.000	-0.040	-0.355	0.351
HS upload	11	-0.107	-2.597	0.005	-0.107	-2.597	0.005
HS logins	15	0.171	2.802	0.003	0.171	2.802	0.003
Wind	6	0.107	2.620	0.004	0.107	2.620	0.004
$PM_{2.5}$	8	0.093	2.482	0.007	.	.	.
$PM_{2.5}$	9	-0.135	-3.607	0.000	-0.042	-0.784	0.217
NO_2	1	0.121	2.572	0.005	0.121	2.572	0.005
Indices	$\hat{\sigma}^2_{u^*}$=0.273	R^2=0.819	$\hat{F}_X$=8.62	$F_{0.01}$=2.07	p-value 0.000		

Legend: $\hat{\sigma}^2_{u^*}$ is the robust residual variance; $\hat{F}_X$ is the statistic of the *global* contribution of all X_{jt-i}; $\hat{g}_j = \sum_i \hat{\delta}_{ji}$ are the gains of the X_{jt} present in the final model (9), computed as in Eqs. (7a) and (8).

Table 5
Robust estimates of the model (9) on YH_t = CoViD hospitalized (H) data, with m=19 covariates and lag orders p=14, q=28. Only 95% significant regressors are reported.

X_{jt-i}	lag i	$\hat{\beta}_l$	$\hat{Z}_l$	p-value	$\hat{g}_j$	$\hat{Z}_g$	p-val.
Const.	0	-1.167	-3.403	0.000			
YH_{t-1}	1	0.424	11.245	0.000			
YH_{t-2}	2	0.103	2.793	0.003			
YH_{t-3}	3	0.228	6.088	0.000			
YH_{t-4}	4	0.186	4.691	0.000			
YH_{t-9}	9	0.183	4.876	0.000			
YH_{t-14}	14	-0.113	-3.969	0.000			
YH Lodi	2	-0.044	-2.432	0.008	.	.	.
YH Lodi	16	-0.038	-2.294	0.011	-0.082	-3.34	0.000
Vaccination	8	-64.295	-3.144	0.001	.	.	.
Vaccination	9	66.867	3.072	0.001	.	.	.
Vaccination	23	-93.925	-4.411	0.000	.	.	.
Vaccination	25	221.863	4.140	0.000	.	.	.
Vaccination	26	-130.567	-3.627	0.000	-0.575	-1.23	0.021
Traffic	0	0.287	6.369	0.000	.	.	.
Traffic	1	-0.250	-6.186	0.000	.	.	.
Traffic	4	-0.083	-2.745	0.003	.	.	.
Traffic	14	0.161	4.988	0.000	0.115	1.75	0.041
Indices	$\hat{\sigma}^2_{u^*}$=0.238	R^2=0.821	$\hat{F}_X$=7.05	$F_{0.01}$=2.29	p-value 0.000		

Legend: see Tables 3 and 4.

Table 6
Robust estimates of the model (9) on YD_t = CoViD deaths (D) data with m=19 covariates and orders p=14, q=35. Only 99% significant regressors are reported.

X_{jt-i}	lag i	$\hat{\beta}_l$	$\hat{Z}_l$	p-value	$\hat{g}_j$	$\hat{Z}_g$	p-val.
Const.	0	1.101	4.194	0.000			
YD_{t-1}	1	0.204	4.938	0.000			
YD_{t-3}	3	0.131	3.268	0.001			
YD_{t-4}	4	0.373	9.115	0.000			
YD_{t-5}	5	0.226	5.357	0.000			
YD Bergamo	6	0.084	2.706	0.003	0.084	2.706	0.003
Vaccination	2	-136.805	-3.334	0.000	.	.	.
Vaccination	3	177.169	3.705	0.000	.	.	.
Vaccination	10	-272.755	-4.042	0.000	.	.	.
Vaccination	11	275.392	3.830	0.000	.	.	.
Vaccination	15	-144.985	-3.484	0.000	.	.	.
Vaccination	17	171.709	4.012	0.000	.	.	.
Vaccination	21	-105.507	-4.058	0.000	.	.	.
Vaccination	30	228.340	3.090	0.001	.	.	.
Vaccination	31	-192.903	-2.876	0.002	-0.345	-1.32	0.093
Restrictions	22	-0.352	-3.890	0.000	-0.352	-3.890	0.000
Sunlight	16	-0.288	-3.035	0.001	-0.288	-3.035	0.001
Indices	$\hat{\sigma}^2_{u^*}$=0.431	R^2=0.653	$\hat{F}_X$=2.89	$F_{0.01}$=2.22	p-value 0.001		

