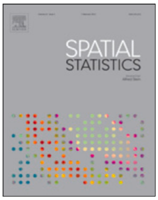
Spatial Statistics 49 (2022) 100549

Contents lists available at ScienceDirect

Spatial Statistics

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



Spatiotemporal variable selection and air quality impact assessment of COVID-19 lockdown

Alessandro Fassò [a,*], Paolo Maranzano [a], Philipp Otto [b]

[a] University of Bergamo, DSE, Via Caniana 2, Bergamo, 24127, BG, Italy
[b] Leibniz University Hannover, Appelstrasse 9a, Hannover, 30167, Lower Saxony, Germany

ARTICLE INFO

Article history:
Received 15 May 2021
Received in revised form 15 October 2021
Accepted 18 October 2021
Available online 29 October 2021

Keywords:
Multivariate spatiotemporal models
Model selection
Particulate matters
Nitrogen dioxide
COVID-19

ABSTRACT

During the first wave of the COVID-19 pandemics in 2020, lockdown policies reduced human mobility in many countries globally. This significantly reduces car traffic-related emissions. In this paper, we consider the impact of the Italian restrictions (lockdown) on the air quality in the Lombardy Region. In particular, we consider public data on concentrations of particulate matters (PM_{10} and $PM_{2.5}$) and nitrogen dioxide, pre/during/after lockdown. To reduce the effect of confounders, we use detailed regression function based on meteorological, land and calendar information. Spatial and temporal correlations are handled using a multivariate spatiotemporal model in the class of hidden dynamic geostatistical models (HDGM). Due to the large size of the design matrix, variable selection is made using a hybrid approach coupling the well known LASSO algorithm with the cross-validation performance of HDGM.

The impact of COVID-19 lockdown is heterogeneous in the region. Indeed, there is high statistical evidence of nitrogen dioxide concentration reductions in metropolitan areas and near trafficked roads where also PM_{10} concentration is reduced. However, rural, industrial, and mountain areas do not show significant reductions. Also, $PM_{2.5}$ concentrations lack significant reductions irrespective of zone. The post-lockdown restart shows unclear results.

[*] Corresponding author.
E-mail address: alessandro.fasso@unibg.it (A. Fassò).

https://doi.org/10.1016/j.spasta.2021.100549
2211-6753/© 2021 Elsevier B.V. All rights reserved.

1. Introduction

The first western country hit by COVID-19 pandemics in early 2020 was Italy and, in particular, the Lombardy Region. During the first pandemic wave the Italian government enforced draconian lockdowns. Hence, between 9th March and 18th May 2020, people and car circulation reduced dramatically (Finazzi and Fassò, 2020). Since vehicular traffic is one of the major air pollutant emission sources in Lombardy, it is not surprising that Collivignarelli et al. (2020) observed an air quality (AQ) improvement in the metropolitan area of Milan, and Cameletti (2020) in the city of Brescia.

In fact, a generalized AQ improvement has been observed worldwide. For example, the COVID-19 impact on AQ at the European level is considered by the European Environmental Agency (2020), and a thorough statistical analysis of the corresponding impact in the North of China is developed in Zheng et al. (2021). Some studies have used time series analysis techniques at the station level, e.g. Cameletti (2020), others aggregated the results according to province partitioning or land use (Maranzano and Fassò, 2021).

Moreover, many studies have addressed AQ spatiotemporal modelling issues before COVID-19 pandemics. Using an idea popular in the regression Gaussian Process approach of Machine Learning (Rasmussen and Williams, 2006; Beauchamp et al., 2017) used an enlarged spatial dimension $2D \times V$ to handle AQ covariance non-stationarity, and Guan et al. (2020) used the same idea for non-stationary spatiotemporal modelling. Wan et al. (2020) provided insights into model seasonality. Calculli et al. (2015) used a multivariate approach to merge a multipollutant AQ network and a non co-located meterological network; Wang et al. (2021) considered a spatiotemporal functional data approach for hourly ozone data.

The idea that the COVID-19 impact on AQ is related to spatial variability and spatial correlation is consistent with the marked spatiotemporal dynamics of the pandemics itself (Jalilian and Mateu, 2021), and the heterogeneity of people mobility during the lockdown (Finazzi and Fassò, 2020).

In this paper, we use multivariate statistical spatiotemporal modelling to understand the COVID-19 joint impact on the concentrations of fine and ultrafine particulate matters (PM_{10} and $PM_{2.5}$) and nitrogen dioxide (NO_2). In particular, we consider the AQ variations during the lockdown period, which started on 9th March 2020, and during the mobility "restart", beginning on 19th May. We describe the spatiotemporal regression function in detail introducing several meteorological, calendar, and land use covariates. The spatiotemporal correlation is covered by the hidden dynamic geostatistical model (HDGM) of Calculli et al. (2015). Mean non-stationarity in space and in time is covered by interacting the covariates with seasons, regional zoning, and local conditions. Similarly the lockdown/restart impact is detailed for zoning and local conditions. It results in a design matrix with about three hundred columns, and a variable selection issue arises.

Considering variable selection, Hoeting et al. (2006) discussed the Akaike information criterion (AIC) and the minimum description length (MDL) criterion for geostatistical models. Moreover, Shen et al. (2021) proposed the Spatial information criterion (SIC).

To avoid the course of dimensionality arising by subset regression implied by the above criteria, we consider regularization techniques. In the linear regression context with numerous covariates, the approach based on penalized least squares known as LASSO (Tibshirani, 1996) is well known and computationally efficient. Its standard implementation is based on k-fold cross-validation to select the regularization coefficient, which gives the best forecast performance in the validation dataset. A more general regularization approach is based on the penalized maximum likelihood estimation (PMLE) (Fan and Li, 2001). Chu et al. (2011) adapted PMLE to geostatistical models.

While research is ongoing to extend PMLE to multivariate spatiotemporal models such as HDGM, in this paper, we use a hybrid approach based on merging the standard LASSO method with the HDGM forecasting performance.

The remaining part of the paper is organized as follows. Section 2 introduces the data sources and gives a preliminary analysis of the impact of lockdown and restart on AQ. Section 4, after recalling the essential elements of the HDGM model and its estimation, explains the hybrid algorithm used for variable selection. Section 5 describes the model selection and estimation process and results for the AQ data. Section 6 discusses the model-based assessment of lockdown and restart for the various zones and local conditions implemented in the model. Finally, Section 7 concludes the paper.

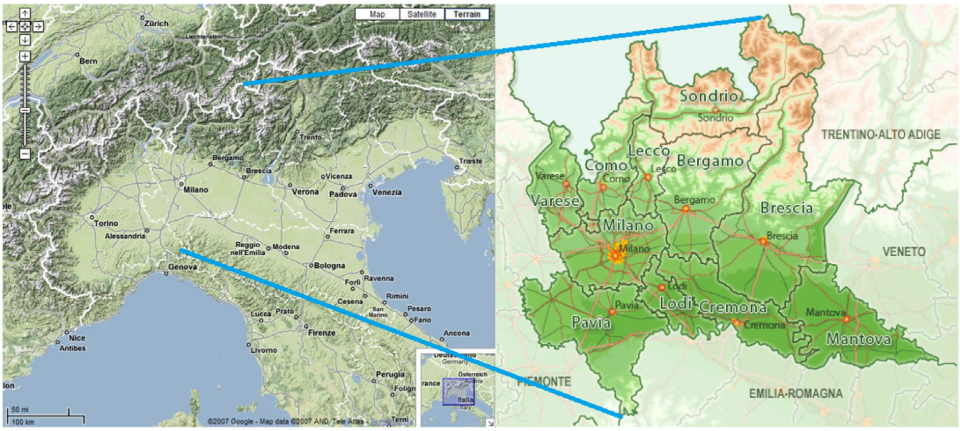


Fig. 1. Left panel: the North of Italy map highlighting the C-shaped mountain range that encloses the Po Valley and prevents Oceanic air circulation. In the map middle, the Lombardy Region and the cities of Milano, Bergamo, and Brescia. Right panel: Lombardy Region details highlighting the province borders and names, the northern Alpine area, the central Po Plain, and the southern Apennine area.

