Spatiotemporal modelling of $PM_{2.5}$ concentrations in Lombardy (Italy): a comparative study

Philipp Otto[1] · Alessandro Fusta Moro[2] · Jacopo Rodeschini[2] ·
Qendrim Shaboviq[3] · Rosaria Ignaccolo[4] · Natalia Golini[4] · Michela Cameletti[2] ·
Paolo Maranzano[5,6] · Francesco Finazzi[2] · Alessandro Fassò[2]

Abstract

This study presents a comparative analysis of three predictive models with an increasing degree of flexibility: hidden dynamic geostatistical models (HDGM), generalised additive mixed models (GAMM), and the random forest spatiotemporal kriging models (RFSTK). These models are evaluated for their effectiveness in predicting $PM_{2.5}$ concentrations in Lombardy (North Italy) from 2016 to 2020. Despite differing methodologies, all models demonstrate proficient capture of spatiotemporal patterns within air pollution data with similar out-of-sample performance. Furthermore, the study delves into station-specific analyses, revealing variable model performance contingent on localised conditions. Model interpretation, facilitated by parametric coefficient analysis and partial dependence plots, unveils consistent associations between predictor variables and $PM_{2.5}$ concentrations. Despite nuanced variations in modelling spatiotemporal correlations, all models effectively accounted for the underlying dependence. In summary, this study underscores the efficacy of conventional techniques in modelling correlated spatiotemporal data, concurrently highlighting the complementary potential of Machine Learning and classical statistical approaches.

Handling Editor: Giada Adelfio

Extended author information available on the last page of the article

1 Introduction

The Lombardy region, situated in the heart of the Po Valley in Northern Italy, is known to be highly polluted due to the natural barrier created by the Alps hindering the dispersion of air pollutants (see, e.g., Pernigotti et al. 2012). Fine particulate matter ($PM_{2.5}$) has been identified as the most hazardous air pollutant (European Environmental Agency 2022a), representing a mix of air pollutants with a diameter less than 2.5 μm (see also Jerrett et al. 2005, for a review). Information about the air quality dynamics is essential for decision-makers to effectively mitigate adverse effects. Statistical and machine learning models can offer valuable insights into this behaviour and its consequences, including the identification of pollution sources and the factors influencing its behaviour and forecasting future pollution levels under various scenarios, such as changes in emissions or weather patterns. Moreover, the combination of different modelling techniques may further enhance the results. By leveraging the strengths of each approach, decision-makers can gain a more comprehensive understanding of air pollution and make more informed choices for mitigation strategies.

This paper compares three statistical and machine learning models with varying degrees of flexibility to elucidate daily $PM_{2.5}$ concentrations in the Lombardy region. More precisely, hidden dynamic geostatistical models (HDGM), generalised additive mixed models (GAMM), and random forest spatiotemporal kriging (RFSTK) were utilised to describe the relationships between a large set of predictors and $PM_{2.5}$ concentrations. All three models employed in this study have been specifically developed to handle spatiotemporal data. HDGM incorporates a latent variable to capture spatiotemporal dependence, while external factors are included in a linear manner within the model. Conversely, GAMM allows for the nonlinear impact of exogenous predictors, which are estimated using splines. It incorporates spatiotemporal dependence by utilising a smoothing spline for spatial variation and a first-order autoregressive process for temporal dependence. Lastly, RFSTK employs a random forest (RF) to model the nonlinear effects of predictors and then a spatiotemporal kriging model to account for the possible spatiotemporal dependence.

These models are frequently applied in diverse areas. First, HDGM has been primarily used for air pollution studies (see, e.g., Najafabadi et al. 2020; Taghavi-Shahri et al. 2020 for air pollution in Iran, and Maranzano et al. 2023; Fassò et al. 2022; Calculli et al. 2015 for Italy). Notably, there are further applications in other fields, such as modelling bike-sharing data or coastal profiles (see Piter et al. 2022; Otto et al. 2021). HDGM is a linear mixed effects model with a specific structure of the random effects capturing the spatiotemporal dynamics of environmental data, which are widely applied in diverse areas (see, e.g., Jiang and Nguyen 2007 for an overview). Second, GAMM has been employed in various areas, such as ecology (Knape 2016; Kneib et al. 2011), psychology (Bono et al. 2021), economics (Fahrmeir and Lang 2001), psycholinguistics (Baayen et al. 2017), or event studies (Maranzano and Pelagatti 2023). Third, needless to say, models based on decision trees have demonstrated their effectiveness in capturing complex patterns in different fields (see, e.g., Belgiu and Drăguţ 2016

for an overview in remote sensing, or Qi 2012 for bioinformatics), particularly in combination with kriging approaches, e.g., in environmental (Sekulić et al. 2020; Chen et al. 2019; Guo et al. 2015), or air pollution studies (Liu et al. 2018, 2019). We refer the interested reader to the systematic literature review of Patelli et al. (2023) for a structured overview of these approaches. Furthermore, an interesting new approach is to use deep neural networks for the prediction and interpolation of spatial data (see Nag et al. 2023; Daw and Wikle 2023). Hybrid models, integrating different models in one single framework and exhibiting good robustness and adaptability, can combine the advantages of different models during the different stages of the modelling phase. Their adoption is rapidly increasing in various fields and predictions, including $PM_{2.5}$ (e.g., Bai et al. 2022; Tsokov et al. 2022; Sun and Xu 2022; Wang et al. 2019; Ding et al. 2021), greenhouse gas emissions (Javanmard and Ghaderi 2022), tea yield (Jui et al. 2022), depopulation in rural areas (Jato-Espino and Mayor-Vitoria 2023), or disease monitoring (Kishi et al. 2023) and calibration of citizen-science air quality data (Bonas and Castruccio 2021).

All three models account for the intrinsic spatial, temporal, and spatiotemporal dynamics of the $PM_{2.5}$ concentrations. This temporal and spatial dependence arises from the persistence of the particles in the atmosphere over a certain time and, simultaneously, from the displacement and spread of the particles to nearby areas, e.g., by wind (Merk and Otto 2020). Previous studies successfully employed several statistical models to model air pollution scenarios in Northern Italy, such as generalised additive models (Bertaccini et al. 2012), Bayesian hierarchical modes based on the stochastic partial differential equation approach (Cameletti et al. 2013; Fioravanti et al. 2021), or random forests (Stafoggia et al. 2019). In a comparative study for Northern Italy, Cameletti et al. (2011) studied the effectiveness of different statistical models in a Bayesian framework. Machine learning algorithms, including random forests, are adept at capturing nonlinearities and interactions. Still, when applied to air quality modelling, the spatiotemporal nature of the phenomenon is often ignored (see, e.g., Fox et al. 2020). Consequently, the model's performance deteriorates, with worse outcomes than those obtained from Kriging with External Drift (KED), considered standard for modelling spatiotemporal phenomena. KED shows better results than random forest in Lombardy (Fusta Moro et al. 2022) and in the USA (Berrocal et al. 2020). On the other hand, machine learning algorithms outperform classical models if spatiotemporal dependence is not considered at all (Kulkarni et al. 2022). Lu et al. (2023) compared geostatistical and ML models for NO_2 concentrations in Germany. Despite the limited number of studies comparing geostatistical and ML models, this subject is gaining increasing interest because the comparison provides valuable insights into the dynamics of the process, as we will illustrate below.

The remaining sections of this paper are organised as follows. Section 2 describes the general framework of the study and the data set used for our comparisons. Then, we explain the theoretical background of all considered models in Sect. 3. The comparative study is presented in Sect. 4, including fitting procedure (Sect. 4.1), residual analysis (Sect. 4.2), prediction performances within the cross-validation scheme (Sect. 4.3), and model interpretation (Sect. 4.4). Section 5 concludes the paper.

2 Data

Our comparative analysis is based on the *Agrimonia data set*, a comprehensive daily spatiotemporal data set for air quality modelling available open-access on Zenodo (Fassò et al. 2023). Specifically, it includes air pollutant concentrations and important covariates for all 141 stations of the air quality monitoring network in the Lombardy region and a 30 km buffer zone around the administrative boundaries. The data originates from multiple sources with different temporal and spatial resolutions. Using suitable aggregation and interpolation techniques described in Fassò et al. (2023), the *Agrimonia data set* is available on a daily basis for all ground-level measurement stations in the study area. It spans six years, from 2016 to 2021, and includes daily air pollutant concentrations, weather conditions, emissions flows, land use characteristics, and livestock densities. We summarise all variables considered in this study in Table 1, including their main descriptive statistics.

The response variable is the $PM_{2.5}$ concentration at the ground described in ensuing Sect. 2.1 in more detail, while the selected remaining variables of the *Agrimonia data set* serve as explanatory variables or features. They are summarised and motivated in Sect. 2.2.

2.1 $PM_{2.5}$ concentrations

The *Agrimonia data set* includes daily observations of several atmospheric pollutants retrieved from the Italian air quality monitoring network. Not all monitoring stations are equipped with the same sensors, so we have excluded locations where stations were not measuring $PM_{2.5}$, which is the target pollutant of this study. The 49 remaining stations are depicted in Fig. 1 (left) along with information about the type of surrounding area (rural, suburban, urban) and the primary nearest emission source (background, industrial, traffic), according to the EU classification (European Environmental Agency 2022b). To depict the spatial variation of the PM concentrations, we coloured the stations according to the average daily concentration across the entire time period on the right-hand map in Fig. 1.

