




# A Review of Air Quality Modeling

K. Karroum<sup>1,4\*</sup> , Y. Lin<sup>2</sup>, Y.-Y. Chiang<sup>2</sup>, Y. Ben Maissa<sup>3</sup>, M. El Haziti<sup>5</sup>, A. Sokolov<sup>4</sup>  
and H. Delbarre<sup>4</sup>

<sup>1</sup>LRIT-CNRST, URAC 29, Rabat IT Center, Faculty of Sciences, Mohammed V University in Rabat, BP 1014 Rabat, Morocco

<sup>2</sup>Spatial Sciences Institute, University of Southern California, Los Angeles, CA, USA

<sup>3</sup>Laboratory of Telecommunications, Networks and Service Systems, National Institute of Posts and Telecommunications, Allal El Fassi Avenue, Rabat, Morocco

<sup>4</sup>Laboratoire de Physico-Chimie de l'Atmosphère, Université du Littoral Côte d'Opale, 59140 Dunkirk, France

<sup>5</sup>Higher School of Technology, BP 227 Sale, Morocco

Received: 31 December 2018 / Accepted: 10 March 2020 / Published online: 20 March 2020

© Metrology Society of India 2020

**Abstract:** Air quality models (AQMs) are useful for studying various types of air pollutions and provide the possibility to reveal the contributors of air pollutants. Existing AQMs have been used in many scenarios having a variety of goals, e.g., focusing on some study areas and specific spatial units. Previous AQM reviews typically cover one of the forming elements of AQMs. In this review, we identify the role and relevance of every component for building AQMs, including (1) the existing techniques for building AQMs, (2) how the availability of the various types of datasets affects the performance, and (3) common validation methods. We present recommendations for building an AQM depending on the goal and the available datasets, pointing out their limitations and potentials. Based on more than 40 works on air quality, we concluded that the main utilized methods in air pollution estimation are land-use regression (LUR), machine learning, and hybrid methods. In addition, when incorporating LUR methods with traffic variables, it gives promising results; however, when using kriging or inverse distance weighting techniques, the monitoring stations measurements of air pollution data are enough to have good results. We aim to provide a short manual for people who want to build an AQM given the constraints at hands such as the availability of datasets and technical/computing resources.

**Keywords:** Air pollution; Air quality models (AQMs); Techniques and validation; Datasets; AQMs recommendations

## 1. Introduction

The long-term exposure to air pollution not only causes deterioration of the respiratory system [1, 2] but also increases the risk of cardiovascular and atherosclerosis diseases [3, 4]. Poor air quality might adversely affect cognition, lead to mental illness such as dementia [5] and others [6], and cause preterm labor and birth [7]. Furthermore, ignoring the principal causes of air pollution or underestimating pollution sources could lead to inaccurate results, such as areas near airports where the air pollution emission can be more serious [8]. Therefore, air quality prediction and forecast are still open, significant, and

challenging tasks. With a growing number of residents living in urban areas, the major concerns become (1) how to identify the primary sources that lead to the air pollution, (2) how to quantify the impacts of these sources on air quality, and (3) how to reduce the threats of air pollution to the human health as well as the environment.

There exist many reviews on air quality modeling (AQM), covering a variety of topics. Ryan et al. [9] summarize the history and applications of land-use regression (LUR) models, which have been widely used to characterize the intra-urban air pollution exposure. The authors also discuss the similarities and the differences in the variables among the six studies. Hoek et al. analyze 25 LUR cases and point out that LUR has a better performance than other traditional techniques, such as kriging and dispersion models in [10]. They also propose a possible

\*Corresponding author, E-mail: khaoula.ari@gmail.com

improvement of LUR, like enabling its transferability to different areas and including more predictors in a LUR model. In contrast to the two aforementioned reviews of AQMs, Zhang et al. focus on real-time air quality forecasting (RT-AQF) [11, 12]. The authors discuss the historical milestones and accessibility of primary existing techniques of RT-AQF in review part I [11], and the possible ways of improving the accuracies of RT-AQF models as well as the challenges that limit their performance and the prospects in part II [12].

This paper aims to provide an extensive literature review based on a multitude of works on air quality prediction and forecasting, pointing out the diverse elements of air quality modeling as follows:

1. reporting the performance and strengths of the existing air quality predicting and forecasting methods;
2. discussing the role of data on these models, including datasets accessibility (free, for a fee, or restricted), types (directly measured or surrogated), and sources (e.g., traffic data and land-use data) since the data availability is typically the most critical aspect when selecting an applicable air quality model;
3. summarizing common ways of assessing and validating air quality models;

To the best of our knowledge, there is not yet a review covering all the aspects mentioned above. The remainder of the paper is organized as follows. Section 2 categorizes the existing techniques for air quality predicting and forecasting, which makes the selection of technical methods easier, based on the intent to build an AQM. Section 3 discusses how data quality and availability play a crucial role in selecting an air quality modeling method. Section 4 presents the common methods of AQMs' validation. Section 5 makes recommendations on AQMs based on the available data inputs and the intended goals. Section 6 presents an example of a real case study of PM<sub>10</sub> concentrations estimation using three techniques: nearest neighbor, IDW, and kriging. Finally, Sect. 7 discusses the review and provides a general conclusion of the work.