2. Environmental conditions and data

As shown in Fig. 1, the climate of Lombardy Region is conditioned by the C-shaped mountain range, which deviates the air masses coming from the Atlantic Ocean and facilitates air stagnation in the Po Plain. The region has three main zones: the Alpine range in the North, which is sparsely populated; the sloping foothills in the middle, which is densely urbanized and industrialized, and comprises the metropolitan areas of Milan, Bergamo, and Brescia; and the southern rural area, which is less densely populated and oriented toward agriculture and farming.

Therefore, it is not surprising that Lombardy is one of the most air-polluted regions in Europe (European Environmental Agency, 2019). Interestingly, Raffaelli et al. (2020) showed that, if Po Valley had the same meteorological conditions typical of central-northern Europe and the observed emission levels of the year 2013, the average monthly concentrations of PM_{10} and NO_2 would lower by 60 to 80%.

2.1. Air quality data

We use daily air quality data on Lombardy Region obtained from the public repository of the regional agency for environmental protection, namely ARPA Lombardia. We focus on the monitoring network depicted in Fig. 3, comprising those 84 control sites that monitor at least one of the three pollutants considered in this paper: nitrogen dioxide (NO_2), particulate matters with a maximum diameter of 10 μm (PM_{10}) and particulate matters with a maximum diameter of 2.5 μm ($PM_{2.5}$). The study period runs from January 2017 to October 2020, which covers a pre-pandemic period appropriate for model estimation, the first COVID-19 wave, and the restart period.

All the 84 monitoring sites include the sensor for the NO_2, while just 62 and 31 sites monitor PM_{10} and $PM_{2.5}$, respectively. Among the stations collecting measurements on $PM_{2.5}$, 30 also collect information on PM_{10}, while there exists a single station that controls $PM_{2.5}$, but not PM_{10}. Hence, we have a heterogeneous network and need a multivariate statistical approach able to handle it.

2.2. Land use

Following the above discussion about the region structure, we aim at understanding if the lockdown effects are homogeneous in the region or are correlated to some important geographical

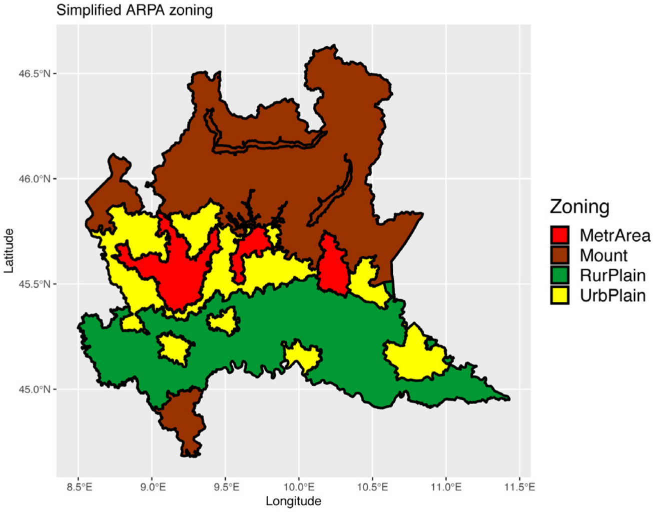


Fig. 2. Lombardy Region with simplified ARPA classification called zoning. The three orange zones (MetrArea) correspond to Milan, Bergamo, and Brescia metropolitan areas. The brown zone (Mount) identifies alpine areas. The green zone (RurPlain) corresponds to the rural areas near the Po river. Eventually, the yellow zone (UrbPlain) corresponds to low density urban areas in the Po plain. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Simplified ARPA classifications of ARPA Lombardia AQ network with the number of monitoring sites.

Site type			Zoning		
Local conditions		# Sites	Area type		# Sites
Background	B	42	Mountain	Mount	10
Industrial	I	6	Urbanized plain	UrbPlain	27
Rural	R	11	Rural plain	RurPlain	18
Traffic	T	25	Metropolitan area	MetrArea	29

structures. To do so, we employ two different territorial classification schemes provided by ARPA Lombardia, both related to land use and considering a mix of pollutant emissions, concentration patterns, and geography. The first taxonomy, called zoning, is a large scale classification, while the second one, called station type, considers local conditions.

For zoning, we use the simplified 4-class taxonomy depicted in Fig. 2 and in the right hand panel of Fig. 3. It considers metropolitan areas (MetrArea), urban plain (UrbPlain), rural plain (RurPlain), and mountain area (Mount) areas. Notably, MetrArea includes the three metropolitan areas of Milan, Bergamo, and Brescia. In addition, UrbPlain includes less dense urbanized areas. RurPlain is characterized by small towns surrounded by agricultural and rural land with a rather low population density. Finally, Mount coincides with the pre-Alps, Alps, and Appenine regions.

Station type classification is related to local conditions near the monitoring site. We use the simplified 4-class taxonomy depicted in the left hand panel of Fig. 3. It considers background sites (B), which are representative of the average population and vegetation exposure; traffic sites (T), which are representative of the local AQ pattern close to trafficked roads; industrial sites (I), which are close to industrial areas; and rural sites (R), which are characterized by sparsely populated areas in a rural context. These two taxonomies and the related number of monitoring sites are reported in Table 1.

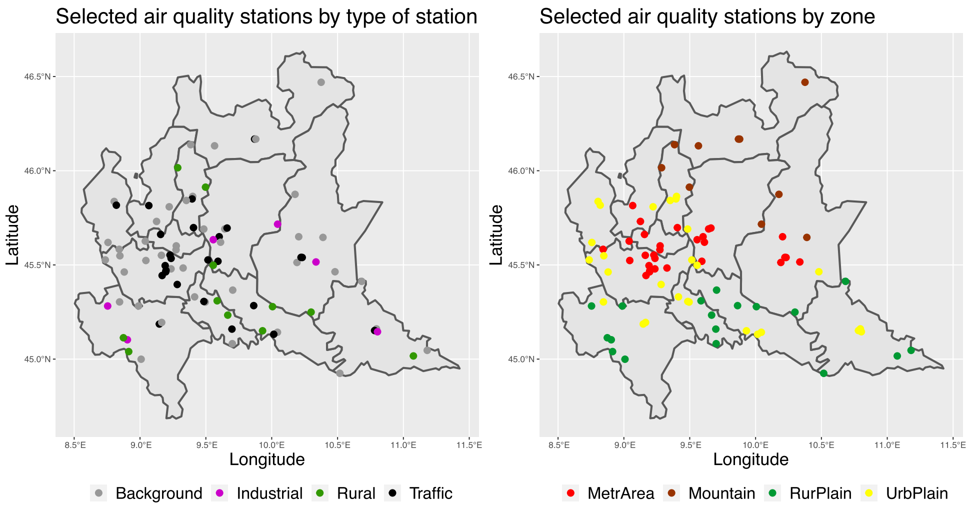

Selected air quality stations by type of station
Selected air quality stations by zone
46.5°N
46.0°N
45.5°N
45.0°N
Latitude
8.5°E
9.0°E
9.5°E
10.0°E
10.5°E
11.0°E
11.5°E
Longitude
● Background ● Industrial ● Rural ● Traffic
● MetrArea ● Mountain ● RurPlain ● UrbPlain

Fig. 3. Lombardy Region with province borders and AQ network classified by station type (left panel) and zoning (right panel).

From the left panel of Fig. 3, one can notice that the background stations (grey points) are spread over the provinces of Lombardy, while the rural sites are mainly placed in the southern territories, which is consistent with the zoning in Fig. 2. Traffic sites mostly coincide with the main urban centres of the region.

2.3. Meteorology and land data

To obtain a reasonable and accurate estimate of the effect of lockdown on air quality in Lombardy, we aim at controlling AQ variation due to meteorology and other confounders. In particular, we use weather and land morphology data from the ERA5-Land database provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) in the frame of the Copernicus program. Notice that ERA5-land products are obtained by atmospheric models that assimilate empirical observations from across the world into a globally complete and consistent dataset using the laws of physics. ERA5-Land offers gridded data at a 0.1° × 0.1° spatial resolution, with hourly temporal resolution from 1981 to present. The regular spatial and temporal resolution of the dataset allows for mapping the output of the statistical model at unmonitored sites.