In the estimation of the model (9) on hospitalized (H) CoViD cases, we extend the order q to 28 days as hospitalization involves longer delay and dynamics (see Table 2). We also exclude from the covariates the series of infected CoViD people YI_t (as YH_t is its subset) to better understand the role of exogenous factors. Table 5 provides robust estimates; the major fact is that all environmental covariates do not enter the model, while the rate of vaccinated people has an important role, with high delay and cumulated negative impact. This confirms the role of vaccination to avoid hospitalization; while the traffic variable confirms its positive role, as indicates that people meet and may contaminate. The global fitting performance is as good as that in Table 4. Finally, we have also checked that these results *do not* substantially change by also including the infected CoViD series (of Milan, Lodi, Bergamo) in the model.

Table 6 provides the model of the death (D) CoViD series YD_t, with lag order q = 35 days and by excluding infected and hospitalized series from the covariates, to better understand the role of exogenous variables. The estimates stress the importance of the vaccination campaign, whose dynamic effects are complex but overall negative on deaths (with global gain $\hat{G}_N = -5.06$), hence positive on surviving. There are other minor covariates with positive effects on survival, such as authority restrictions and solar irradiance, that occur with 2-3 weeks delay. As in Table 5, the great absents from the model are climatic (except sunlight) and pollution factors. Although weaker than previous cases, the model of YD_t has an acceptable global fitting performance with R^2= 0.65 and 99% significant statistic $\hat{F}_X$.

4.3. Main findings

Table 7 summarizes, in *analogical* form, the main results of Tables 3–6 and extends the bivariate analysis to the series YH_t, YD_t. The significance symbols (*) concern the statistic $\hat{F}_X$ (6) of models (2) and the Normal statistics $\hat{Z}_l$ of parameters $\hat{\beta}_l$ in the multivariate model (9). The sign symbols (±) regard the gains $\hat{g}_j$ (7a) when they are significant: $|\hat{Z}_g| > 1.96$; a question mark is added when the sign does *not* agree with the mainstream contagion hypothesis. For example, wind and ozone should have positive signs as may "carry" the viruses (e.g. Belosi et al. 2021); however, also the opposite thesis can be stated as they "clean" the air.

Table 7 can be read both by columns (models) and rows (covariates). By columns, one can see that bivariate models (2) are relatively homogenous on the 3 dependent variables Y_t and identify various sources

Table 7
Synthesis of the estimates in Tables 3–6, with indicators of fitting (*) and impact (±).

Group	X_{jt} \ Model	YI_t (2)	YH_t (2)	YD_t (2)	YI_t (9)	YH_t (9)	YD_t (9)
X1 Spatial	Y-Lodi	*** +	–	–	** +	* –	
"	Y-Bergamo	***	*	*	**		* +
X2 Authorit.	Vaccinations	*	*	*		**	**
"	Restrictions	*** –	** ––	–			* –
"	Traffic vehic.	*** +	*** +	+	**	**	
X3 Internet	HS download						
"	HS upload			*	* –		
"	HS n. users	** +	** ++	* +			
"	HS n. logins	** +	** ++	+	* +		
X4 Meteo	Sunlight	* –	* –	**			* –
"	Temperature						
"	Humidity		* +				
"	Wind	*	–?	–?	* +		
"	Pressure						
"	Rain			* +			
X5 Pollution	PM_{10}	*	*				
"	$PM_{2.5}$	*	*		*		
"	NO_2	** +	** +	* +	* +		
"	O_3	* –?	*	–?			

Legend: (*) Significance of the *F*-statistic (6) in the models (2) of Table 3, and *Z*-statistics of parameters in the models (9) of Tables 4–6. (±) Sign of the 95% significant Z_g-statistic (8); the question mark mean that the sign of the gain *g* does *not* agree with the mainstream contagion hypotheses.

of causality. Instead, when fitting multivariate models (9), the number of significant covariates drastically reduces. This is due to the fact that some relationships between X_{jt}, Y_t are indirect, as induced by the dependence among the X_{jt}, and the robust subset algorithms of Sect. 3.4 are effective in removing them. The final outcome is the predominance of social factors over the physical ones, especially in the hospitalized and deaths series. These also prove the efficacy of the undertaken public health measures.