We consider the period from 2016 to 2020. The temporal variation of the observed $PM_{2.5}$ concentrations, grouped by months and by type of area, is displayed through a series of boxplots in Fig. 2. The colours are chosen according to the type of the surrounding area. Not surprisingly, there is a clear seasonality with higher concentrations in winter due to meteorological conditions resulting in reduced air circulation. The median concentrations range between 10 μg/m^{-3} and 40 μg/m^{-3} across the year. Thus, throughout the year, the median concentration was beyond the threshold of 5 μg/m^{-3} considered hazardous by the World Health Organisation guidelines (WHO 2021). From Fig. 2, it is clear that all different types of areas are similarly affected by poor air quality. This spatial homogeneity is also visually confirmed in the map of Fig. 1 where clusters of similar neighbouring concentrations can be seen, suggesting a pronounced spatial dependence.
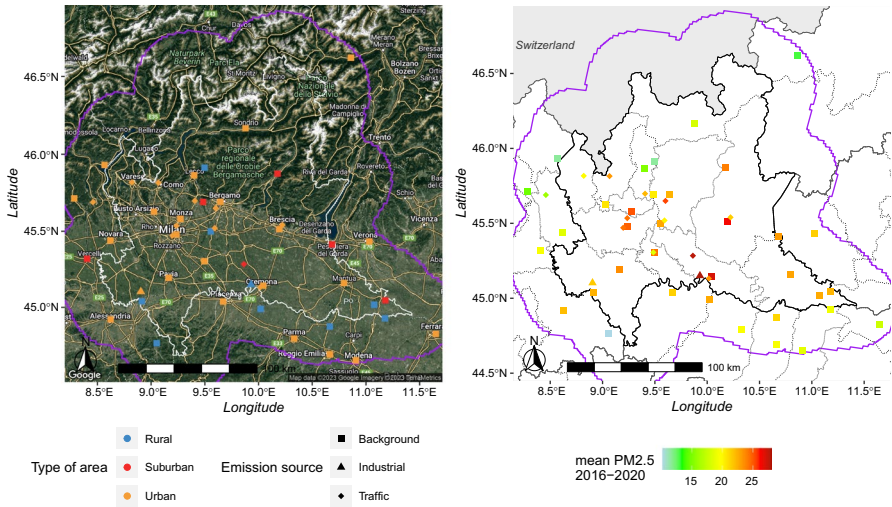
Fig. 1 Map of the 49 PM$_{2.5}$ monitoring stations extracted from the *Agrimonia data set*. The purple line represents a 0.3° buffer around the administrative boundaries of the Lombardy region, the latter represented by the white line. Left: Stations are coloured according to the type of area. The shape indicates the main emission sources. Right: the stations are coloured according to the 2016–2020 average PM$_{2.5}$ concentrations [μg/m^{-3}]
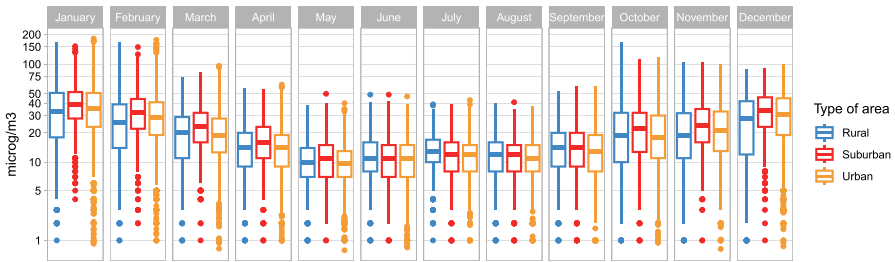

Fig. 2 Monthly boxplots of PM$_{2.5}$ concentrations (on a log scale) measured in Lombardy, including the buffer area

To explore a possible spatiotemporal correlation, we estimate a spatiotemporal variogram $\gamma(h, \tau)$ based on the sample variance of observations within certain distance ranges in space and time h and τ (see, e.g., Cressie and Wikle 2015). Smaller values of the variogram for smaller distances indicate (short-term) statistical dependence. The spatiotemporal variogram of observed PM2.5 concentrations is depicted in Fig. 3. As expected, the variogram identifies an apparent correlation of the PM$_{2.5}$ concentrations across time and space. More precisely, the values of the variogram for the first temporal lags indicate a pronounced temporal dependence within the first 5-6 days, i.e., approximately one week. Furthermore, we observe a noticeable spatial dependence since the variogram increases with
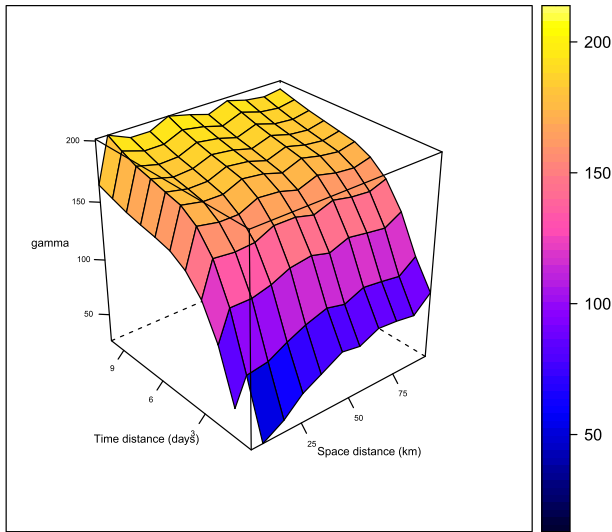
Fig. 3 Spatiotemporal variogram of the PM$_{2.5}$ concentrations

increasing spatial distances. It is important to note that this variation still includes spatial and temporal seasonalities and variations caused by exogenous factors.

2.2 Regressors

All models, which will later be used for the comparison, share the same set of regressors, including weather conditions and livestock densities. Based on an extensive literature review on air pollution modelling, we carefully selected key weather variables and incorporated information regarding local-scale animal breeding. The specific variables considered are presented in Table 1 and their corresponding descriptive statistics. Furthermore, monthly indicator variables are included to capture the seasonality, as highlighted by Fig. 2. Below, we will provide a brief motivation for each variable and the descriptive statistics in Table 1 to offer an intuitive understanding of the explanatory variables.

Firstly, several studies have found that weather is a crucial factor in air quality modelling (Bertaccini et al. 2012; Ignaccolo et al. 2014; Merk and Otto 2020; Fassò et al. 2022; Grange et al. 2023; Chang and Zou 2022). Changes in weather conditions such as temperature, precipitation, and wind speed and direction can affect atmospheric stability and turbulence, which can influence the transport and deposition of

Table 1 Variables selected from the *Agrimonia data set*

Variable name and description [unit of measurement]	Min	Mean	SD	Max
Altitude	4	171.458	200.004	1194
Height in relation to sea level [m]				
AQ_pm25	0.506	20.896	16.353	182
Fine particulate matter concentrations [µg/m^{-3}]				
WE_temp_2m	-20.650	12.916	8.233	32.880
Air temperature at 2 m [°C]				
WE_tot_precipitation	0	0.003	0.008	0.172
Total precipitation [m]				
WE_rh_mean	19.490	74.433	12.299	99.520
Relative humidity [%]				
WE_wind_speed_100m_mean	0.564	2.550	1.326	11.930
Average wind speed at 100 m [m/s]				
WE_blh_layer_max	13.790	1, 039.402	556.877	4, 421.000
Daily maximum height of the air mixing layer layer m]				
LI_pigs_v2	0.022	115.215	159.666	652.100
Average density of pigs bred for the area (10 km^2) surrounding the measurement stations [number/ km^2]				
LI_bovine_v2	1.543	46.241	47.463	178.800
Average density of bovine bred for the area (10 km^2) surrounding the measurement stations [number/km^2]				
LA_hvi	0.861	2.324	0.804	5.034
High vegetation abundance [m^2/m^2]				
LA_lvi	0.865	2.208	0.560	3.662
Low vegetation abundance [m^2/m^2]				

Names are consistent with the data set (Fassò et al. 2023) and the accompanying data descriptor (Fassò et al. 2023)

pollutants. For instance, temperature and boundary layer height are usually negatively related to air pollutant concentrations. Similarly, we typically observe reduced PM concentrations during periods with increased precipitation or wind speed. On the contrary, the direction and size of the effect of the relative humidity are still debated, but it undoubtedly affects the PM$_{2.5}$ concentrations (Zhang et al. 2017).

Secondly, we considered agricultural influences, which appear to impact air quality (e.g., Thunis et al. 2021; Lovarelli et al. 2020). The Livestock (LI) data used in this study provide information on the average density of pigs and cattle per municipality (expressed as animals per km^2) in the vicinity of each station (within a radius of 10 km^2). Including livestock data is essential to capture the impact of ammonia (NH$_3$) emissions on air quality, as livestock farming is the major source of NH$_3$ emissions (up to 95%). Therefore, including LI data in air quality modelling can help better understand and mitigate livestock's impact on air pollution levels.

3 Spatiotemporal statistical models and machine learning techniques

We consider the $PM_{2.5}$ concentrations as realisations of a spatiotemporal stochastic process $\{Z(s,t) : s \in D, t = 1, 2, \ldots, T\}$, where D is the spatial domain that contains a set of locations $\{s_i : i = 1, \ldots, n\}$ (i.e., the ground-level measurement stations) and the temporal domain is discrete $t = 1, \ldots, T$ (i.e., daily observations). Furthermore, we posit that $Z(s, t)$ might be influenced by external variables related to weather conditions, emissions, or agricultural activities. Throughout the remainder of the paper, the terms regressors, covariates, and features are used interchangeably. The spatiotemporal proximity of observations typically induces statistical dependence, and thus, the selected model candidates should appropriately incorporate this inherent spatiotemporal dependence. To structure the model alternatives, we can decompose the models into three terms, i.e.,

$$Z(s,t) = S(s,t) + U(s,t) + \varepsilon(s,t), \tag{1}$$

where $S(s, t)$ is the large-scale component including the regressors, $U(s, t)$ includes small-scale spatiotemporal effects, and $\varepsilon(s, t)$ comprises the measurement and modelling errors, which are assumed to be a zero-mean white noise process.