We consider the following set of weather measurements: temperature [°C] and temperature dew point [°C] at 2 metres from the ground, the eastward and northward components of the wind at 10 metres altitude [m/s], surface pressure [hPa], and total precipitation [mm]. Using the temperature (T) and the dew point temperature (T_d), we compute the relative humidity (RH) using the well known approximation:

$$RH = 100 \times e^{\frac{17.625 \cdot T_d}{243.04 + T_d} - \frac{17.625 \cdot T}{243.04 + T}}, \tag{1}$$

where 17.625 and 243.04 are the physical transformation constants used by the August-Roche-Magnus approximation formula for computing the saturation vapour pressure (Alduchov and Eskridge, 1996). Thus, relative humidity depends on the difference between the dewpoint-induced saturation and the saturation generated by atmospheric temperature.

In addition to the atmospheric measurements, we consider the effects of soil type and orography on air quality. In particular, we consider the leaf area index for both high vegetation and low vegetation [m^2/m^2] and the geopotential height [m^2/s^2]. The leaf area indices are measured as one-half of the total green leaf area per unit horizontal ground surface area for vegetation type, whereas total precipitation includes the accumulated liquid and frozen water, including rain and snow, which fall to the Earth's surface. The geopotential height measures the gravitational potential energy of a unit mass at a particular location relative to mean sea level. At the surface of the Earth, this parameter shows the variations in geopotential (height) of the surface and is often referred to as the orography.

3. Preliminary policy assessment

Following the rapid spread of the COVID-19 virus during the first two months of 2020, especially in northern and central Italy, the Italian government ordered a total national lockdown from 9th March to 18th May, totalling 71 days. The lockdown imposed many restrictions, the more important being the use of masks everywhere outside private households, the closure of every food-service activity (restaurants, cafes, and pubs), the substitution of learning activities at schools and universities with 100% e-learning activities, forbidding public events and stopping private visits in nursing homes. All movements were forbidden except for work, health, or other urgent reasons, and all the commercial activities were suspended, except those selling foodstuffs and basic necessities. Markets were also closed, except for stands exclusively selling food. Newsstands, tobacco shops, and pharmacies remained open. Both public and private workplaces were strongly incentivized to postpone all frontal activities or rather to remote their operations and avoid personal contacts.

As a direct consequence of these restrictions, there has been a huge drop in mobility movement, consumption (especially fuel consumption), and economic output throughout Italy. Using a sample of 20 thousands Italian users of an earthquake-tracker smartphone application, Finazzi and Fassò (2020) estimated that, at the peak of the lockdown period, the daily mean distance travelled by users decreased by approximately 50%, and the percentage of users who did not move within 24 h reached 65%.

At the end of the lockdown period, the government gradually reduced the restrictions, leading to a situation of free movement within the whole national territory from 3rd June and lasting until the end of October. This period of low restrictions and free movement is addressed here as the restart period. Although, new closures and severe restriction policies were imposed at the regional level in November, we do not consider here the second and third COVID-19 waves.

3.1. Preliminary data analysis

As a preliminary analysis of lockdown impact, we consider the difference of the pollutant levels averaged during the lockdown period (9th March–18th May) of 2020 and the average in the same period of years 2017–2019. Similarly, for the restart impact, we consider the average differences between the restart period (19th May–31th October) of the year 2020 and the average in the same period of years 2017–2019.

The first column of Fig. 4 depicts the lockdown variations by station type: NO_2 generally shows the largest decreases, with a maximum of 15 $\mu g/m^3$ in traffic stations and a smaller or zero reduction in industrial and rural sites. Moreover, PM_{10} and $PM_{2.5}$ show smaller reductions also with a maximum in traffic stations. The first column of Fig. 5 depicts the lockdown variations by station type and shows that the largest reductions are in metropolitan areas for all three pollutants.

The restart impact is shown in the second column of Figs. 4 and 5. Overall, we have negative differences, meaning that the pre-pandemics pollution levels are not attained despite the restart. This may be due to various factors: a persisting reduction in mobility related to distance working, meterological changes for year 2020 compared to the previous triennium, and an overall decreasing trend of pollutant concentrations (Maranzano and Fassò, 2021).

In conclusion, the above results may be biased by various factors, including spatial and temporal correlation and trends, network heterogeneity, and meteorological confounders. Hence, in Section 5 we build a statistical model able to handle all the above factors and confounders, including two

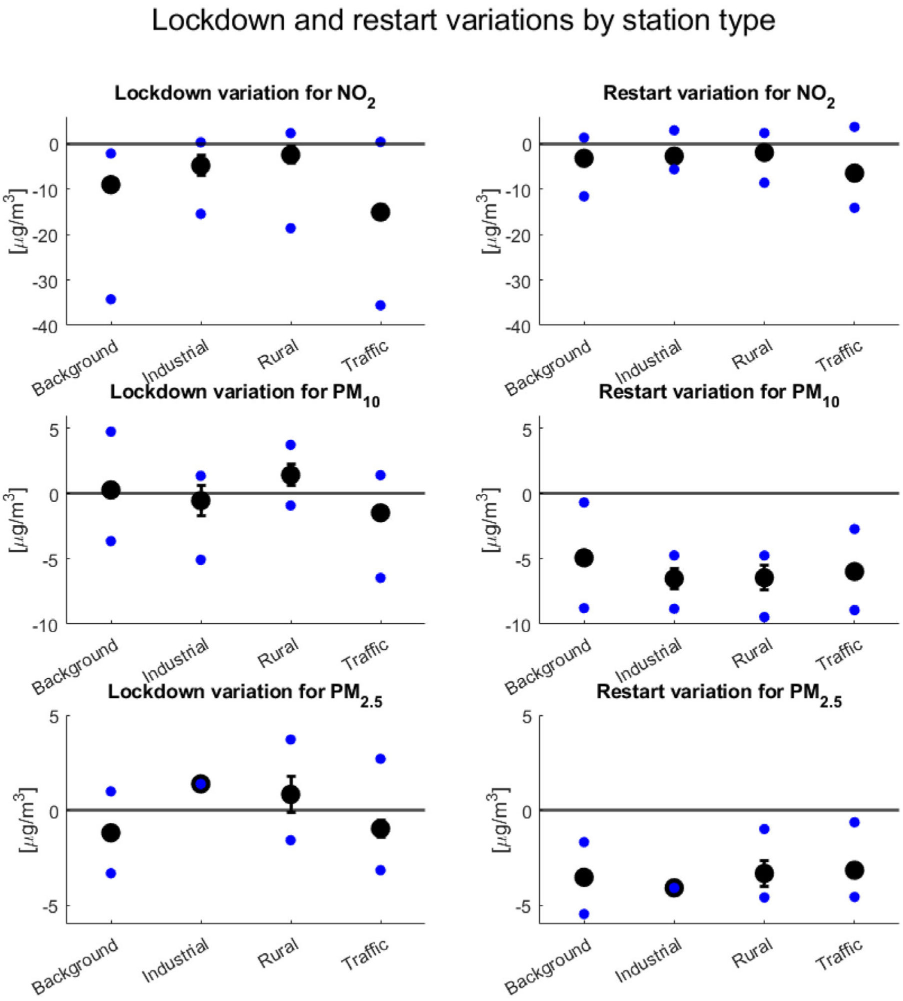


Fig. 4. Average variations between year 2020 and triennium 2017–2019 by station type. First column: NO$_2$; second column: PM$_{10}$; third column: PM$_{2.5}$. First row: lockdown period (9th March–18th May); second row: restart period (19th May–31th October). Black points represent the average variation, blue points are the minimum and the maximum variations, while the error bars represent ±1 standard error of the mean, computed under random sampling assumptions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

dummy variables, one the lockdown and one for the restart. Using this in Section 6, we perform a model-based impact assessment.

4. Variable selection for multivariate spatiotemporal models

In this section, we introduce a model selection technique suitable for multivariate spatiotemporal models. In particular, we focus on the multivariate hidden dynamic geostatistical models (HDGMs) introduced by Calculli et al. (2015), which are described subsequently. Regarding many potential covariates, the selection of the most suitable regressors is computationally challenging. Therefore, we show how to employ a hybrid penalized approach to choose the best models. In Section 4.2, this model selection technique is explained in detail.