To clearly evaluate the effect of the exogenous variables on the CoViD worst cases, we have omitted the infection series *YI* in the models of hospitalized and death cases *YH, YD*; in fact, these series are just subsets of *YI* (see Fig. 1a, b). However, *YI* also regards the monitoring activity of health authorities and the people's behavior (through the number of swabs made); hence, *YI* could have a prevention role in pandemic worst cases. However, we have checked that the inclusion of *YI* in the multivariate models for *YH, YD* does *not* substantially modify the estimates in Tables 5–6.

5. Discussion and conclusion

5.1. Comparisons

Recent studies on the analysis of causality between CoViD and environmental time series may be grouped according to the type of covariates X_{jt} and models they use. In the first group, Delnevo et al. (2020), Zheng et al. (2021), Martinez et al. (2022) consider pollution indicators (as X5) and use classical AR models; while Sarkodie et al. (2020), D'Albis et al. (2021), Sharma et al. (2021) focus on climatic conditions (as X4) and dynamic panel-data models. In the second group, Magyar et al. (2021), García et al. (2021), Sato et al. (2021) consider social interaction indicators based on Internet (as X3); while Habib et al. (2021), Li et al. (2021), Mastakouri and Schölkopf (2020) use real mobility and authority restrictions data (as X2). Table A in the Appendix summarizes their technical features and confirms the existence of causality between CoViD pandemic and various environmental conditions. However, their findings may be affected by the data quality (short time series), model misspecification (small lag-orders and bivariate models), non-stationarity of residuals (heteroskedastic and with outliers) and the absence of health policy measures.

Compared to these works, the improvements of our study consist of using:

(1) Complete dataset with 3 dependent and 19 independent factors, grouped in 5 categories;
(2) ARX models with long memory and large dimensions, which host all 19 covariates and their lags;
(3) Robust estimators, resistant to outliers, and algorithms for subset model identification.

When properly applied, these solutions provide unbiased estimates of the causality indicators; hence, the empirical results of Section 4 can be evaluated with confidence:

(1) *Spatial factors*: Due to the spatial autocorrelation, the CoViD series of neighboring areas have an important role in the spread of a pandemic. Lodi and Bergamo have intense social and economic exchanges with Milan, through commuter workers. When the outbreak started, the first identified case was in Lodi, while the first deadly wave (mainly in the elderly hospices) was in Bergamo (see Fig. 2a). From the regional ICU place planning, the 3 cities have also exchanged hospitalized patients, and therefore deaths. Table 7 shows a significant role of the neighboring series on that of Milan, which also persists in the multivariate models (9). This result agrees with the studies which have used panel data models, that include the spatial autocorrelation of CoViD series in order to filter the confounding factors of causality; see Sarkodie et al. (2020), D'Albis et al. (2021) and Sharma et al. (2021).
(2) *Authority measures*: These variables deal with one of the main questions in the CoViD research: Are restrictions by health authorities effective in contrasting the pandemic? From Table 7, the answer is positive, both in bivariate and multivariate models (9). Alarm levels, mobility restrictions and vaccination campaigns are the major factors that influence (and control) the CoViD spread in Milan city. In particular, the role of vaccinations is enhanced in the multivariate models of the crucial series YH_t, YD_t. The cited papers did not consider vaccination, as they concerned the initial period of the pandemic, but the role of alarm levels and mobility trends are significant in Li et al. (2021), Mastakouri and Schölkopf (2020) and García-Cremades et al. (2021) and Habib et al. (2021).
(3) *Internet usage*: To have indicators of smart working and social distancing (i.e. the actual fulfillment of authority restrictions), we have

considered Internet data on the use of free WiFi nets of municipal hotspots. The total gigabyte flows in download and upload are not significant for the CoViD series, and are not indicators of smart working; instead, the number of users and logins are very significant at the bivariate level and may be actual measures of social distancing and lockdown fulfillment. This result agrees with the analysis Magyar et al. (2021) and Sato et al. (2021); but it considerably weakens when is inserted into the multivariate models for the presence of cofactors, such as vehicular traffic and alarm levels (group 2).