3.1 Hidden dynamic geostatistical model

The first model selected is the HDGM, which serves as the comparative analysis' starting point or baseline method. It is a widely applied geostatistical model, first considered by Huang and Cressie (1996) as an extension of classical mixed-effects models for univariate spatiotemporal data. Calculli et al. (2015) extended the HDGM to multivariate data. This modelling approach proved particularly useful for air quality modelling (e.g., Fassò and Finazzi 2011; Finazzi et al. 2013), as the comparative study of Cameletti et al. (2011) confirmed. The HDGM specifies the large-scale effects as a linear regression model, i.e.,

$$S(s,t) = \mathbf{X}_\beta(s,t)'\boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ is a vector of p fixed-effect coefficients, including the model intercept β_0, and $\mathbf{X}_\beta(s, t)$ is the (s, t)-th entry of the fixed design matrix of the selected covariates/features. In other words, $\mathbf{X}_\beta(s, t)$ is the vector of the observed covariates at location s and time point t.

The spatiotemporal dependence is modelled as small-scale effects by a geostatistical process

$$U(s,t) = v\xi(s,t), \tag{3}$$

where v is an unknown, homoscedastic scaling factor, which has to be estimated and describes the degree of the small-scale effects. Further, $\xi(s, t)$ is a latent random variable with Markovian temporal dynamics given by

$$\xi(s,t) = g_{HDGM}\xi(s,t-1) + \eta(s,t), \quad \eta(s,t) \sim GP, \tag{4}$$

where $g_{HDGM}\xi(s,t-1)$ is a hidden first-order autoregressive process with coefficient g_{HDGM}. The temporal dependence is separated from the spatial interactions, which are modelled in $\eta(s,t)$. It is worth noting that this implies a separable space-time covariance. Specifically, $\eta(s,t)$ is a Gaussian process (GP) with zero mean, unit variance, and covariance matrix determined by an exponential spatial correlation function

$$\rho(||s-s'||;\theta_{HDGM}) = exp(-||s-s'||/\theta_{HDGM}) \tag{5}$$

with θ_{HDGM} being the range parameter, s and s' are two distinct spatial locations, and the distance between them is given by the vector norm $|| \cdot ||$. For this study, we will always employ the distance on the great circle, i.e., the length of the geodesic between s and s'. The parameters of the random effects process $\xi(s,t)$ are assumed to be in a space leading to a weakly stationary spatiotemporal process. Finally, $\varepsilon(s,t)$ is an identically distributed random error independent across space and time with zero mean and constant variance σ_ϵ^2.

The model parameter set $\boldsymbol{\Phi} = \{\boldsymbol{\beta}, g_{HDGM}, \theta_{HDGM}, v, \sigma_\epsilon^2\}$ is estimated by the maximum-likelihood method using an expectation-maximisation (EM) algorithm (Calculli et al. 2015). The estimation procedure is computationally implemented in the MATLAB software package D-STEM (see Wang et al. 2021).

3.2 Generalised additive mixed model

Compared to generalised additive models (GAM, Hastie and Tibshirani 1987), generalised additive mixed models (GAMM) include a random-effects component to describe correlated response variables, such as time series, spatial or spatiotemporal data. It extends the HDGM by allowing for linear and nonlinear regressive effects in a GAM fashion, i.e., the response variable linearly depends on smooth functions of the predictors. To be precise, the large-scale components are given by

$$S(s,t) = \mathbf{X}_{linear}(s,t)'\boldsymbol{\beta}_{linear} + \underbrace{\sum_{j=1}^{m} \alpha_{(j)}(\mathbf{X}_{nonlinear,j}(s,t))}_{\text{nonlinear effects}} \tag{6}$$

with $\mathbf{X}_{linear}(s,t)\boldsymbol{\beta}_{linear}$ being a linear parametric regression term of the first k covariates, with a parameter vector $\boldsymbol{\beta}_{linear} = (\beta_0, \beta_1, \ldots, \beta_k)'$, including the intercept term β_0. Moreover, $\sum_{j=1}^{m} \alpha_{(j)}(\mathbf{X}_{nonlinear,j}(s,t)))$ is an additive term with nonlinear influence functions $\alpha_{(j)} : \mathbb{R} \to \mathbb{R}$ of the j-th column in $\mathbf{X}_{nonlinear,j}$ for the remaining m regressors. These nonlinear influences can be estimated along with the other model coefficients, e.g., as regression splines or penalised splines (Fahrmeir et al. 2004).

The small-scale effects of the GAMM are specified as a first-order autoregressive model for the temporal dependence and a smooth spatial surface for the spatial dependence, that is,

$$U(s,t) = g_{GAMM}(Z(s,t-1) - S(s,t-1) - C(s)), \tag{7}$$

where g_{GAMM} is the parameter representing the temporal dependence, where zero indicates no temporal correlation. The spatial dependence is modelled as a smooth surface $C(s)$, which follows a Gaussian process with exponential covariance function with range parameter θ_{GAMM} (Handcock and Wallis 1994). This structure is identical to the spatial term of the random effects model in HDGM as given by equation (5). In our case, θ_{GAMM} is estimated as proposed by Kammann and Wand (2003). The model estimation is computationally implemented in the package `mgcv` available in R (Wood 2017).

3.3 Random forest spatiotemporal kriging

For the third approach, RFSTK, we increase the flexibility of the model in the large-scale component by considering a random forest (RF) algorithm. In other words, the third hybrid model combines an RF for the large-scale component $S(s,t)$ with that of a spatiotemporal kriging model for $U(s,t)$. The idea traces back to the combination of random forests and kriging, the so-called random forest residual kriging, which—even if only considering spatial dependence—showed promising results compared to RF alone (e.g. Wang et al. 2019; Viscarra Rossel et al. 2014). For spatiotemporal data, RFSTK has been considered to model air quality, again showing good performances (see Zhan et al. 2018; Shao et al. 2020).

Random forests are widely used tools in machine learning as an ensemble of multiple decision trees (Breiman 2001). They are constructed as an ensemble of multiple decision trees, making them highly versatile and robust for a wide range of predictive tasks. In a regression problem, the prediction of the large-scale effect is obtained by averaging across the predictions of n_{tree} decision trees. Each of these decision trees is trained or estimated from independent bootstrap samples Z_j^* of the input data. These bootstrap samples are created through random resampling without replacement (recommended for dependent variables, see Strobl et al. 2008) from the original dataset. The predictions of these individual decision trees $\widehat{E}(Z_j^*(s_i,t) \mid \mathbf{X}(s_i,t))$ are then averaged to produce the final prediction. This ensemble approach helps to improve the robustness and generalisation of the model, reducing the risk of overfitting. Thus, the large-scale model is given by

$$S(s_i,t) = \frac{1}{n_{\text{tree}}} \sum_{j=1}^{n_{\text{tree}}} \widehat{E}(Z_j^*(s_i,t) \mid \mathbf{X}(s_i,t) : i = 1, \ldots, n, t = 1, \ldots, T), \tag{8}$$

where $\widehat{E}(Z_j^*(s_i,t) \mid \cdot)$ is the prediction of the j-th decision tree. It is essential to emphasise that for regression trees within the random forest framework, the averaging process should be conducted for each region of interest in the covariate space. This means that the model considers the different regions of the input space and provides predictions tailored to the characteristics of each region. Random forests excel at handling complex, nonlinear relationships and are widely used in various

applications, including classification and regression tasks, as well as feature selection and data exploration.

The small-scale model of RFSTK is assumed to be a zero-mean, weakly stationary spatiotemporal Gaussian process

$$U(s,t) = \tilde{\eta}(s,t) \sim GP, \tag{9}$$

where the covariance matrix is obtained from a separable space-time correlation function given by

$$\rho(||s-s'||, |t-t'|) = \rho(||s-s'||, \theta_{RFSTK_s}) \cdot \rho(|t-t'|, \theta_{RFSTK_t}) \tag{10}$$

with $||s-s'||$ and $|t-t'|$ representing spatial and temporal distances between (s,t) and (s',t'), respectively. That is, the exponential correlation functions are equivalent to the spatial correlation function of the HDGM and GAMM, but the other two approaches consider an autoregressive temporal dependence, while the RFSTK employs a continuous correlation function for both the temporal and spatial dependence. The parameters and decision trees are estimated in a two-step procedure. First, predictions for the large-scale component are obtained using an RF, computationally implemented in the R package `randomForest` (Liaw and Wiener 2002). Second, to adjust the predictions of the RF accounting for space-time interactions, the parameters of the separable space-time correlation function in (10) are estimated by variography on RF residuals, implemented in the R package `gstat` (Gräler et al. 2016).

4 Comparative study

In the subsequent section, we present an application of each methodology on air quality data extracted from the *Agrimonia data set* (see Sect. 2). We start with exploring the model fitting process and examining the residuals. Subsequently, we evaluate the predictive performances of the models through cross-validation. Finally, our attention shifts to the interpretation of each model and its comparison. Drawing upon this comparison, we offer practical suggestions for integrating the modelling outcomes into other environmental analyses. All source codes can be found in the supplementary material.

4.1 Model fitting

To ensure the comparability of the results across all three model alternatives, we considered all predictors listed in Table 1 in their original scale. Transformations such as a logarithmic transformation of the response variable did not generally yield better prediction results and model fits. This suggests an additive structure in the large-scale effects, and its coefficients can directly be interpreted as marginal (linear) effects.

The large-scale component $S(s,t)$ of HDGM models the regressors' influence through linear relationships while the small-scale $U(s,t)$ captures the

spatiotemporal dependence through a latent variable. The large-scale component $S(s, t)$ of GAMM captures nonlinear effects by estimating a functional relationship between the predictors and the response variable. Identifying predictors requiring nonlinear relationships entailed simulating model residuals from a linear regression model and graphing them alongside their corresponding predictors, including confidence intervals, as suggested in Fasiolo et al. (2020). A smooth nonlinear effect is estimated if a pattern outside the confidence bands is detected. In this study, all continuous variables, except for altitude, required a smooth term represented by a penalised thin plate regression spline. The spatial dependence is modelled within the GAMM as a two-dimensional smooth surface $C(s)$ governed by a zero-mean Gaussian process with an exponential covariance function, while the temporal dependence is represented by an autoregressive term of order one. The model is estimated through the restricted maximum likelihood method using the package `mgcv` in R (Wood 2011).