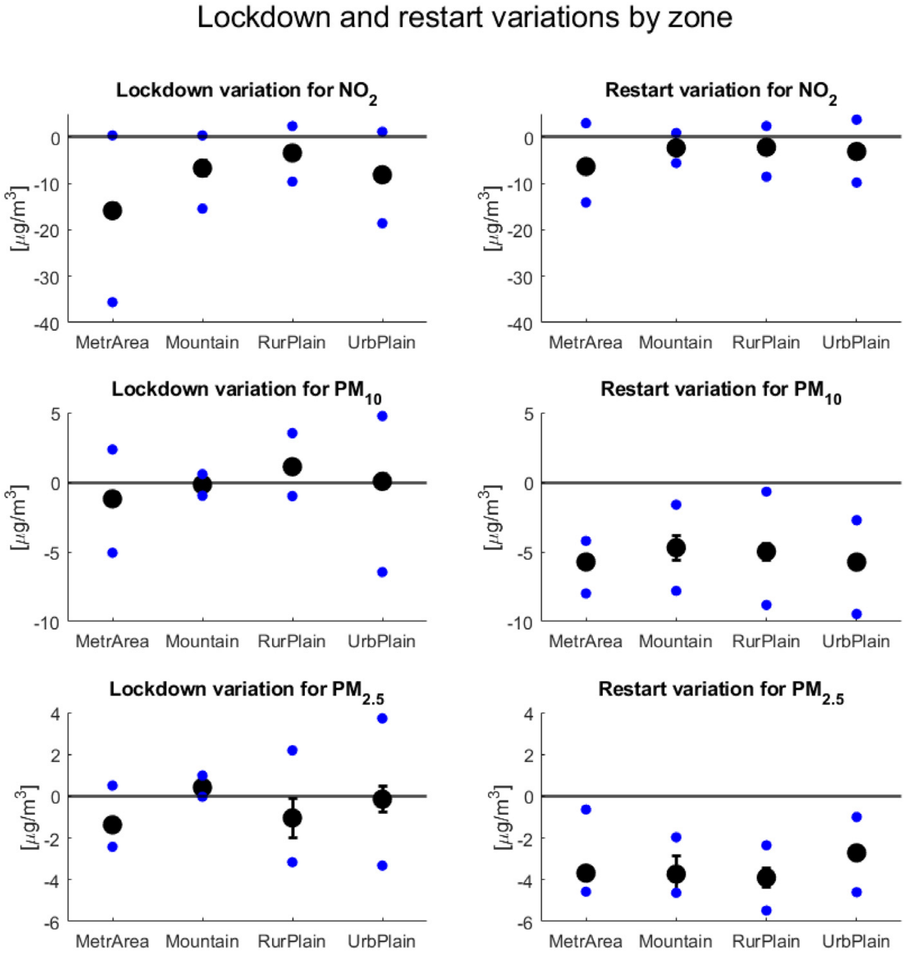


Fig. 5. Average variations between year 2020 and triennium 2017–2019 by zone. First column: NO$_2$; second column: PM$_{10}$; third column: PM$_{2.5}$. First row: lockdown period (9th March–18th May); second row: restart period (19th May–31th October). Black points represent the average variation, blue points are the minimum and the maximum variations, while the error bars represent ±1 standard error of the mean, computed under random sampling assumptions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.1. Hidden dynamic geostatistical model

We consider that the data is driven by q-variate spatiotemporal process $\{Y(s,t) \in \mathbb{R}^q : s \in D, t = 1, \ldots, T\}$, where D is the spatial domain and t represents a discrete point of time. We refer to each component of the multivariate response as $Y_i(s,t)$ with $i = 1, \ldots, q$. More precisely, a HDGM is used, in which a random effects term $w(s,t)$ covers all spatial and temporal dependence, while the fixed effects term $v(s,t)$ accounts for all exogenous, regressive effects. That is,

$$Y(s,t) = v(s,t) + w(s,t) + \varepsilon(s,t) \tag{2}$$

with $\varepsilon(s,t)$ being q-dimensional error vectors that are assumed to be independent and identically distributed across space and time with mean zero and a constant, diagonal covariance matrix

$\Sigma_\varepsilon = diag(\sigma_1, \dots, \sigma_q)$. More specifically, the fixed-effects mean term can be specified as

$$v(s,t) = diag(X_1(s,t), \dots, X_k(s,t))\beta \,, \tag{3}$$

$X_i(s,t)$ are $(n \times p_i)$-dimensional matrices of regressors for each component $i = 1, \dots, q$, location s and time point t, so that $diag(\cdot)$ forms a block diagonal matrix. Each of these so-called design matrices contains a column of ones representing the component-specific intercept. Moreover, β is a $(p_1 + \cdots + p_q)$-dimensional vector of regression coefficients, with dimensions ps allowed to be very large.

This random effects term $w(s,t)$ accounts for the spatiotemporal dependence in the random process $Y(s,t)$. More precisely,

$$w(s,t) = Gw(t-1,s) + \omega(s,t)$$

and $\omega(s,t) = (\omega_1(s,t), \dots, \omega_q(s,t))'$ is a q-dimensional Gaussian random variable with zero mean and covariance matrix

$$\Gamma_i = V\rho(\|s - s'\|, \theta_i, \nu)$$

for each variable i, where V is a correlation matrix and ρ is a Matern covariance function with shape parameter ν. Notably, this spatial dependence might differ for each variable because the θ_i can vary. The temporal dynamics are covered by the Markovian process with a matrix $G = (g_{ij})_{i,j=1,\dots,q}$ for all potential temporal dependencies and interactions. In the application of Sections 5 and 6, we will use a simplified HDGM with a diagonal persistence matrix G and a constant spatial correlation range θ. The former assumption is not serious because the cross-correlation among the response components is handled by the matrix V. The assumption of constant θ may be more critical, but in our application, we will get a very good fitting for all response variables. Moreover, let ξ be a vector of all coefficients of the random-effects term, i.e., $\xi = (g_{1,1}, \dots, g_{q,q}, v_{1,1}, \dots, v_{q,q}, \theta_1, \dots, \theta_q, \nu)'$.

The maximum likelihood estimates of ξ may be computed using the EM algorithm, which is implemented together with the parameter variance–covariance matrix computation in D-STEM software (Finazzi and Fassò, 2014; Wang et al., 2021).

4.2. Hybrid variable selection method

For the multivariate setting, the number of regressive parameters can become very large, and the selection of the covariates may be computationally demanding. Thus, we use a hybrid two-step approach.

At the first step, we use a simple penalized regression approach to select the regressors of the fixed-effects term $v(s,t)$ for each fixed penalty parameter λ. More precisely, we use the least absolute shrinkage and selection operator (LASSO) for each component of the multivariate response to select relevant regressors. That is, we exploit the ability of the LASSO estimator to shrink irrelevant coefficients to zero. However, this step is only applied to the regression coefficients β, while the spatial, temporal dependence structure and the error variance is ignored in the first step (Tibshirani, 1996).

At the second step, using the selected βs, we estimate the HDGM using the classical maximum likelihood approach. Eventually, goodness-of-fit measures of the HDGM are computed for each λ to select the best penalty parameter by cross-validation.

In more detail, let $\mathcal{D}_0 = \{y(s,t)\}$ be the set of all observations at all time points t and locations s, let $\vartheta = (\xi, \sigma_1, \dots, \sigma_q)'$ be the parameters of the random-effects term and errors, and let β be the vector of regression parameters. Furthermore, let $\mathcal{D}_1$ be a partitioning of data $\mathcal{D}_0$ used for selecting the penalty parameter in cross-validation as discussed below. Moreover, let $\mathcal{A}_0 = \{j : \beta_j^{TRUE} \neq 0\}$ denote the "true" active set, i.e., the set of all regressors having an influence on the dependent variable in model (2). Further, a LASSO approach is applied to estimate the vector β as in (3) and simultaneously shrink irrelevant regressors to zero. We initially neglect any correlation between the response variables and apply separated LASSO to each of them:

$$\beta_i^{(LASSO)}(\lambda) = \arg\min_{\beta_i} \|Y_i(s,t) - X_i(s,t)\beta_i\|_2^2 + \lambda\|\beta_i\|_1 \,, \text{ for } i = 1, \dots, q \,.$$

For simplicity, we use a constant penalty parameter λ for all components Y_i here. The generalization to q different penalties is conceptually straightforward but could be computationally demanding for large q.