(4) *Climatic conditions*: The hypothesis that the CoViD pandemic depends on climatic conditions is widely debated in environmental studies; however, as shown in Table A, the findings are not homogeneous. We considered 6 weather indicators that may act on CoViD time series either directly (as carriers of the virus, as the wind) or indirectly (through people gathering, as the cold). In Table 7, both bivariate and multivariate models agree to show that only the solar activity acts on (and against) the pandemic, while the sign of the wind impact is not uniform. Using different models, Lolli et al. (2020), Sarkodie et al. (2020), D'Albis et al. (2021) find that humidity and temperature cause CoViD series, but with different signs. Instead, Zheng et al. (2021) did not find causality of *all* climatic factors in the main cities of China.

(5) *Pollution conditions:* Even the role of air pollutants is widely debated, either as direct (virus carriers) or indirect (immune depressive) causes of a pandemic. The four typical indicators in Table 7 show a significant role in the bivariate models, but their effects vanish in the multivariate systems, for the prevailing force of other covariates, such as vehicle traffic and wind. Even the sign of the pollutant impacts generates doubts, given the negative effects of O_3 and $PM_{2.5}$ (in Table 4) on the CoViD series, which were also found by Sharma et al. (2021) and Tizaoui et al. (2022). However, many authors did not use multivariate models and did not indicate the sign of the impacts, e.g. Delnevo et al. (2020) and Martinez et al. (2022).

In summary, our numerical results show the supremacy of the social covariates over the physical ones in explaining the dynamic of the CoViD pandemic in Milan. In particular, the major neighboring cities, the alarm levels (Fig. 2b, where the maximum value corresponds to the total lockdown), the access of vehicles to downtown (Fig. 2c), the public hotspot usage (Fig. 2d, as indicators of social distancing), significantly explain the spread of the CoViD infection (Tables 3, 4). Further, direct medical measures, such as the vaccination rate, contrast both hospitalizations and deaths (Tables 5, 6); this is a clear statistical evidence of the effectiveness of the public health measures undertaken by Italian authorities.

As regards the combined effect of physical factors, we have seen that climatic and pollution variables are absent or weak in the global models (Tables 4–6). This result contrasts with the findings of some studies in Table A, especially those based on panel models (which combine the data of non-homogeneous regions), those with a short period of time (i.e. small sample size), and those with small lag orders *p,q* (with consequent model misspecification). Further, as shown in some studies in Table A, the sign of the coefficients of physical variables may not agree with that expected from their common-sense relationships. This is the case, for example, of the negative effect of PM on the CoViD infected cases, which is also present in Table 4. The fact that pollution variables weaken their effect in the joint models (and their impact may change sign with respect to Table 3), is probably due to their correlation with the vehicle traffic, whose volume strongly depends on the authority restrictions.

5.2. Conclusions

The topic of the relationships between CoViD spread and social and environmental conditions is widely debated in the literature, both theoretically and empirically, and is important to set up control actions. Statistical methods are used to test for the existence of causality, as those based on time-series models. Our study has concerned the urban area of Milan which was the epicenter of the first CoViD outbreak in Italy, during the period 2020–2021. Although specific, this area holds millions of people and is homogeneous; the dataset is comprehensive as the number of periods is about 500 and the number of variables is 20.

Owing to robust estimation methods and subset identification algorithms we have obtained multivariate AR models which exhibit the predominance of social factors and public health measures over climatic and pollution conditions. This proves the validity of the actions undertaken by the Italian authorities, such as: borders lockdown, mobility limitations, smart working, closure of activities, social distancing, hygienic behavior, etc. These are non-medical preventive measures that have, however, important costs on social and economic activities and had to be carefully managed.

More importantly, our study has shown the role of vaccines and vaccination campaigns in reducing hospitalizations and deaths, and thus in defeating the CoViD pandemic. This is is important evidence against the disinformation carried out by no-vax and skeptical movements. By the way, the role of statistics is finding objective signals in complex phenomena (Appendix A1 Appendix A2, Eqs. (7b), (9)).