To optimise the RF algorithm, we evaluate the out-of-bag root mean squared error across various hyperparameter settings, including the number of trees n_{tree}, the number of candidate predictors for building each tree, and the size of the final leaves of each tree. Our findings indicate that utilising default settings (500 trees, the number of candidate variables equal to one-third of the number of the predictors, and final leaf size of 5) within the R package `randomForest` (Liaw and Wiener 2002) is suitable. Finally, RF predictions are adjusted by adding RF residual predictions obtained by fitting an ordinary spatiotemporal kriging model by using the R package `gstat` (Gräler et al. 2016).

To highlight the difference between the model fits, we compared the in-sample (i.e. using the entire data set) predictive performance of all the models, separately for the large-scale component (LS) and the full model (FM), including the space-time effects. The comparison was based on the root mean squared errors (RMSE), mean absolute errors (MAE), and the coefficient of determination R^2. The results are reported in Table 2. HDGM generally had better prediction capabilities and lower computational costs than the other models when considering the full model. For comparison, RFSTK model had relatively good prediction capabilities but was computationally intensive, particularly when including the spatiotemporal kriging. For our dataset, the GAMM model had the lowest in-sample fit. Since $C(s)$ serves as the constant model intercept in GAMM, we incorporated

Table 2 In-sample performance of the three models assessed by RMSE (in μg/m^3), and the adjusted coefficient of determination R^2

|	HDGM		GAMM		RFSTK	
	LS	FM	LS	FM	LS	FM
RMSE [μg/m^{-3}]	11.92	1.814	11.456	8.468	8.245	5.361
MAE [μg/m^{-3}]	0.068	− 0.002	− 0.002	− 0.003	− 0.129	− 0.019
R^2	0.469	0.988	0.509	0.732	0.746	0.893

The fit of the only large-scale component (LS) is compared to the full model (FM)

$C(s)$ in the LS component (i.e., $S(s,t) + C(s)$) to enable direct comparisons with the regression terms in HDGM.

In general, the informativeness of the in-sample results can be questionable due to the potential sensitivity of models to the training data or overfitting. To address this issue, we evaluated prediction performances within a cross-validation scheme, which is explained in Sect. 4.3. For instance, this analysis revealed that the HDGM could generalise the estimated relation to obtain outperforming out-of-sample predictions across space, while we observe a serious overfit in the in-sample case due to the flexibility of the random-effects model. We will focus on this result in more detail below.
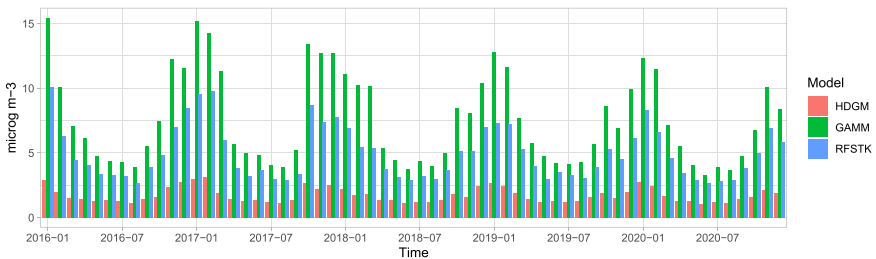
(a) Standard deviation of FM residuals grouped by months and models.
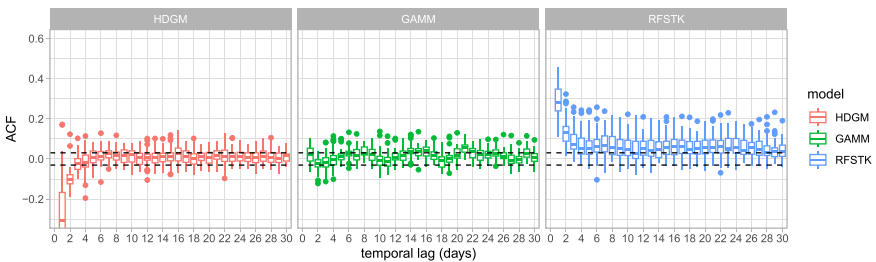
(b) Boxplots (across stations) of FM residuals' temporal correlograms for the three models.
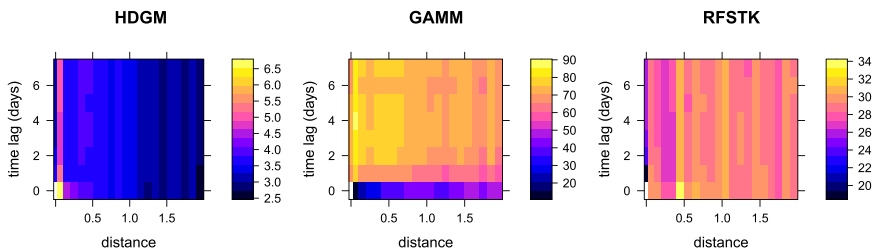
(c) Sample spatiotemporal variogram of the FM residuals of each model. Spatial distances are expressed in Earth degrees.

Fig. 4 In-sample residual diagnostics for the three models

4.2 Residual analysis

The residual distributions of the full model (FM) are symmetric and slightly lepto-kurtic, which means there is a greater chance of extreme values than a normal distribution, indicating that all models are less reliable at predicting extreme events. This is not surprising as they are designed to predict the mean level of the distribution and not for modelling the extremes.

The model uncertainties across time, shown in Fig. 4a, where the standard deviation of the full model residuals is depicted for each month, reveal that all three models have varying uncertainty throughout the year. More precisely, the PM concentrations in the winter periods could be less accurately predicted than in the summer periods. Therefore, when the models are implemented for forecasting or scenario analysis, it is recommended to use a heteroscedastic model to not underestimate the prediction accuracy in the winter months (or overestimate for the summer period). For instance, spatiotemporal stochastic volatility models could be estimated for the residual process, as demonstrated in Otto et al. (2023) for much simpler mean models. However, in this paper, our focus will be on the comparison of the mean predictions of the three models.

Furthermore, we investigated the spatial and temporal dependence of the residuals, estimating temporal autocorrelation functions (ACF) and spatiotemporal variograms. The results are shown in Fig. 4c and b. Different patterns were observed across the three models. While HDGM shows a small negative correlation at the beginning, indicating a slight overestimation of the temporal dependence, the spatial correlation is satisfactorily captured. On the contrary, GAMM leads to significantly lower temporal correlations in the residuals but does not capture the spatial dependence, as highlighted by the variogram through the bottom line for time lag 0. The RFSTK is characterised by a more pronounced positive autocorrelation for the first 3-4 lags, as shown by the correlogram (ACF), consistently with the model specification that does not consider an autoregressive term. The spatiotemporal correlation was clearly captured, as confirmed by the flat variogram. It is worth noting that the scales of the variograms differ significantly. This is because the prediction performances of the models also differ significantly in the in-sample case.

4.3 Cross-validation and comparison of predictive performance

Three factors—randomness of partition, mutual independence of test errors, and independence between the training set and test set—are crucial considerations in the context of cross-validation. We employed the leave-one-station-out cross-validation (LOSOCV) scheme, which is a variation of the commonly used leave-one-out cross-validation (LOOCV) approach applied in the spatiotemporal framework (e.g., Meyer et al. 2018; Nowak and Welsh 2020). For this method, a complete time series of a single station withheld is not used in the model's training but is used to evaluate the model's prediction performance. In this way, the validation blocks are sufficiently large not to destroy the spatiotemporal dependence. In general, utilising a LOOCV

provides an assessment of model performance but may overlook the temporal correlation of errors. By adopting a LOSOCV scheme, we gain the ability to examine the autocorrelation of errors at each station. This approach unveils distinct behaviours among stations, influenced, for instance, by factors like atmospheric stability, thereby accounting for temporal dependence. All stations within the Lombardy region were used for validation, except for the station "Moggio." It is located in the mountains with unique climatic conditions that are not well-represented by all other stations. After implementing the LOSOCV approach, we obtained prediction results from the full models for the 31 stations included in the validation process. It is worth noting that we applied the identical cross-validation scheme for each model so that the results are directly comparable.

The prediction performances assessed in the LOSOCV scheme in terms of mean squared errors (MSE), RMSE, MAE, and R^2 are summarised in Table 3. HDGM is confirmed to be the best model, but, compared to the in-sample residuals in Table 2, the uncertainty is on a realistic level with an RMSE comparable to the other models. That is, the overfit in the in-sample data did not affect the generalisation ability of the HDGM. This could be due to the linear structure in the large-scale component. While a more flexible model (e.g., random forest or artificial neural networks) could produce extremely bad predictions in areas of insufficient training data or overfitting, the linear structure of the HDGM regression term prevents us from obtaining such extreme predictions. Generally, we observe satisfactory prediction performances for all three model alternatives, with GAMM and the RFSTK approach being in second and third place, respectively. The substantial difference between the in-sample residuals from the model trained on the entire dataset and the errors from the LOSOCV scheme highlights the importance of validating prediction uncertainty through a cross-validation scheme, which accounts for the spatiotemporal nature of the data. Interestingly, the GAMM obtained a similar fit in terms of the coefficient of determination in both the in-sample case and the cross-validation. Thus, we would not overestimate the prediction capabilities when only looking at the in-sample fit. In contrast to the other models, the GAMM incorporates a relatively 'weaker' small-scale structure, capturing spatiotemporal dependence with a smooth spatial surface, serving as a relatively simple constant intercept. This design prevents overfitting and, thereby, results in similar training and testing performance.