When the cross-correlation is positive as in our case, the separated LASSO solution is shrinking the parameters less compared to joint scheme accounting for cross-correlation between the response variables (Lee and Liu, 2012). Since this approach is used for selecting the covariates, applying a selection-consistent approach is imperative. That is, the probability that the correct coefficients are shrunk to zero should tend to one, so that only the coefficients of the "true" active set $\mathcal{A}_0$ remain. For the considered spatiotemporal data, it is very likely that the covariates are correlated, so that classical LASSO procedures are not necessarily selection-consistent. Conditions for selection consistency of the general LASSO are extensively discussed in Zhao and Yu (2006) among others. For multivariate models Lee and Liu (2012) showed that a joint scheme accounting for cross-correlation between the response variables is selection-consistent. In our case, an adaptive LASSO procedure as proposed by Zou (2006) and Zou and Li (2008) can be applied to obtain selection-consistent estimates of the regressive parameters, which will be the subject of a future paper.

In this paper, however, we initially use the simpler approach applying a separated LASSO approach independently for each component and selecting λ, such that the out-of-sample fit of the HDGM is maximized. Thus, the penalized coefficients $\beta_i^{(LASSO)}(\lambda)$ are estimated for λ in a sequence $\{0, \lambda_{min}, \ldots, \lambda_{max}\}$ and $i = 1, \ldots, q$ to obtain an active set for each choice of λ. We denote this active set by $\mathcal{A}_\lambda = \{(i, j) : \beta_{i,j}^{(LASSO)}(\lambda) \neq 0\}$, where $\beta_{i,j}^{(LASSO)}(\lambda)$ is the jth element of $\beta_i^{(LASSO)}(\lambda)$.

Then, the HDGM given by (2) is estimated by the EM algorithm with only the regressors of the active set $\mathcal{A}_\lambda$. Thus, we obtain consistent estimates of all parameters. Since all the coefficients depend on the active set implied by the penalty parameter λ, we denote these estimates by $(\hat{\beta}_\lambda, \hat{\vartheta}_\lambda)'$. This estimation is repeated for each cross-validation partition in $\mathcal{D}_1$ to obtain out-of-sample goodness-of-fit measures as described in Section 4.3. Eventually, an optimal λ^* is chosen according to this goodness-of-fit of the spatiotemporal model.

At the end of the variable selection iterations, the spatiotemporal model is re-estimated to get the final parameters $(\hat{\beta}_{\lambda^*}, \hat{\vartheta}_{\lambda^*})'$ and their variance–covariance matrix. Following Belloni and Chernozhukov (2013) and Chzhen et al. (2019), since these final parameters are not resulting from a penalized regression, all results from Calculli et al. (2015) are valid, allowing for classical statistical inference on these parameters. The computation of the covariance matrix of the regressors can be computationally demanding, especially for a large number of locations and time points, e.g. Piter et al. (2020), and should therefore be avoided as far as possible. For the suggested model selection procedure, this only has to be done once in the final step. Hence, it is computationally very efficient.

Besides, it is important to note that some values of λ can lead to the same out-of-sample fit since the final estimates $(\hat{\beta}_\lambda, \hat{\vartheta}_\lambda)'$ only depend on the active set $\mathcal{A}_\lambda$ but not on λ directly. To be precise, if $\mathcal{A}_{\lambda_1} = \mathcal{A}_{\lambda_2}$ for any pair $\lambda_1 \neq \lambda_2$, we get the same estimates $(\hat{\beta}_{\lambda_1}, \hat{\vartheta}_{\lambda_1})' = (\hat{\beta}_{\lambda_2}, \hat{\vartheta}_{\lambda_2})'$ of the spatiotemporal model. That is, the penalty parameter λ is not uniquely defined but only the active set $\mathcal{A}_\lambda$. The Algorithm 1 summarizes the above procedure.

4.3. Spatiotemporal partitioning and cross-validation

Feature selection is performed evaluating the out-of-sample forecasting performances of each spatiotemporal model associated to the active sets provided by the separated LASSO estimates. We consider a cross-validation setup using the standard spatiotemporal random k-fold scheme: the full dataset $\mathcal{D}_o$ is split into k equally-sized folds, $\mathcal{D}_1(j),\ j = 1, \ldots, k$ by randomly assigning each observation to a specific group.

Model performances are compared according to an L_1 metric, i.e. the mean absolute error (MAE), and an L_2 metric, i.e. the root mean squared error (RMSE). To see this, let $Y_{f,i}(s, t),\ i = 1, \ldots, q$ denote the elements of $\mathcal{D}_1(f)$, i.e. the set of observations associated to the fth fold and having length n_f, which is assumed not to depend on i for exposition simplicity. The remaining $n - n_f$ observations are used for model estimation and to predict the out-of-sample observations. These

Algorithm 1 Hybrid LASSO-HDGM procedure for variable selection of multivariate spatiotemporal models

```
Set the initial active set $\mathcal{A}$ to all regressors
Define the k-fold data partitioning $\mathcal{D}_1(1), ..., \mathcal{D}_1(k) \subset \mathcal{D}_0$
for $\lambda = 0, \lambda_{min}, \dots, \lambda_{max}$ do
    Estimate $\beta^{(LASSO)}(\lambda)$
    Define the active set $\mathcal{A}_\lambda = \{(i,j) : \beta^{(LASSO)}_{i,j}(\lambda) \neq 0\}$
    if $\mathcal{A}_{\geq} \neq \mathcal{A}_{\bar{\lambda}}$ for $\bar{\lambda} < \lambda$ or $\lambda = 0$ then
        for fold $f = 1, ..., k$ do
            Estimate $(\hat{\beta}_\lambda, \hat{\vartheta}_\lambda)$ using HDGM and data in $\mathcal{D}_0 \setminus \mathcal{D}_1(f)$
            Predict the observations in $\mathcal{D}_1(f)$ using $HDGM(\hat{\beta}_\lambda, \hat{\vartheta}_\lambda)$
            Compute $RMSE_{f,i}(\lambda),\ i = 1, ..., q$
        end for
    end if
end for
Select $\lambda^*$ such that the average $RMSE(\lambda)$ is minimal
Re-estimate the HDGM with $\mathcal{A}_{\lambda^*}$ to obtain the final parameters $(\hat{\beta}_{\lambda^*}, \hat{\vartheta}_{\lambda^*})'$
Compute the covariance matrix of the estimates $(\hat{\beta}_{\lambda^*}, \hat{\vartheta}_{\lambda^*})'$
```

predictions are denoted by $\hat{Y_{f,i}}(s,t)$ with $f = 1, \dots, k$. The two above metrics are then computed comparing observed and predicted values as follows:

$$RMSE^2_{f,i} = \frac{\sum_{n_f} \left(\hat{Y_{f,i}}(s,t) - Y_{f,i}(s,t)\right)^2}{n_f}$$

and

$$MAE_{f,i} = \frac{\sum_{n_f} \left|\hat{Y_{f,i}}(s,t) - Y_{f,i}(s,t)\right|}{n_f}.$$

Eventually, these metrics are averaged across the k folds and the q responses. As well argued by Roberts et al. (2017), ignoring the dependence structure of the data when designing the cross-validation scheme generates two undesirable effects: first, the underestimation of the prediction error and, second, the potential overfitting for non-causal predictors. In fact, The central assumption for an appropriate cross-validation is that training and evaluation data are independent. To overcome these issues, several solutions have been proposed depending on the type of data structure. If both spatial and temporal dimensions are considered, one can choose among several target-oriented cross-validation strategies (Meyer et al., 2018). Example of such spatiotemporal cross-validation schemes are the leave-location-out (i.e., single locations or groups are excluded from the training set and used to validate the model) (Le Rest et al., 2014), the leave-time-out (i.e., single time stamps are dropped from the training sample and used as test sample), and the leave-time-and-location-out scheme (i.e., at each iteration, model validation is performed predicting unknown locations and unknown points in time). Compared to the random K-fold CV schemes, the three approaches tend to provide higher out-of-sample error estimates, reducing potential spatiotemporal over-fitting of the models (Meyer et al., 2018). Thus, they should be considered key performance items for spatial and spatiotemporal models. However, random K-fold cross-validation remains a useful tool for assessing the spatiotemporal interpolation capability of models and is employed in spatiotemporal statistics literature (Gasch et al., 2015; Hengl et al., 2018; Ließet al., 2016; Yeşilkanat, 2020). Regarding the spatial correlation, Pohjankukka et al. (2017) suggested a spatial K-fold CV scheme ensuring that the training dataset is composed only by containing points that are at least a certain spatial or temporal distance away from the test dataset. Another possible approach is presented by Meyer and Pebesma (2021) who define the concept of applicability area, i.e., the area in which the model learns the

Table 2
Variables list.