Declaration of Competing Interest

As regards the paper
"Robust time-series analysis of the effects of environmental factors on the CoViD-19 pandemic in the area of Milan (Italy) in the years 2020-21" submitted for publication in "Hygiene and Environmental Health Advances "
I declare that

- The research received No external financial or non-finacial support
  There are No additional relationships to disclose
  There are No patents to disclose
  There are No additional activities to disclose

The author,
Carlo Grillenzoni

Declaration of Competing Interest

The author declares that He has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Literature review

Table A summarizes the main features of the studies which apply the analysis of Granger causality to CoViD pandemic and environmental time series. It displays a variety of models and results which can be classified into two broad groups, according they use physical (climate and pollution) or social (human interactions and mobility) covariates. In general, they show the existence of statistical dependence, but there are limitations due to the short observation period (few months) and the high spatial aggregation (at the country level) of the datasets. Also, the models used have small lag orders and are mainly fitted at bivariate level (with one covariate at time). Further, the analysis of residuals of regression is not displayed, for checking the presence of outliers which make parameter estimates and test statistics biased.

A.1. Physical factors

One of the first works was carried out by Delnevo et al. (2020) on a limited temporal (3 months) and spatial (a region) scale in Italy. They detect a significant dependence between daily air pollutants ($PM_{2.5}$, PM_{10}, NO_2) and daily new CoViD-19 infections, either in the pre- or

Table A
Main technical features of the studies which apply Granger causality to CoViD data.

Authors	Space span	Time span	CoViD variables	Covariates	Model type	Lag order	Conclusions
Delnevo et al. (2020)	9 cities, 1 region, Italy	daily, Feb-Apr, 2020	Infected	Pollutants PM, NO_2	bivariate AR	14.	Causality of PM and Infections, pre- and post- lockdown (Tables 4, 6)
Zheng et al. (2021)	265 cities, China	daily, Jan-Feb, 2020	Infected	PM, AQI and Meteo	bivariate AR	3.	Non causality of Meteo data on Infections in the main cities (Table 4)
Martinez et al. (2022)	2 cities, GR, IR	daily, Mar-Oct, 2020	Infected and Deaths	proxy $PM_{2.5}$	bivariate AR	30.	Significant uni-directional causality PM25 on deaths (Table 1)
Sarkodie et al. (2020)	20 countries	daily, Jan-Apr, 2020	Infected, Hospitalized, Deaths	Meteo (temp, hum, wind, etc.)	dynamic panel with fixed-effects	NA	Significant Causality of all meteo variables on all CoViD (Tables 3, 4)
Sharma et al. (2021)	10 countries	daily, Feb-Jun 2020	Infected and Deaths	all Meteo and Pollutants	dynamic panel with fixed-effects	NA	Non-causality betwen PM, wind and the Deaths (Tables 7, 8)
D'Albis et al. (2021)	54 regions, France	weekly, Mar 2020, Jan 2021	Hospitalized and Deaths	Temp, Hum, PM, NO	dynamic panel and multivariate AR	3.	Causality of meteo, but not of pollution (Table A3)
Magyar et al. (2021)	3 countries	daily, Mar-Dec 2020	Deaths	Social Media Indicators	multivariate AR	2.	Bi-directional causality of Entry counts and Deaths (Table 3)
Sato et al. (2021)	9 countries	weekly, Jan-Oct 2020	Infected	54 Google search keywords	multivariate AR	4-8.	Only "loss of smell" keyword present in all countries (Table 3)
Mastakouri and Schölkopf (2020)	12 regions, Germany	daily, Jan-May 2020	Infected	Restrictions and Contiguity	dynamic graphycal model	NA	Strong causal role of spatial intercations (Fig. 3)
Habib et al. (2021)	10 countries	daily, Mar-Jul 2020	Infected	Index of transport mobility	nonlinear by quantiles	1.	Bi-directional causality at all quantilies (Table 3)
García et al. (2021)	Spain	daily, Jul 2020, Jan 2021	14-day cumulative incidence	6 Google mobility indices	bivariate AR	5-14.	Causality from recreation, parks, stations mobility (Table 6)
Li et al. (2021)	50 US states	daily, Mar-Aug 2020	Infected and Deaths	19 socio-economic and mobility	bivariate AR	NA	Causality from various variables in various US states (Tables 2, 3)

post- lockdown periods. Instead, Zheng et al. (2021) analyzed the time series of 265 cities in China in the short period Jan-Feb 2020. They find that the number of confirmed CoViD cases is weakly correlated with air quality indicators (AQI and PM), whereas the Granger dependence on meteorological factors (such as temperature) varies geographically.