Eventually, we compare the prediction performances for each station separately because we observed that the order of the best-fitting model is not homogeneous across space. For this reason, Fig. 5 displays the cross-validation RMSE on a map by the size of differently coloured circles. That is, the colour of the smallest circle at each station corresponds to the model with the best prediction performance,

Table 3 Prediction performance indices evaluated with the LOSOCV scheme [μg/m³]

	MSE	RMSE	MAE	R^2
HDGM	35.373	5.948	4.376	0.879
GAMM	78.042	8.834	6.239	0.733
RFSTK	53.099	7.286	5.119	0.819

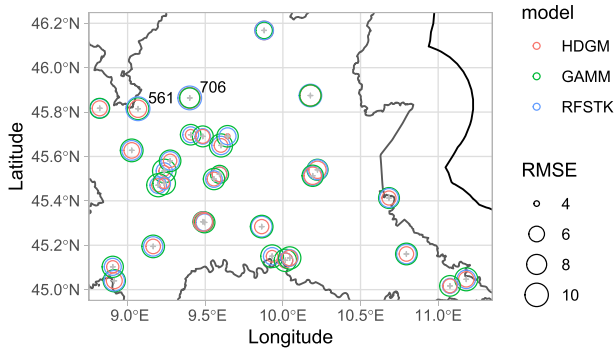**Fig. 5** Prediction performances expressed as RMSE calculated for each station within the LOSOCV scheme. The stations "Lecco - Via Sora" (ID 706) and "Como - Via Cattaneo" (ID 561) are labelled
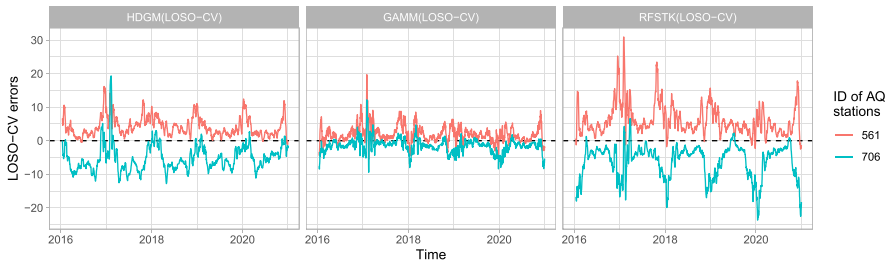

Fig. 6 15 days moving averages on the prediction errors (in μg/m^{-3}) for each model (HDGM, GAMM and RFSTK from left to right)

whereas the largest circles show the worst predictions. Below, we will discuss two selected cases with interesting behaviour: "Lecco-Via Sora" (station 706) and "Como - Via Cattaneo" (station 561). Furthermore, we depict the cross-validation prediction errors across time for these two stations in Fig. 6.

HDGM and RFSTK performed worse than GAMM at the station "Lecco - Via Sora" (station ID 706). This is because Lecco is characterised by good air quality, but its neighbouring areas are affected by high PM$_{2.5}$ concentrations. The GAMM model, which does not include strong, time-varying spatial interactions, was able to capture this difference in air quality better than the HDGM and RFSTK models. This is confirmed by the fact that the 15-day moving average of the test errors (calculated as observed minus predicted) displayed in Fig. 6 shows that both HDGM and RFSTK overestimate the PM$_{2.5}$ concentrations at this station.

At the other selected station, "Como - Via Cattaneo" (station ID 561), the RFSTK model performed the worst while HDGM showed the best performance. The reason may lie in its poor ability to capture temporal dependence well. The concentrations of PM$_{2.5}$ at this station are very stable over time, and the RFSTK model does not fully capture this stability. This is shown by the 15-day moving average of the test errors in Fig. 6, which shows that RFSTK underestimates

$PM_{2.5}$ concentrations, especially in the winter periods, when the air circulation is at its lower limit and temporal stability is the highest.

These results highlight the need to select the model according to the local conditions carefully. The best model for one location may not be the best for another. Moreover, model averaging could additionally improve the predictions.

4.4 Model interpretation

In the following three sections, we delve into the outcomes derived from our models, offering a comprehensive interpretation of each estimated model.

4.4.1 HDGM

Table 4 summarises the estimated β parameters of the large-scale component of HDGM. Except for bovine density (LI_bovine_v2), all coefficients significantly differ from zero. The signs of the majority of the coefficients are consistent with

Table 4 Estimated coefficients of the large-scale component of the HDGM model

Variable	Coefficient	Std. Err	t-value	p-value
(Intercept)	39.626	1.399	28.332	< 0.0001
February	− 7.222	1.316	5.490	< 0.0001
March	− 15.298	1.474	10.378	< 0.0001
April	− 21.569	1.552	13.898	< 0.0001
May	− 27.608	1.596	17.295	< 0.0001
June	− 27.662	1.666	16.602	< 0.0001
July	− 27.493	1.690	16.270	< 0.0001
August	− 29.316	1.681	17.436	< 0.0001
September	− 27.611	1.628	16.956	< 0.0001
October	− 20.335	1.555	13.080	< 0.0001
November	− 16.602	1.493	11.117	< 0.0001
December	− 8.526	1.349	6.318	< 0.0001
Altitude	− 0.007	< 0.001	20.180	< 0.0001
WE_wind_speed_100m_mean	− 1.946	0.048	40.520	< 0.0001
WE_tot_precipitation	− 159.416	7.145	22.310	< 0.0001
WE_temp_2m	0.505	0.037	13.595	< 0.0001
WE_rh_mean	0.186	0.008	24.188	< 0.0001
WE_blh_layer_max	− 0.003	< 0.001	19.628	< 0.0001
LI_pigs_v2	0.005	0.001	9.288	< 0.0001
LI_bovine_v2	0.003	0.002	1.147	0.251
LA_lvi	− 4.441	0.225	19.698	< 0.0001
LA_hvi	−0 .494	0.141	3.507	< 0.0001

our expectations: summer months are related to reductions of $PM_{2.5}$ concentrations (about $-27/29$ μg/m^{-3}), every 1 m/s of wind speed is related to a decrease of 2 μg/m^{-3} of $PM_{2.5}$, and every 10 mm of precipitation are related to an expected decrease of 1.6 μg/m^{-3} of $PM_{2.5}$. Moreover, each degree Celsius increase in temperature is related to an increase of 0.5 g/m^{-3} of $PM_{2.5}$, which seems counter-intuitive at first glance. However, the temperature effect should be interpreted together with the monthly fixed effects. The relative humidity is positively related to $PM_{2.5}$, so high humidity levels (100%) are associated with an increase of 18 μg/m^{-3} with respect to extremely dry air. The maximum height of the boundary layer is negatively associated with $PM_{2.5}$; every increase of 1000 m is related to an expected decrease of 3 g/m^{-3} of $PM_{2.5}$.

Regarding the agricultural impact, we observe that the number of pigs in the territory is positively associated, and an increase of 1000 animals per km^2 corresponds to an expected increase of 5   g/m^{-3} of $PM_{2.5}$. Both vegetation indices are negatively related to $PM_{2.5}$, while low vegetation (e.g. bushes) has a stronger effect than higher vegetation (e.g. trees).

The small-scale effect of the HDGM is defined by a latent variable $\xi(s,t)$ in (4), which has an autoregressive structure of order one and a Gaussian process with exponential covariance function given by (5). The spatial range parameter θ_{HDGM} describes the decay of the exponential correlation function and is estimated to be equal to 0.79°. Thus, there is a large spatial correlation (i.e., > 0.37) for surrounding stations in an area of 80 kilometres. The estimate of the temporal autoregressive parameter is equal to $\hat{g}_{HDGM} = 0.72$. This indicates that the time series has low-frequency components with relatively gradual changes over time.

4.4.2 GAMM

The estimated coefficients of our second model, the GAMM, are presented in Table 5. In the first section of the table, the estimated coefficients for the linear part of the model (including the monthly fixed effects) are reported, while the second part summarises the effective degrees of freedom of the nonlinear effects as a measure of complexity/non-linearity. Compared to HDGM, the monthly fixed effects are slightly smaller, indicating that the seasonal variation is better captured by the weather variables, which enter the model nonlinearly. Complex nonlinear relationships with large degrees of freedom characterised the smooth terms of the penalised thin plate regression splines. All of them are significant except for the density of bovine. To illustrate the difference between the linear effects in the HDGM and the nonlinear effects in GAMM, we depict some selected regression splines in Fig. 7 along with the estimated linear functions of the HDGM. These curves correspond to the marginal effect of variables neglecting the spatiotemporal correlation (i.e., without the influence of neighbouring sites). In general, we observe a similar tendency for both models. An exceptional notice would be the height of the boundary layer, which has a negative effect for up to 500 kilometres, and afterwards, the effect changes to be positive. By contrast, the effect is negative for the HDGM, which mimics the effect in the areas where most observations are located. The grey contour lines in Fig. 7 additionally illustrate the estimated kernel density of the couple

Table 5 Estimated coefficients of the large-scale component of the GAMM

A: Linear effects	Coefficient	Std. Err.	t-value	p-value
(Intercept)	37.3052	0.4541	82.1536	< 0.0001
February	− 7.2787	0.3559	− 20.4496	< 0.0001
March	− 11.6052	0.4169	− 27.8335	< 0.0001
April	− 14.6751	0.5173	− 28.3679	< 0.0001
May	− 20.5547	0.6097	− 33.7147	< 0.0001
June	− 22.0366	0.6978	− 31.5800	< 0.0001
July	− 23.2506	0.7287	− 31.9076	< 0.0001
August	− 24.5223	0.6813	− 35.9908	< 0.0001
September	− 22.7519	0.5743	− 39.6143	< 0.0001
October	− 18.9167	0.4687	− 40.3605	< 0.0001
November	− 15.8692	0.4005	− 39.6225	< 0.0001
December	− 11.4140	0.3541	− 32.2379	< 0.0001
Altitude	− 0.0023	0.0012	− 1.9378	0.0526
B: Nonlinear effects	edf	Ref.df	F-value	p-value
WE_temp_2m	8.5851	8.5851	267.3673	< 0.0001
WE_tot_precipitation	7.2891	7.2891	210.3836	< 0.0001
WE_rh_mean	8.0060	8.0060	602.4805	< 0.0001
WE_wind_speed_100m_mean	6.5219	6.5219	214.0715	< 0.0001
WE_blh_layer_max	8.8552	8.8552	278.6934	< 0.0001
LI_pigs_v2	7.9422	7.9422	7.2958	< 0.0001
LI_bovine_v2	3.7865	3.7865	1.2020	0.2962
LA_hvi	8.4121	8.4121	23.1429	< 0.0001
LA_lvi	7.4114	7.4114	82.1857	< 0.0001
Longitude, Latitude	27.7758	32.0000	19.0023	< 0.0001