Class	Variable name
Meteorology	Temperature Relative Humidity Pressure Rainfall Eastward component of wind Northward component of wind
Land	Vegetation High Vegetation Low Geopotential Height
Calendar	Week-end
Policy	Lockdown Restart

relationships based on the training data while the estimated cross-validation performance holds. Finally, it is also worth mentioning stratified (Zeng and Martinez, 2000) schemes, which ensure that the excluded observations in each fold maintain the informativeness of the overall sample according to one or more categorical variables, e.g., the distribution of station types among traffic, rural, and background.

In this paper, we employ a simpler random k-fold cross validation strategy to select the penalty parameter. We use a 10-fold approach, and given the large number of locations (84 sites) and time stamps (almost 4 years of daily data), for each fold, only 1-out-of -10 of the available space–time points are randomly eliminated. Since this strategy is prone to overestimate the predictive performance, we suggest to evaluate the predictive power in final re-estimation step of the HDGM using more sophisticated CV approaches. Therefore, in future studies, the above methods are expected to be employed to analyse how model selection (i.e., choice of penalty parameter) is affected when ignoring spatiotemporal dependence at this stage.

5. Modelling

In this section, we discuss the model implementation used in Section 6 to quantify the effects of the lockdown restriction on air quality. We consider a trivariate HDGM in which the responses of interest are the daily concentrations of NO_2, PM_{10}, and $PM_{2.5}$ observed from January 2017 to October 2020, i.e., 1400 days, from 84 ground measurement stations. All responses and covariates have been standardized. The variables used as predictors are listed in Table 2. To introduce interactions, some considerations are required.

Seasonal effects are important for empirically proving lockdown effects. For instance, Maranzano and Fassò (2021) showed that temperature plays a key role in improving air quality during winter and spring, but not during summer. Besides being influenced by the season, meteorology depends also on the geographic area. Therefore, we use also the area type (zoning) in the HDGM as a categorical variable. To do this, three of the four zones discussed in Section 2 and their corresponding interactions with environmental variables and week-end are included as covariates. The excluded zone, rural plain, is retained as the reference level, while metropolitan areas, mountains, and rural plains are directly included as dummy variables.

Moreover, one could expect varying lockdown effects for different urbanization conditions. That is, larger reductions could be expected in urbanized and heavily trafficked settings, while more modest shifts are expected in rural areas. Thus, the policy impact is differentiated by local conditions. To do this, we add as covariates the interactions between the lockdown and restart dummies and both the station type (small scale context) and the ARPA zoning (large scale context). The reference level for station type is background.

In summary, for each response variable Y_i, the full HDGM includes a constant term and 94 predictors. Let R be the set of six meteorological, three land and one weekend covariates, i.e., $R =$

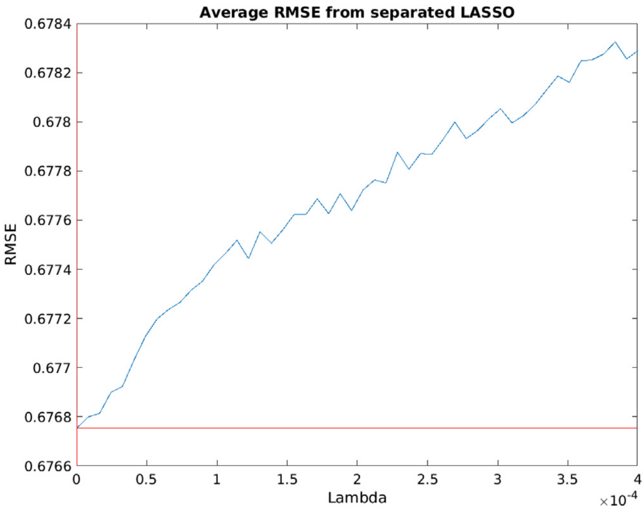


Fig. 6. Average RMSE (cross-validation) associated to each penalty parameter included in the dense grid and computed using the separated LASSO algorithm.

[Meteorology, Land, Calendar]. Moreover, consider the following dummy vectors: let S be the season with the spring as a reference, $S =$ [Autumn, Summer, Winter]; let T be the station type with background as a reference, $T =$ [Industrial, Rural, Traffic]; and let Z be the zoning with rural plain as a reference $Z =$ [MetrArea, Mountain, UrbPlain]. Eventually, let *Lockdown* be the lockdown dummy variable and similarly *Restart*. Then, using the Wilkinson notation, the full fixed-effect component of the HDGM can be expressed as:

$$\begin{aligned} Y \sim\ & R + S + Z + T + R \times S + R \times Z \\ & + Lockdown + Lockdown \times Z + Lockdown \times T \\ & + Restart + Restart \times Z + Restart \times T. \end{aligned} \tag{4}$$

Moreover, the multivariate spatiotemporal dynamics is described by other thirteen parameters, so that the total number of coefficients amounts to 295, and the overall data matrix has more than 11 million elements.

5.1. Variable selection

The overall dimension of the application for observations and variables is very high and computationally intensive. This leads us toward the use of variable selection techniques described in Section 4.2, which combines LASSO and HDGM cross-validation. As an illustration, we start first by conducting an ordinary LASSO optimization and then compare the two approaches.

The penalty parameter is identified using two sequential exponentially decaying grids for λ: the first coarse grid with a large range and widely spaced values and a second more dense grid around the point of optimum found with the first grid. The coarse grid ranges from 0 to 4. The minimum average RMSE across all three responses and $k = 10$ folds is obtained at $\lambda = 0$, indicating the ordinary least square estimate as the best. The search for the optimum is then refined through a denser grid with 50 values ranging from 0 to 10^{-4}. Fig. 6 shows the average RMSE computed for each of the fifty penalty coefficients.

The minimum of this curve is still obtained for $\lambda = 0$. This result allows us to highlight some important considerations. First, it is evident that the separate LASSO algorithms for each response uniquely suggest using a full model, without excluding any covariates among those

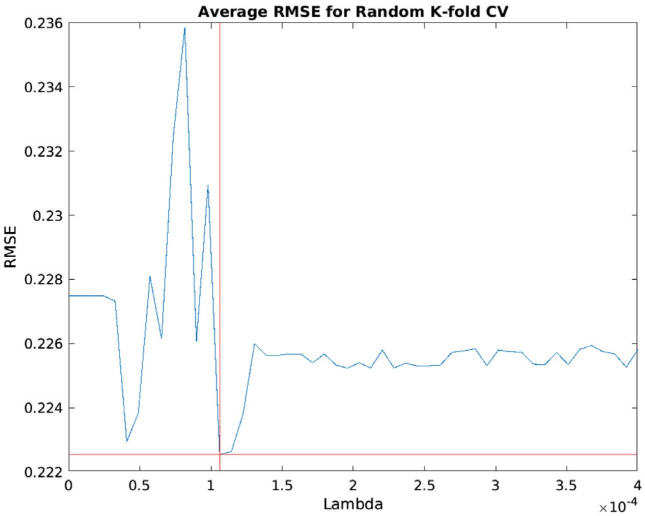


Fig. 7. Average RMSE (cross-validation) associated with each penalty parameter included in the dense grid and computed using the HDG model.