Further, D'Albis et al. (2021) use a dynamic *panel* model with specific regional and temporal fixed effects on the data of 54 French regions and 42 weeks. They show the existence of Granger causality between a composite indicator of temperature and humidity (IPTCC) on the number of hospitalized and death cases in the period Mar 2020 - Jan 2021. However, they find the absence of causality between pollution indicators (particulate, ozone and nitrogen) when measured jointly with the composite climatic indicator (D'Albis et al., 2021, Table A3). They also estimate the impact of climatic factors with the dynamic multipliers of a multivariate AR model; obtaining a positive response of temperature and IPTCC on both CoViD indicators (this means that CoViD pandemic spreads in France, as the temperature increases).

Similarly, Sarkodie and Owusu (2020) employ a dynamic panel-data model with specific fixed effects to deal with country heterogeneity and apply the panel Granger test of Dumitrescu and Hurlin (2012) to control cross-section dependence between countries due to common shocks. They find a significant Granger causality between climatic indicators and all CoViD daily data (infected, hospitalized and deaths) of 20 top countries in the period Jan-Apr 2020. However, the model exhibits different signs of the instantaneous elasticities with respect to the 3 dependent variables and they conclude that high temperature and relative humidity reduce cases, whereas the others favor virus survival and hence infectivity.

Even Sharma et al. (2021) use a dynamic panel data model with cross-sectional dependence and heterogeneity effects (Chudik and Pesaran, 2015), on the daily data of the 10 most infected countries in the period Feb-June 2020. They find *bi-directional* Granger causality between climatic and pollution variables and the number of infected and death cases applying the test by Dumitrescu and Hurlin (2012). The sole exceptions are $PM_{2.5}$ and wind variables, which may be attributed to lockdown measures. However, it is difficult to understand the negative elasticity of $PM_{2.5}$ on the CoViD series and the feedback of these on other climate variables.

A.2. Social factors

Regarding social causes, Magyar et al. (2021), analyze the effect of social and internet indicators on the number of infected and deaths of CoViD-19. They consider daily data from Mar-Dec 2020 of 3 countries (US, Spain and Hungary) using as covariates mobility restrictions, internet traffic and sentiment indicators derived from social media. Using a multivariate AR model of lag-order 2, they find that the relationship between internet data and epidemiological indicators is bi-directional. In particular, the Twitter entry count has a positive sign on CoViD deaths (perhaps as a consequence of anti-mask initiatives) and was stronger during the first pandemic wave and Spain. Instead, a minor role is played by authority restrictions and people's sentiments.

Habib et al. (2021) investigate the relationship between CoViD infections and the mobility of people in 10 countries, in the period Mar-Jul 2020. A synthetic transportation index is obtained from the principal component analysis of 3 main series: private driving, public transport and people walking. Their modelling is nonlinear and is applied to 19 quantiles of the CoViD and transport series of each country; a positive bi-directional association is detected in nearly all quantiles, with the exception of France (see their Table 2).

Mastakouri and Schölkopf (2020) investigate the effect/efficacy of restriction policies by health authorities on the spread of CoViD pandemic in Germany. They consider the number of reported infections in the period Jan-May 2020 of 12 federal states and build 9 indicators of public restrictions (such as closure of activities, ban of gatherings, etc.) which are defined as binary step-variables. The modeling is done for each state by considering the dependence on neighbors and using a graphical algorithm to detect direct causal factors, avoiding confounding ones. The main findings are that regional interactions have a major role in explaining the pandemic spread, whereas the effect of internal

restrictions is minor and varies at the local level. However, they claim that states that timely close schools, less influence the others.

Li et al. (2021) study the effects of 19 social-economic variables on the CoViD pandemic in 50 US states in the period Mar-Aug 2020. Applying a bivariate AR model to each state and to each variable, they find that the most significant risk factors for CoViD deaths are: active cases, unemployment rate, imported CoViD cases, unemployment claims, tests done, testing capacity, percentage of change in consumption, percentage of working from home, and 6 mobility indexes (for transit stations, workplaces, residential places, retail and grocery shops). These factors act together in 17 states but in the 10, most populated states (as FL), they are *not* significant. Further, since mobility variables are related to authority restrictions, they claim that social distancing has a significant role in mitigating both infections and deaths.