Linear relationships (A) are identified by the β coefficients, while for nonlinear relationships (B), the complexity of the curve is described by the effective degrees of freedom (edf)
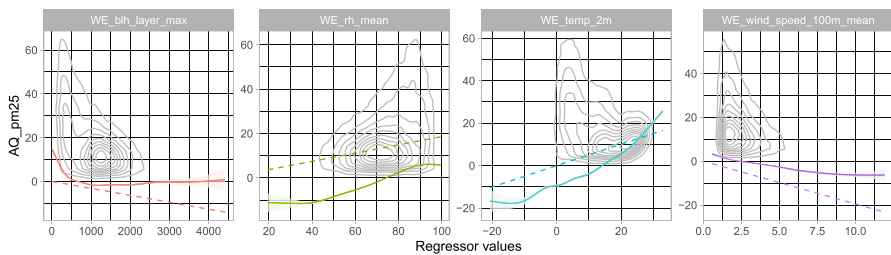

Fig. 7 Regression splines (continuous lines), including their 95% confidence intervals (coloured shadows), and HDGM regression coefficients (dashed lines) for relevant weather regressors (i.e., boundary layer height, relative humidity, temperature, and wind speed). Grey contour lines represent the estimated two-dimensional kernel densities of the PM$_{2.5}$ concentrations and the corresponding regressors
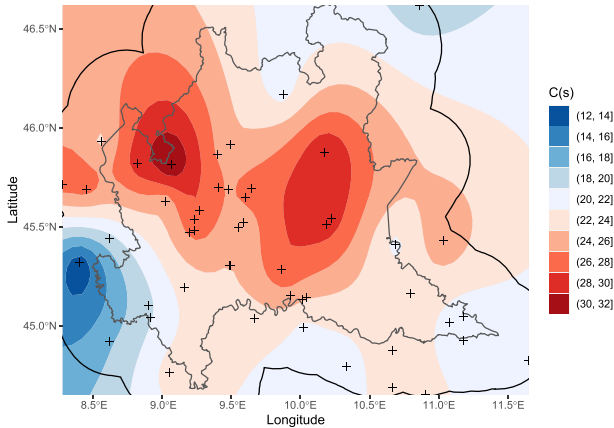
Fig. 8 Estimated smoothing spline $\hat{C}(s)$ of GAMM, which corresponds to the PM$_{2.5}$ predicted on a regular grid using the large scale of GAMM, where all regressors are set to 0. Stations are marked with a black cross, Lombardy boundaries are shown in grey, and the black line marks the surrounding buffer zone

(PM$_{2.5}$, WE_regressor). We note that the confidence intervals around the fitted curves are smaller in areas with higher density.

The GAMM smooth spatial surface $C(s)$ is displayed in Fig. 8. This smoothing spline $C(s)$ capturing the spatial dependence identifies correlated areas. Our study shows higher concentrations of PM$_{2.5}$ in the area of Como and the area of Brescia, while in the southwest, corresponding to the Ligurian border, lower concentrations. Furthermore, the estimated range parameter of the exponential covariance function is $\hat{\theta}_{GAMM} = 1.16°$, which corresponds to approximately 110 km. Hence, it is in a similar range to the other two models. Furthermore, the autoregressive parameter is estimated as $\hat{g}_{GAMM} = 0.67$, similar to HDGM and indicating a medium temporal persistence across one day. In this sense, the models show similar spatiotemporal dynamics as the HDGM.

4.4.3 RFSTK

The interpretability of the RFSTK model can be challenging due to its intricate nature. However, the variable importance factor (VIF) can help to identify the most important variables. To determine the VIF, the mean decrease accuracy technique is employed, which assesses the variables' importance by measuring the increase in MSE (IncMSE) when the values of the regressors are permuted (Breiman 2001). The results are depicted in Fig. 9.

Like the previous models, the random forest cannot fully capture the seasonality by the included (weather) regressors, as we can see by the high importance of the monthly effects. The most important variables after the month are the height of the boundary layer, the temperature and the (low) vegetation index. The bovine density is the least important variable, and indeed, it was not significantly different from zero for the other two models.
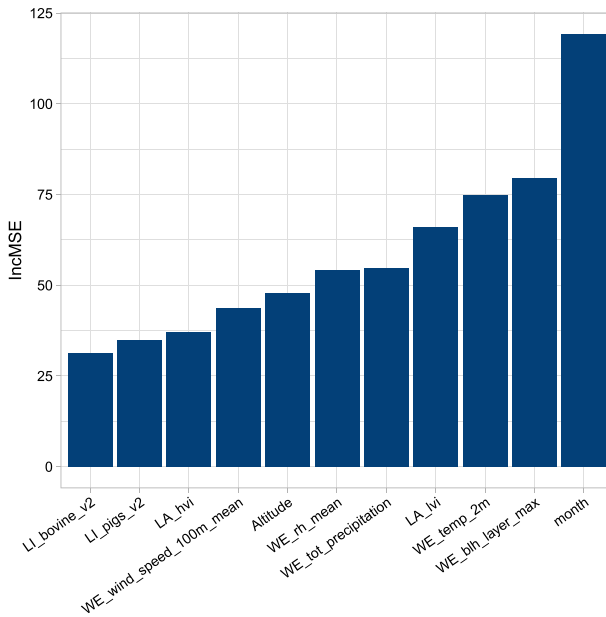
Fig. 9 Variable Importance Factor (measured as IncMSE) for the 11 selected features in the large-scale component of the RFSTK

For the RFSTK, the spatiotemporal dependence is captured by considering a spatiotemporal Gaussian process with a separable exponential covariance function for space and time. The corresponding variogram model is fitted to the RF residuals. For our analysis, we obtained exponential decay parameters of $\hat{\theta}_{RFSTK_t} = 0.78$ days and $\hat{\theta}_{RFSTK_s} = 0.48°$ for time and space, respectively, the latter corresponds to approximately 50 km. The spatial range parameters are smaller than HDGM because the large scale (RF) explains more variation than the linear large-scale term of HDGM, as shown in Tab. 2.

4.5 Model comparison

Intriguing observations arise when comparing the results of the three models. First, the temporal variation plays an essential role in all three models, as evidenced by the magnitude of the temporal dummy coefficients in HDGM and GAMM (with January as the worst month for air quality) and 'month' receiving the highest ranking in the VIF of RFSTK. This suggests that weather variables alone are insufficient for capturing all seasonal variability, even considering the most important weather variables. Interestingly, temperature, which is usually negatively associated with $PM_{2.5}$, was found to be positively related to $PM_{2.5}$ conditional on the month. This is due to the opposing effect of temperature and monthly indicator variables. Furthermore, surprisingly, the significance tests in both HDGM and GAMM on livestock densities
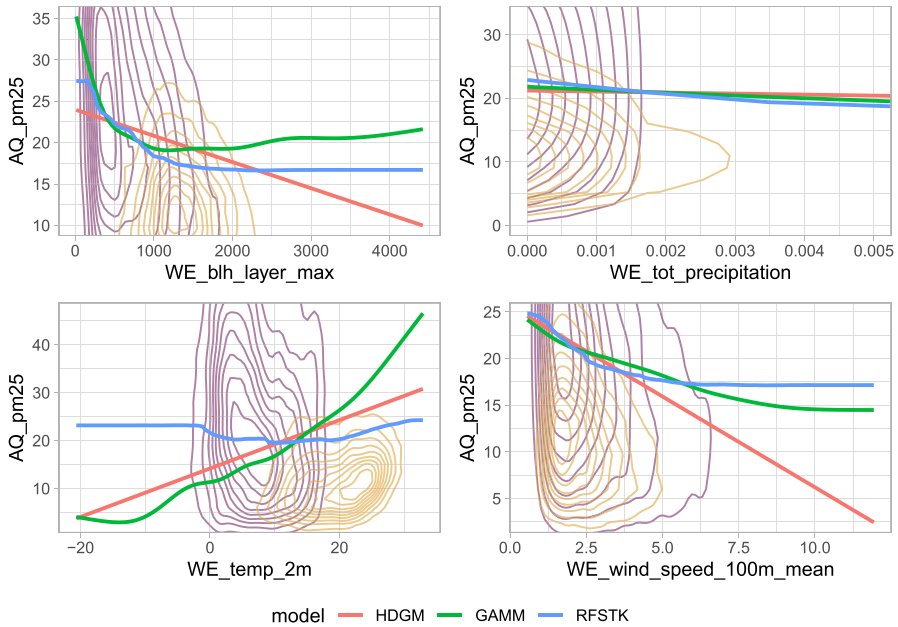
Fig. 10 PDP calculated on the large-scale of all three models. Contour lines represent kernel density estimation of the couple (PM$_{2.5}$, WE_regressor), in yellow for hot months and purple for cold months

suggested that bovine livestock density is not related to PM$_{2.5}$ concentrations, while pig density is.