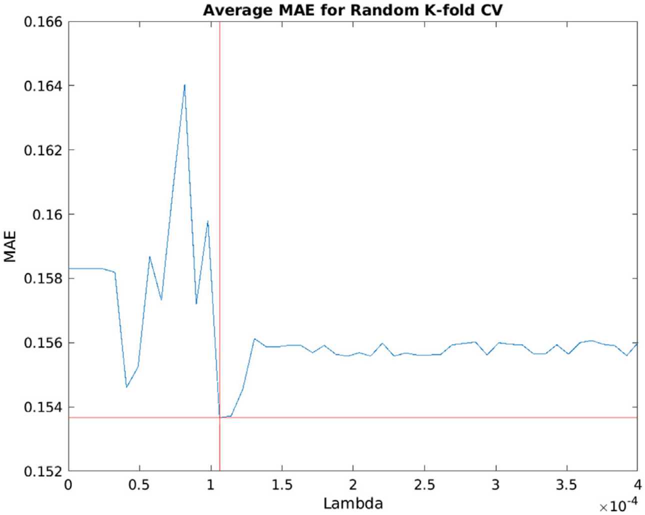


Fig. 8. Average MAE (cross-validation) associated with each penalty parameter included in the dense grid and computed using the HDG model.

included. Second, regarding the optimal RMSE value, we have $RMSE = 0.67$ giving $R^2 = 55\%$ as a determination coefficient averaged for the three responses.

Now, we move to the hybrid approach applied to the dense grid defined above. Using the matrix of LASSO coefficients associated with each λ, the distinct active sets were constructed and their predictive performances in the HDGM model were evaluated. The dense grid gives 44 distinct active sets. For each active set, the cross-validation prediction performance can be computed for the model that accounts for all spatial and temporal interactions. The results are reported in Figs. 7 and 8 for the RMSE and MAE, respectively.

This induces an RMSE that is much lower compared to the standard separated LASSO of Fig. 6. That is, a large share of the variability of the data is explained by the spatiotemporal dependence. The optimal value of $\lambda = 1.0612 \times 10^{-4}$ is different from zero and exceeds the optimal value for separated LASSO, meaning that some of the covariates can be excluded by the modelling obtaining an improvement in prediction accuracy. In the specific case, the optimal active set associated with the minimum RMSE excludes 9 covariates. Given the large number of predictors, the unselected variables seem to be a relatively minor part. Compared to the full model, in the optimal restricted model, only one of the covariates that identify the effect of lockdown and restart is removed. This is the interaction among the restart period and the mountain zone for NO_2.

Another point is that the prediction accuracy gain induced by the optimal reduced model seems to be very small. In general, this fact holds for all the active sets. Observing the vertical axis, we note that the values range from 0.222 to 0.236. Hence, at least, for forecasting, the models considered in this grid are equivalent. The RMSE values show that around the 95% of the total variability is explained by the spatiotemporal model (compared to the 55% of the pure LASSO model). That is, passing from a purely linear model to a geostatistical model, the prediction error reduces by 2/3, and the fit almost doubles.

The parameter estimates and standard errors of the HDGM used in the next section for COVID-19 impact assessment are reported in the auxiliary material.

5.2. Computational complexity

The model selection algorithm applied to the case study presented above requires a considerable computating time and memory. The calculations were conducted using a cluster machine comprising 64 CPUs that allow both the parallelization and the distribution on several machines of the tasks. The data preparation and MLE estimation of the full model (with 282 parameters) takes about 40 min. The estimates require around 160 EM iterations. The most time- and memory-demanding step of the algorithm is the computation of the variance–covariance matrix of the parameters taking another 32 min. This step cannot be run in parallel, but it is also required only once in the final re-estimation step. The phases of separated LASSO and cross-validation on the acrive stats are parallelizable by employing all the available cores. Considering the dense grid for the penalty term λ (51 distinct values), the separated LASSO step is completed in 34 min, most of which refers to values very close to zero. The active set cross-validation step is the most intensive and requires around 6 h to be completed, mainly due to CV computation. Considering that all the active sets are evaluated in parallel, the performances of each of the 10 folds are evaluated sequentially and take 35 min on average. Finally, the selected model is re-estimated with ML using the same time as the full model. Considering the available IT infrastructure, the entire process typically requires 8 h of computation time.

6. The impact of COVID-19 lockdown

Table 3 reports the estimated COVID-19 impacts on NO_2, PM_{10}, and $PM_{2.5}$, classified as marginal, zonal, and local effects. According to the interactions included in the model, the marginal effect of lockdown and restart coincides with the estimated variations at rural-background locations. Thus, they can be interpreted as the changes in background concentration levels in low-urbanized areas, providing a benchmark baseline of the effects. None of the coefficients associated with these areas are statistically significant. Meaning that where human presence is limited, the effects of shutdowns on pollution are marginal. Two other important outcomes are highlighted in the table. The first is that trafficked roads experienced a significant and large reduction of NO_2 concentration and a more modest but significant reduction of PM_{10}, while the $PM_{2.5}$ has non significant variations. The second important result concerns the significant reduction of NO_2 in metropolitan areas in general. Aggregating the coefficients, the estimated variation of NO_2 levels due to the traffic in metropolitan areas reduced around 15.5 $\mu g/m^3$, while in rural areas and peripheral districts the variation is -5.46 $\mu g/m^3$. This is important for exposure because most of the Lombardy population lives there. The other station types and zones have non significant changes. Importantly, the marginal variation

Table 3
Estimated lockdown and restart impact on the three pollutants. Marginal effects and interactions with land use at large scale (zone) and at small scale (site type) are reported.

	Lockdown impact [$\mu g/m^3$]		
	NO_2	PM_{10}	$PM_{2.5}$
Marginal effect	−7.6	+1.9	−0.09
×Metr. Area	-5.1^{***}	−3.3	−1.1
×Mountain	−1.5	−1.2	−1.5
×Urban Plain	−0.99	−0.84	+0.7
×Industrial sites	+1.7	−0.71	+0.7
×Rural sites	+2.2	+0.85	+1.5
×Traffic sites	-2.8^{***}	-1.5^{*}	+0.01
	Restart impact [$\mu g/m^3$]		
Marginal effect	−2.5	−2.3	−4.4
×Metr. Area	−2.7	−2.1	+1.5
×Mountain	0	−0.93	+0.4
×Urban Plain	−0.32	−1.2	+2.0
×Industrial sites	+1.3	−1.4	+2.8
×Rural sites	+1.6	−1.3	+2.5
×Traffic sites	−1.6	−0.91	+1.3

Note: Stars stand for statistical significance: '***' *pvalue* < 0.01, '**' *pvalue* < 0.05, '*' *pvalue* < 0.10.

of NO_2 is large but not statistically significant. This is due to the heterogeneity of the considered air quality monitoring network. Estimates for particulate matter show very weak statistical significance. The reduction of PM_{10} concentrations in congested metropolitan areas is statistically significant at 5% and amounts to 2.72 $\mu g/m^3$, while for $PM_{2.5}$, the reductions is 1.22 $\mu g/m^3$. In rural areas, a slight increase in concentrations is estimated. The findings for NO_2 parallel those of Agresti et al. (2020) and Maranzano and Fassò (2021), reporting reductions of the same scale. The results on particulate matter, however, support what has already been pointed out by the Lombardy Environmental Protection Agency, which identified minor reductions in concentrations in urbanized areas and steady or mildly increased levels in the agricultural plain ARPA Lombardia (2020).

As already mentioned in Section 2, the restart effects are more difficult to interpret, and none is statistically significant. Hence further research is needed to interpret them.

7. Conclusions and future works

Although one might suspect that air quality generally improved during the COVID-19 lockdown due to reduced human mobility, this effect is not so clear when seasonal effects, weather influences and generic spatiotemporal dependencies are also included. We studied these effects for fine and ultra fine particulate matters and nitrogen dioxide in an area with one of the most severe air pollution in Europe - Lombardy in northern Italy (a detailed overview of the region was provided in Section 2). Because the number of potential covariates that influence air quality is high, we proposed a hybrid approach for variable selection in multivariate geostatistical models. It combines the standard LASSO method and the HDGM. In this way, we could efficiently select relevant regressors in a first step and employ the improved forecasting performance by including spatiotemporal interactions in the second step. More precisely, after selecting the relevant regressors in linear regression framework with multiple interactions effects, the HDGM is re-estimated using the classical maximum likelihood approach. To choose the LASSO penalty parameter, we minimized the out-of-sample prediction performance of the HDGM obtained from random k-fold cross-validation study.

After adjusting for covariate effects and all spatial and temporal interactions, we could see that mainly the NO_2 concentrations were reduced during the lockdown period, however, only in metropolitan areas or for traffic-close stations. For all remaining types of stations, we could not see statistically significant reductions. This also underlines the influence of traffic on the NO_2 concentrations. Furthermore, we observed significant reductions of PM_{10} concentrations at traffic stations. Compared to the large NO_2 reductions, the fine particulate matter concentrations decreased less in absolute terms. The concentration of ultra fine particles was not significantly improved — neither for rural nor urban stations.