Sato et a. (2021) analyze the relationship between Google search trends and CoViD infection series in 9 countries of 5 continents on the weekly data in the period Jan-Oct 2020. Google trends in relative search volume regard 54 keywords on various CoViD symptoms; using a bivariate AR model with 4-8 lags, they find that the number of keywords which have a significant Granger causality with CoViD trend infections ranges from 4 (of India) to 29 (of USA). Instead, classical correlation tests indicate a minimum of 21 (for Singapore) to 53 (India) keywords; this means that Pearson and Spearman tests (which do not take into account the serial correlation), may be biased upward and predicting pandemic trends on the basis of Google search trends may be wrong. Among the most frequent Granger-significant keywords (which are present in at least 5 countries), there are classical CoViD symptoms, such as loss of smell and taste, cough and sore throat, etc..

Finally, García-Cremades et al. (2021) consider the CoViD 14-day cumulative incidence (CI) data of Spain in the period Jul. 2020 to Jan. 2021 and aim to find the best out-of-sample predictive model among various alternatives. These include regression models having as inputs the Google mobility data (GMD, with respect to retail and recreation, parks, stations, supermarkets, workplaces and residential), on which Granger causality tests are also conducted. They show that the prediction performance of AR-based models outperforms the others (in particular that of neural networks) and a significant Granger causality is detected only for recreation, parks and station mobility indices.

References

- - Belosi, F., Conte, M., Gianelle, V., Santachiara, G., Contini, D., 2021. On the concentration of SARS-CoV-2 in outdoor air and the interaction with pre-existing atmospheric particles. Environ. Res. 193. doi:10.1016/j.envres.2020.110603.

Borro, M., Di Girolamo, P., Gentile, G., De Luca, O., Preissner, R., Marcolongo, A., Ferracuti, S., Simmaco, M., 2020. Evidence-based considerations exploring relations between SARS-CoV-2 pandemic and air pollution: involvement of $PM_{2.5}$-mediated upregulation of the viral receptor ACE-2. Int. J. Environ. Res. Public Health 17 (15). doi:10.3390/ijerph17155573.

Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control, 2nd ed. Holden Day, San FranciscoCA.

Chudik, A., Pesaran, M.H., 2015. Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. J. Econom. 188, 393–420. doi:10.1016/j.jeconom.2015.03.007.

Copiello, S., Grillenzoni, C., 2020. The spread of 2019-nCoV in China was primarily driven by population density. Comment on “Association between short-term exposure to air pollution and CoViD-19 infection: evidence from China” by Zhu et al. Sci. Total Environ. 744, 141028. doi:10.1016/j.scitotenv.2020.141028.

D’Albis, H., Coulibaly, D., Roumagnac, A., De Carvalho Filho, E., Bertrand, R., 2021. Quantification of the effects of climatic conditions on French hospital admissions and deaths induced by SARS-CoV-2. Sci. Rep. 11, 21812. doi:10.1038/s41598-021-01392-2.

Delnevo, G., Mirri, S., Roccetti, M., 2020. Particulate matter and CoViD-19 disease diffusion in Emilia-Romagna (Italy): already a cold case? Computation 8 (2), 59. doi:10.3390/computation8020059.

Dumitrescu, E.I., Hurlin, C., 2012. Testing for Granger non-causality in heterogeneous panels. Econ. Model. 29, 1450–1460. doi:10.1016/j.econmod.2012.02.014.

García-Cremades, S., Morales-García, J., Hernández-Sanjaime, R., Martínez-España, R., Bueno-Crespo, A., Hernández-Orallo, E., López-Espín, J.J., Cecilia, J.M, 2021. Improving prediction of CoViD-19 evolution by fusing epidemiological and mobility data. Sci. Rep. 11, 15173. doi:10.1038/s41598-021-94696-2.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37 (3), 424–438. doi:10.2307/1912791.

Granger, C.W.J., 2012. Useful conclusions from surprising results. J. Econom. 169 (2), 142–146. doi:10.1016/j.jeconom.2012.01.031.

Grillenzoni, C., 1997. Recursive generalized M-estimators of system parameters. Technometrics 39 (2), 211–224. doi:10.1080/00401706.1997.10485086.

Grillenzoni, C., 1998. Forecasting unstable and nonstationary time series. Int. J. Forecast. 14 (4), 469–482. doi:10.1016/S0169-2070(98)00039-9.