In Fig. 10, we depict partial dependence plots (PDPs) (Friedman 2001) for selected covariates of all three models along with two-dimensional kernel density estimates of the pairs of PM$_{2.5}$ concentrations and the corresponding regressors in the hot (summer and spring) and cold (autumn and winter) months as yellow and purple contours, respectively. PDPs allow us to compare the relationships identified within the large-scale of each model and highlight the transition from linear to more complicated relationships. Contrary to the marginal effects, these PDPs account for the typical range of the other predictors. This is accomplished by associating a fixed value of a regressor across all observations with the mean of predicted PM$_{2.5}$ concentrations. The mean of prediction is calculated for different fixed values of the regressor, typically moving from the minimum to the maximum on an equidistant grid. In other words, PDPs show how the predicted outcome of the changes as a single predictor variable is varied while all other variables are held constant.

It is worth noting that all three models demonstrate similar trends, even though they have different levels of flexibility. For example, they all show that PM$_{2.5}$ concentrations are higher in cold periods. However, slight differences exist in the models' behaviour, especially for temperature. This suggests there may be nonlinear influences or interactions between temperature and other variables. For example, temperature may have a different impact on PM$_{2.5}$ concentrations in different altitudes or seasons. RFSTK can capture these interactions more effectively than the

other two models, which is why its PDP is flatter. This suggests that RFSTK is better at capturing the complex relationships between $PM_{2.5}$ concentrations and other variables.

This finding has important implications for the development of air quality models. It suggests that machine learning (ML) techniques can be used to improve the performance and interpretability of classic geostatistical approaches, such as HDGM or GAMM. This is because ML techniques can identify nonlinearities and interactions that are difficult to identify using traditional methods. In the second step, the more interpretable models could include the nonlinear effects and interactions, e.g. GAMM. The comparison of PDPs also highlights the complementary nature of ML techniques and classic approaches. ML techniques are better at capturing complex relationships, while traditional approaches are easier to interpret and allow for straightforward uncertainty estimation. Thus, we advertise combining ML techniques and classic approaches for modelling and predicting PM concentrations.

5 Conclusion

This study compares three statistical models to model and predict Lombardy's daily $PM_{2.5}$ concentrations and simultaneously provide an intuitive interpretation of the influencing factors. The models considered are HDGM, GAMM and RFSTK. All three models used are designed to handle spatiotemporal data, although each employs different methods to model external factors and spatiotemporal dependence. The models can generally be divided into large-scale components, small-scale spatiotemporal effects, and measurement and modelling errors. The large-scale components account for external influences, whereas the small-scale components model the spatiotemporal correlation, and the modelling errors contain the unexplained variation of the process.

The large-scale component of the three models showed significant monthly fixed effects for all three models, with negative coefficients for all months. As expected, January was confirmed to be the month with the worst air quality. Furthermore, we used partial dependence plots to compare relationships within the large-scale of the three models and highlight the transition from linear to more complex relationships. The generalised additive mixed model and the random forest approach exhibit similar patterns as they can handle nonlinear relationships. The geostatistical model is constrained by its linear specification, which fits in areas with sufficiently many observations of the covariates. At the same time, linear specification prevents making unreliable predictions, even in regions with few observations for the model estimation. Thus, the hidden dynamic geostatistical showed the best performances in the cross-validation study with an average RMSE of 6.31 μg/m^3.

By comparing marginal effects, it is possible to better understand potential interactions and nonlinearities, thereby improving the model's specification. Indeed, the HDGM can produce good results due to its ability to incorporate latent variables. Still, at the same level of predictive performances, it is preferable to have a model that explains more on a large scale, extending the degree of interpretability. Thus, it can be highly beneficial to detect nonlinear behaviours or interactions by machine

learning, which can then be used to improve the specification of the "simpler" but faster model integrated with a more efficient spatiotemporal correlation structure. ML techniques and classic statistical models can be used in complementary ways. To further improve the forecasting performance, ensemble forecasts from different statistical and ML models could be considered in future research.

The comparison of models in the field of air quality highlighted that the spatiotemporal correlation is a crucial aspect that requires careful consideration. However, this correlation is also very sensitive. If not handled properly, it can lead to overfitting the model to the specific data used, thereby hindering its ability to generalise the discovered relations. This limitation was illustrated by the discrepancy in the performance of the HDGM model when evaluated on the entire data set versus when assessed using the leave-one-station-out cross-validation approach. In our analysis, we applied similar models for the spatiotemporal interactions and small-scale effects, which are readily implemented in existing software. In future research, a more detailed comparison of the small-scale model specification across different model alternatives would be interesting.

In conclusion, our findings suggest that classic approaches, such as the hidden dynamic geostatistical model, yield the best predictive performance while being computationally efficient. However, more complex algorithms like random forest can enhance the identification of nonlinear and interaction effects. Therefore, these methods can be used complementary, ushering in a new era where newly developed techniques support traditional and well-established practices.

Author contributions All: Conceptualisation, checking empirical results, reviewing PO: supervision, writing original version, guidance empirical analysis, revisions AFM: writing original version, computations, empirical analysis, graphics, revisions JR: writing original version, computations, empirical analysis, graphics PM: computations, empirical analysis AF: project lead

Declarations

Competing interest The authors declare no competing interests

References

Baayen H, Vasishth S, Kliegl R et al (2017) The cave of shadows: addressing the human factor with generalized additive mixed models. J Mem Lang 94:206–234

Bai L, Liu Z, Wang J (2022) Novel hybrid extreme learning machine and multi-objective optimization algorithm for air pollution prediction. Appl Math Modell 106:177–198

Belgiu M, Drăguţ L (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31

Berrocal VJ, Guan Y, Muyskens A et al (2020) A comparison of statistical and machine learning methods for creating national daily maps of ambient $PM_{2.5}$ concentration. Atmos Environ 222:117,130

Bertaccini P, Dukic V, Ignaccolo R (2012) Modeling the short-term effect of traffic and meteorology on air pollution in Turin with generalized additive models. Adv Meteorol 4:8–78

Bonas M, Castruccio S (2021) Calibration of spatio-temporal forecasts from citizen science urban air pollution data with sparse recurrent neural networks. arXiv preprint arXiv:2105.02971

Bono R, Alarcón R, Blanca MJ (2021) Report quality of generalized linear mixed models in psychology: a systematic review. Front Psychol 12(666):182

Breiman L (2001) Random forests. Mach Learn 45:5–32

Calculli C, Fassò A, Finazzi F et al (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, Ignaccolo R, Bande S (2011) Comparing spatio-temporal models for particulate matter in Piemonte. Environmetrics 22(8):985–996

Cameletti M, Lindgren F, Simpson D et al (2013) Spatio-temporal modeling of particulate matter concentration through the SPDE approach. AStA Adv Stat Anal 97:109–131

Chang L, Zou T (2022) Spatio-temporal analysis of air pollution in North China Plain. Environ Ecol Stat 29(2):271–293

Chen L, Wang Y, Ren C et al (2019) Assessment of multi-wavelength SAR and multispectral instrument data for forest aboveground biomass mapping using random forest kriging. For Ecol Manage 447:12–25

Cressie N, Wikle CK (2015) Statistics for spatio-temporal data. Wiley, New York

Daw R, Wikle CK (2023) REDS: random ensemble deep spatial prediction. Environmetrics 34(1):e2780

Ding C, Wang G, Zhang X et al (2021) A hybrid CNN-LSTM model for predicting PM2.5 in Beijing based on spatiotemporal correlation. Environ Ecol Stat 28(3):503–522

European Environmental Agency (2022a) Air quality in Europe. Accessed 18 March 2023. https://doi.org/10.2800/488115

European Environmental Agency (2022b) Classification of monitoring stations and criteria to include them in EEA's assessments products. https://www.eea.europa.eu/ds_resolveuid/cb32af951deb4e40aef444bdd37d9306. Accessed 12 March 2023

Fahrmeir L, Lang S (2001) Bayesian inference for generalized additive mixed models based on Markov random field priors. J R Stat Soc Ser C 50(2):201–220

Fahrmeir L, Kneib T, Lang S (2004) Penalized structured additive regression for space-time data: a Bayesian perspective. Stat Sin 89:731–761

Fasiolo M, Nedellec R, Goude Y et al (2020) Scalable visualization methods for modern generalized additive models. J Comput Graph Stat 29(1):78–86

Fassò A, Finazzi F (2011) Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. Environmetrics 22(6):735–748

Fassò A, Maranzano P, Otto P (2022) Spatiotemporal variable selection and air quality impact assessment of COVID-19 lockdown. Spat Stat 49(100):549

Fassò A, Rodeschini J, Fusta Moro A et al (2023) Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region Italy. Sci Data 10(1):143

Fassò A, Rodeschini J, Fusta Moro A et al (2023) AgrImOnIA: Open Access dataset correlating livestock and air quality in the Lombardy region. Italy. https://doi.org/10.5281/zenodo.7956006

Finazzi F, Scott EM, Fassò A (2013) A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. J R Stat Soc Ser C 62(2):287

Fioravanti G, Martino S, Cameletti M et al (2021) Spatio-temporal modelling of PM_{10} daily concentrations in Italy using the SPDE approach. Atmos Environ 248(118):192

Fox EW, Ver Hoef JM, Olsen AR (2020) Comparing spatial regression to random forests for large environmental data sets. PLoS ONE 15(3):e0229,509

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 45:1189–1232

Fusta Moro A, Salis M, Andrea Z et al (2022) Ammonia emissions and fine particulate matter: some evidence in Lombardy. In: Book of Short Papers of the ASA Conference 2022-Data-Driven Decision Making, pp 1–6

Grange SK, Sintermann J, Hueglin C (2023d) Meteorologically normalised long-term trends of atmospheric ammonia (NH3) in Switzerland/Liechtenstein and the explanatory role of gas-aerosol partitioning. Sci Total Environ 900:165844