In summary, penalized regression approaches are promising attempts for model selection in geostatistics. Instead of the hybrid approach, joint procedures for model selection and estimation is an important open topic, such as penalized maximum likelihood methods for the HDGM. In this regard, more advanced block-wise and stratified sampling strategies for cross validation should also be analysed. For the impact of the COVID-19 lockdown, we estimated significant reduction effects for selected stations and pollutants, while the restart effects (after the lockdown) are insignificant and generally more difficult to interpret. This point also requires future studies in research.

References

Agresti, V., Balzarini, A., Bonanno, R., Collino, E., Colzi, F., Lacavalla, Pirovano, M., Riva, G., Toppetti, A., Riva, F., Piccoli, A., 2020. Gli effetti del lockdown sulla qualitá dell'aria a Milano e in Lombardia. Report URL: https://dossierse.it/05-2020-gli-effetti-del-lockdown-sulla-qualita-dellaria-a-milano-e-in-lombardia/.

Alduchov, O.A., Eskridge, R.E., 1996. Improved magnus' form approximation of saturation vapor pressure. J. Appl. Meteorol. 35, 601–609.

ARPA Lombardia, 2020. Analisi Preliminare Della Qualità Dell'aria in Lombardia Durante L'emergenza COVID-19. ARPA Lombardia, URL: https://www.arpalombardia.it/Pages/Qualit%C3%A0-dell%E2%80%99aria-durante-l%E2%80%99emergenza-Covid-19,-l%E2%80%99analisi-di-Arpa-Lombardia-.aspx.

Beauchamp, M., de Fouquet, C., Malherbe, L., 2017. Dealing with non-stationarity through explanatory variables in kriging-based air quality maps. Spatial Stat. 22, 18–46. http://dx.doi.org/10.1016/j.spasta.2017.08.003.

Belloni, A., Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19 (2), 521–547.

Calculli, C., Fassò, A., Finazzi, F., Pollice, A., Turnone, A., 2015. Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. Environmetrics 26 (6), 406–417.

Cameletti, M., 2020. The effect of corona virus lockdown on air pollution: Evidence from the city of Brescia in Lombardia region (Italy). Atmos. Environ. 239, 117794. http://dx.doi.org/10.1016/j.atmosenv.2020.117794.

Chu, T., J., Z., H., W., 2011. Penalized maximum likelihood estimation and variable selection in geostatistics. Ann. Statist. 39 (5), 2697.

Chzhen, E., Hebiri, M., Salmon, J., 2019. On lasso refitting strategies. Bernoulli 25 (4A), 3175–3200.

Collivignarelli, M.C., Abbà, A., Bertanza, G., Pedrazzani, R., Ricciardi, P., Carnevale-Miino, M., 2020. Lockdown for COVID-2019 in Milan: What are the effects on air quality? Sci. Total Environ. 732, 139280. http://dx.doi.org/10.1016/j.scitotenv.2020.139280.

European Environmental Agency, E., 2019. Air quality in Europe - 2019. Report.

European Environmental Agency, 2020. Air quality in Europe - 2020 report. Report http://dx.doi.org/10.2800/786656.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.

Finazzi, F., Fassò, A., 2014. D-STEM: A Software for the Analysis and Mapping of Environmental Space-Time Variables. Vol. 62. No. 6. p. 29. http://dx.doi.org/10.18637/jss.v062.i06.

Finazzi, F., Fassò, A., 2020. The impact of the COVID-19 pandemic on Italian mobility. Significance 17 (3), 17, Oxford, England.

Gasch, C.K., Hengl, T., Gräler, B., Meyer, H., Magney, T.S., Brown, D.J., 2015. Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: The cook agronomy farm data set. Spatial Stat. 14, 70–90. http://dx.doi.org/10.1016/j.spasta.2015.04.001, URL: https://www.sciencedirect.com/science/article/pii/S2211675315000251.

Guan, Y., Johnson, M.C., Katzfuss, M., Mannshardt, E., Messier, K.P., Reich, B.J., Song, J.J., 2020. Fine-scale spatiotemporal air pollution analysis using mobile monitors on google street view vehicles. J. Amer. Statist. Assoc. 115 (531), 1111–1124. http://dx.doi.org/10.1080/01621459.2019.1665526.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.

Hoeting, J., Davis, R.A., Merton, A.A., Thompson, S.E., 2006. Model selection for geostatistical models. Ecol. Appl. Publ. Ecol. Soc. Am. 16 1, 87–98.

Jalilian, A., Mateu, J., 2021. A hierarchical spatio-temporal model to analyze relative risk variations of COVID-19: a focus on Spain, Italy and Germany. Stoch. Environ. Res. Risk Assess. 35 (4), 797–812.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., Bretagnolle, V., 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. Global Ecol. Biogeogr. 23 (7), 811–820. http://dx.doi.org/10.1111/geb.12161.

Lee, W., Liu, Y., 2012. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. J. Multivariate Anal. 111, 241–255.

Ließ, M., Schmidt, J., Glaser, B., 2016. Improving the spatial prediction of soil organic carbon stocks in a complex tropical mountain landscape by methodological specifications in machine learning approaches. Plos One 11 (4), e0153673. http://dx.doi.org/10.1371/journal.pone.0153673.

Maranzano, P., Fassò, A., 2021. The impact of the lockdown restrictions on air quality during COVID-19 pandemic in lombardy, Italy. In steland eds. (2021). In: Steland, A. (Ed.), Artificial Intelligence, Big Data, Data Science and Machine Learning in Statistics - Challenges, Perspectives and Solutions in Environmental Science, Natural Sciences and Technology. in Printing.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Model. Softw. 101, 1–9.

Piter, A., Otto, P., Alkhatib, H., 2020. A spatiotemporal functional model for bike-sharing systems – an example based on the city of Helsinki. arXiv:2012.10746.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., Heikkonen, J., 2017. Estimating the prediction performance of spatial models via spatial k-fold cross validation. Int. J. Geogr. Inf. Sci. 31 (10), 2001–2019. http://dx.doi.org/10.1080/13658816.2017.1346255.

Raffaelli, K., Deserti, M., Stortini, M., Amorati, R., Vasconi, M., Giovannini, G., 2020. Improving air quality in the Po valley, Italy: Some results by the LIFE-IP-PREPAIR project. Atmosphere 11 (4), 429.

Rasmussen, C., Williams, C., 2006. Gaussian Processes for Machine Learning. MIT Press.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40 (8), 913–929.

Shen, C.-W., Chen, Y.-H., Chen, C.-S., 2021. Distribution-free regression model selection with a nested spatial correlation structure. Spatial Stat. 41, 100476. http://dx.doi.org/10.1016/j.spasta.2020.100476.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Wan, Y., Xu, M., Huang, H., Chen, S.X., 2020. A spatio-temporal model for the analysis and prediction of fine particulate matter concentration in Beijing. Environmetrics 32 (1), 1–16.

Wang, Y., Finazzi, F., Fassò, A., 2021. D-STEM v2: A software for modelling functional spatio-temporal data. 99, (10), pp. 1–29. http://dx.doi.org/10.18637/jss.v099.i10, URL: https://www.jstatsoft.org/index.php/jss/article/view/v099i10/0,

Yeşilkanat, C.M., 2020. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos Solitons Fractals 140, 110210. http://dx.doi.org/10.1016/j.chaos.2020.110210, URL: https://www.sciencedirect.com/science/article/pii/S0960077920306068.

Zeng, X., Martinez, T.R., 2000. Distribution-balanced stratified cross-validation for accuracy estimation. J. Expe. Theor. Artif. Intell. 12 (1), 1–12. http://dx.doi.org/10.1080/095281300146272.

Zhao, P., Yu, B., 2006. On model selection consistency of lasso. J. Mach. Learn. Res. 7, 2541–2563.

Zheng, X., Guo, B., He, J., Chen, S.X., 2021. Effects of corona virus disease-19 control measures on air quality in north China. Environmetrics 32 (2), 1–16.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101 (476), 1418–1429.

Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. Ann. Statist. 36 (4), 1509.