Grillenzoni, C., Carraro, E., 2021. Sequential tests of causality between environmental time series: With application to the global warming theory. Environmetrics 32 (1). doi:10.1002/env.2646.

Grillenzoni C. (2021). Robust identification of large subset ARX systems. MATLAB Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/100104

Habib, Y., Xia, E., Hashmi, S.H., Fareed, Z., 2021. Non-linear spatial linkage between CoViD-19 pandemic and mobility in ten countries: a lesson for future wave. J. Infect. Public Health 14 (10), 1411–1426. doi:10.1016/j.jiph.2021.08.008.

Hou, C., Qin, Y., Wang, G., Liu, Q., Yang, X., Wang, H., 2021. Impact of a long-term air pollution exposure on the case fatality rate of CoViD-19 patients - a multicity study. J. Med. Virol. doi:10.1002/jmv.26807.

Huber, P.J., 1981. Robust Statistics. Wiley, New York doi:10.1002/0471725250.

Li, Z., Xu, T., Zhang, K, Deng, H.-W., Boerwinkle, E., Xiong, M., 2021. Causal analysis of health interventions and environments for influencing the spread of CoViD-19 in the United States of America. Front. Appl. Math. Stat. doi:10.3389/fams.2020.611805.

Lolli, S., Chen, Y.C., Wang, S.H., Vivone, G., 2020. Impact of meteorological conditions and air pollution on CoViD-19 pandemic transmission in Italy. Sci. Rep. 10, 16213. doi:10.1038/s41598-020-73197-8.

Magyar, M., Kovács, L., Burka, D., 2021. Forecasting the spread of the CoViD-19 pandemic based on the communication of coronavirus sceptics. Eng. Proc. 5, 35. doi:10.3390/engproc2021005035.

Martinez-Boubeta, C., Simeonidis, K., 2022. Airborne magnetic nanoparticles may contribute to CoViD-19 outbreak: relationships in Greece and Iran. Environ. Res. 204, 112054. doi:10.1016/j.envres.2021.112054, Part B.

Mastakouri, A.A., Schölkopf, B., 2020. Causal analysis of CoViD-19 spread in Germany. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS). Vancouver. CAN.

Sarkodie, S.A., Owusu, P.A., 2020. Impact of meteorological factors on CoViD-19 pandemic: evidence from top 20 countries with confirmed cases. Environ. Res. 191, 110101. doi:10.1016/j.envres.2020.110101.

Sato, K., Mano, T., Iwata, A., Toda, T., 2021. Need of care in interpreting Google trends-based CoViD-19 infodemiological study results: potential risk of false-positivity. BMC Med. Res. Methodol. 21, 147. doi:10.1186/s12874-021-01338-2.

Setti, L., Passarini, F., De Gennaro, G., et al., 2020. SARS-Cov-2 RNA found on particulate matter of Bergamo in northern Italy: first evidence. Environ. Res. 188. doi:10.1016/j.envres.2020.109754.

Sharma, G.D., Bansal, S., Yadav, A., Jain, M., Garg, I., 2021. Meteorological factors, CoViD-19 cases, and deaths in top 10 most affected countries: an econometric investigation. Environ. Sci. Pollut. Res. 28, 28624–28639. doi:10.1007/s11356-021-12668-5.

Sutherland E., Headicar J. & Delong P. (2021). Coronavirus (CoViD-19) infection survey technical article: waves and lags of CoViD-19 in England, Office for National Statistics, https://www.ons.gov.uk/

Tizaoui, C., Stanton, R., Statkute, E., Rubina, A., Lester-Card, E., Lewis, A., Holliman, P., Worsley, D., 2022. Ozone for SARS-CoV-2 inactivation on surfaces and in liquid cell culture media. J. Hazard. Mater. 428. doi:10.1016/j.jhazmat.2022.128251.

Wang, C.C., Prather, K.A., Sznitman, J., Jimenez, J.L., Lakdawala, S.S., Tufekci, Z., Marr, L.C., 2021. Airborne transmission of respiratory viruses. Science 373, 981. doi:10.1126/science.abd9149.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48 (4), 817–838. doi:10.2307/1912934.

Zheng, Z., Dou, J., Cheng, C., Gao, H., 2021. Correlation and causation analysis between CoViD-19 and environmental factors in China. Front. Clim. doi:10.3389/fclim.2021.619338.