Gräler B, Pebesma E, Heuvelink G (2016) Spatio-temporal interpolation using gstat. R J 8:204–218

Guo PT, Li MF, Luo W et al (2015) Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. Geoderma 237:49–59

Handcock MS, Wallis JR (1994) An approach to statistical spatial-temporal modeling of meteorological fields. J Am Stat Assoc 89(426):368–378

Hastie T, Tibshirani R (1987) Generalized additive models: some applications. J Am Stat Assoc 82(398):371–386

Huang HC, Cressie N (1996) Spatio-temporal prediction of snow water equivalent using the Kalman filter. Comput Stat Data Anal 22(2):159–175

Ignaccolo R, Mateu J, Giraldo R (2014) Kriging with external drift for functional data for air quality monitoring. Stoch Environ Res Risk Assess 28:1171–1186

Jato-Espino D, Mayor-Vitoria F (2023) A statistical and machine learning methodology to model rural depopulation risk and explore its attenuation through agricultural land use management. Appl Geogr 152(102):870

Javanmard ME, Ghaderi S (2022) A hybrid model with applying machine learning algorithms and optimization model to forecast greenhouse gas emissions with energy market data. Sustain Cities Soc 82(103):886

Jerrett M, Arain A, Kanaroglou P et al (2005) A review and evaluation of intraurban air pollution exposure models. J Expo Sci Environ Epidemiol 15(2):185–204

Jiang J, Nguyen T (2007) Linear and generalized linear mixed models and their applications, vol 1. Springer, New York

Jui SJJ, Ahmed AM, Bose A et al (2022) Spatiotemporal hybrid random forest model for tea yield prediction using satellite-derived variables. Remote Sens 14(3):805

Kammann E, Wand MP (2003) Geoadditive models. J R Stat Soc Ser C 52(1):1–18

Kishi S, Sun J, Kawaguchi A et al (2023) Characteristic features of statistical models and machine learning methods derived from pest and disease monitoring datasets. R Soc Open Sci 10(6):230,079

Knape J (2016) Decomposing trends in Swedish bird populations using generalized additive mixed models. J Appl Ecol 53(6):1852–1861

Kneib T, Knauer F, Küchenhoff H (2011) A general approach to the analysis of habitat selection. Environ Ecol Stat 18:1–25

Kulkarni P, Sreekanth V, Upadhya AR et al (2022) Which model to choose? Performance comparison of statistical and machine learning models in predicting $PM_{2.5}$ from high-resolution satellite aerosol optical depth. Atmos Environ 282:119,164

Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22

Liu Y, Cao G, Zhao N et al (2018) Improve ground-level $PM_{2.5}$ concentration mapping using a random forests-based geostatistical approach. Environ Pollut 235:272–282

Liu Y, Zhao N, Vanos JK et al (2019) Revisiting the estimations of $PM_{2.5}$-attributable mortality with advancements in $PM_{2.5}$ mapping and mortality statistics. Sci Total Environ 666:499–507

Lovarelli D, Conti C, Finzi A et al (2020) Describing the trend of ammonia, particulate matter and nitrogen oxides: the role of livestock activities in northern Italy during Covid-19 quarantine. Environ Res 191(110):048

Lu M, Cavieres J, Moraga P (2023) A comparison spatial and nonspatial methods in statistical modeling of NO_2: Prediction accuracy, uncertainty quantification, and model interpretation. Geograph Anal 4:89–96

Maranzano P, Pelagatti M (2023) Spatiotemporal event studies for environmental data under cross-sectional dependence: an application to air quality assessment in Lombardy. J Agric Biol Environ Stat 89:1–22

Maranzano P, Otto P, Fassò A (2023) Adaptive LASSO estimation for functional hidden dynamic geosta-tistical models. Stochastic Environmental Research and Risk Assessment pp 1–23

Merk MS, Otto P (2020) Estimation of anisotropic, time-varying spatial spillovers of fine particulate matter due to wind direction. Geograph Anal 52(2):254–277

Meyer H, Reudenbach C, Hengl T et al (2018) Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ Modell Softw 101:1–9

Nag P, Sun Y, Reich BJ (2023) Spatio-temporal DeepKriging for interpolation and probabilistic forecasting. Spat Stat p 100773. https://doi.org/10.1016/j.spasta.2023.100773, https://www.sciencedirect.com/science/article/pii/S2211675323000489

Najafabadi AM, Mahaki B, Hajizadeh Y (2020) Spatiotemporal modeling of airborne fine particulate matter distribution in Isfahan. Int J Environ Health Eng 9(July):1–7

Nowak G, Welsh A (2020) Improved prediction for a spatio-temporal model. Environ Ecol Stat 27:631–648

Otto P, Piter A, Gijsman R (2021) Statistical analysis of beach profiles: a spatiotemporal functional approach. Coast Eng 170(103):999

Otto P, Doğan O, Taşpınar S (2023) A dynamic spatiotemporal stochastic volatility model with an application to environmental risks. Econom Stat. https://doi.org/10.1016/j.ecosta.2023.11.002

Patelli L, Cameletti M, Golini N, et al. (2023) A path in regression random forest looking for spatial dependence: a taxonomy and a systematic review. arXiv:2303.04693

Pernigotti D, Georgieva E, Thunis P et al (2012) Impact of meteorological modelling on air quality: summer and winter episodes in the Po valley (Northern Italy). Int J Environ Pollut 50(1–4):111–119

Piter A, Otto P, Alkhatib H (2022) The Helsinki bike-sharing system-insights gained from a spatiotemporal functional model. J R Stat Soc Ser A 185(3):1294–1318

Qi Y (2012) Random forest for bioinformatics. Ensemble Mach Learn 89:307–323

Sekulić A, Kilibarda M, Heuvelink GB et al (2020) Random forest spatial interpolation. Remote Sens. 12(10):1687

Shao Y, Ma Z, Wang J et al (2020) Estimating daily ground-level PM2.5 in China with random-forest-based spatiotemporal kriging. Sci. Total Environ. 740:139,761

Stafoggia M, Bellander T, Bucci S et al (2019) Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environ Int 124:170–179

Strobl C, Boulesteix AL, Kneib T et al (2008) Conditional variable importance for random forests. BMC Bioinform 9:1–11

Sun W, Xu Z (2022) A hybrid daily PM2.5 concentration prediction model based on secondary decomposition algorithm, mode recombination technique and deep learning. Stoch Environ Res Risk Assess 4:1–20

Taghavi-Shahri SM, Fassò A, Mahaki B et al (2020) Concurrent spatiotemporal daily land use regression modeling and missing data imputation of fine particulate matter using distributed space-time expectation maximization. Atmos Environ 224(117):202

Thunis P, Clappier A, Beekmann M et al (2021) Non-linear response of $PM_{2.5}$ to changes in NO_x and NH_3 emissions in the Po Basin (Italy): consequences for air quality plans. Atmos Chem Phys 21(12):9309–9309. https://doi.org/10.5194/acp-21-9309-2021

Tsokov S, Lazarova M, Aleksieva-Petrova A (2022) A hybrid spatiotemporal deep model based on CNN and LSTM for air pollution prediction. Sustainability 14(9):5104

Viscarra Rossel RA, Webster R, Kidd D (2014) Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging. Earth Surf Process Landf 39(6):735–748

Wang L, Wu W, Liu HB (2019) Digital mapping of topsoil ph by random forest with residual kriging (RFRK) in a hilly region. Soil Res 57(4):387–396

Wang Y, Finazzi F, Fassò A (2021) D-STEM v2: a software for modeling functional spatio-temporal data. J Stat Softw 99:1–29

WHO (2021) WHO global air quality guidelines: particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary

Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc B 73(1):3–36

Wood SN (2017) Generalized additive models: an introduction with R. Chapman and Hall/CRC, Boca Raton

Zhan Y, Luo Y, Deng X et al (2018) Satellite-based estimates of daily NO2 exposure in China using hybrid random forest and spatiotemporal kriging model. Environ Sci Technol 52(7):4180–4189

Zhang L, Cheng Y, Zhang Y et al (2017) Impact of air humidity fluctuation on the rise of PM mass concentration based on the high-resolution monitoring data. Aerosol Air Quality Res 17(2):543–552

Authors and Affiliations

Philipp Otto[1] · Alessandro Fusta Moro[2] · Jacopo Rodeschini[2] · Qendrim Shaboviq[3] · Rosaria Ignaccolo[4] · Natalia Golini[4] · Michela Cameletti[2] · Paolo Maranzano[5,6] · Francesco Finazzi[2] · Alessandro Fassò[2]

✉ Philipp Otto
  philipp.otto@glasgow.ac.uk

  Alessandro Fusta Moro
  alessandro.fustamoro@gmail.com

  Jacopo Rodeschini
  jacopo.rodeschini@unibg.it

  Qendrim Shaboviq
  qendrim.shaboviq@ikg.uni-hannover.de

  Rosaria Ignaccolo
  rosaria.ignaccolo@unito.it

  Natalia Golini
  natalia.golini@unito.it

  Michela Cameletti
  michela.cameletti@unibg.it

  Paolo Maranzano
  paolo.maranzano@unimib.it

  Francesco Finazzi
  francesco.finazzi@unibg.it

  Alessandro Fassò
  alessandro.fasso@unibg.it

[1]   School of Mathematics and Statistics, University of Glasgow, University Place, Glasgow G12 8QQ, UK

[2]   Department of Economics, University of Bergamo, Via dei Caniana, 2, 24127 Bergamo, Italy

[3]   Institute of Cartography and Geoinformatics, Leibniz University Hannover, Hannover, Germany

[4]   Department of Economics and Statistics "Cognetti de Martiis", University of Turin, Lungo Dora Siena, 100A, 10153 Torino, Italy

[5]   Department of Economics, Management and Statistics, University of Milan-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy

[6]   Fondazione Eni Enrico Mattei (FEEM), Corso Magenta, 63, 20123 Milan, Italy