## 2. Air Quality Predicting/Forecasting Techniques

Having a clear idea (given in this Sect. 2) about the techniques of modeling and analyzing air pollutants can lead to a better interpretation of AQM results, help select a proper air quality model, and also identify the primary contributions that intensify air pollution proportion. Although various techniques have been developed to predict or forecast air pollutants, it is essential to choose an appropriate method based on the goal of the application (e.g., predicting/forecasting air pollution, revealing air pollution

contributors, etc.) and the data (e.g., the network density of monitoring stations and land-use information). For example, if we have a dense network or aim at adding a monitoring site to the network, the dispersion modeling (e.g., CALINE: California Line Source Dispersion Model) is the right fit, due to its easy adaptation to new pollutants or geographic areas without adding more monitoring sites in the studied area.

This section classifies AQM techniques into four categories: land-use regression (LUR), machine learning, hybrid techniques, and other techniques that are less frequent in the articles we review.

### 2.1. Land-Use Regression (LUR)

Land-use regression (LUR) is a technique that develops stochastic models to predict air pollutant concentrations at a given site by utilizing the predictor variables (e.g., the surrounding land use, road network, traffic, physical environment, and population) based on geographic information systems (GIS) and the monitoring of air pollutant measurements. The SAVIAH (Small Area Variations in Air quality and Health) project first studies to model small-scale variations of air pollutants by LUR [13] and demonstrates that the GIS-based regression mapping is a robust tool when predicting air quality at a fine-spatial resolution with limited monitoring data. When combined with effective strategies, LUR can be used to explain the air pollution conditions (e.g., seasonal variations in human activities: population density is a significant air pollution source only in winter, inversely in summer where industrial indicators are more influencing in this season) and to reveal the main causes of air pollution [14]. Besides predicting air pollutant concentrations, LUR can also infer the influence of the surrounding environment. Chen et al. reveal that the exposure to heavy traffic might affect human cognition, by utilizing the satellite-based image of PM<sub>2.5</sub> measurements with other variants (e.g., road length, age, and sex) as inputs of the LUR model [5]. The authors find that living close to a major road might cause the increased incidence of dementia. The “right choice” of variables plays an important role in building LUR models. Ross et al. affirm in [15] that with the traffic information and land-use variables, the LUR model reaches more than 60% of the explained variation in PM<sub>2.5</sub> concentrations over a wide area for three different periods of the year and counties. Also, adding more relevant variables in LUR model would affect performance positively. Moreover, Ross et al. raise the explained variation of air pollution from 54 to 79% [16] with a traffic variable (traffic within 300 m buffer to the monitoring stations) as the addition input. ADDRESS (A Distance Decay REgression Selection Strategy) is a strategy to select optimized buffer distances for potential

predictors to maximize model performance. ADDRESS models traffic-related air pollutant in Los Angeles, an extraordinarily complex cityscape. By computing the correlation coefficients of spatial covariates (commercial, residential, industrial, and open land-use data) with residuals of exposure concentrations, the model creates a series of distance decay. This strategy enhances the traditional LUR with normally distributed prediction and accuracy varying between 87 and 91% [17].

LUR performance can also be enhanced by coupling it with some air pollution developed approaches. Lee et al. [18] use the ESCAPE (the European Study of Cohorts for Air Pollution Effects) modeling approach with LUR and show its effectiveness in spatial variation explanation, even when it comes to a high density of traffic roads and population area like Taipei city. ESCAPE is a study of long-term air pollution exposures effecting on human health for 15 European countries. The authors apply this approach in developing LUR models of Taipei which select the top relevant inputs to take in account in the final model, by conducting multiple analyses on the intended predictors to get a model with better accuracy. Furthermore, the LUR-ESCAPE-estimated exposures have a wider scope of pollutants variability and provide results with better spatial resolution compared to the two classic spatial interpolation algorithms, i.e., ordinary kriging and the nearest neighbor (e.g., measurement site) methods. The results could be helpful in assessing the influence of long-term exposure to nitrogen dioxide (NO<sub>2</sub>) and nitrogen oxides (NO<sub>x</sub>) on the epidemiological cohorts in the Taipei Metropolis.

LUR method strongly relies on the availability of spatial data (land uses) without considering the spatial effects, such as spatial non-stationarity and spatial autocorrelation, that limit the LUR performance by reducing the prediction accuracy and increasing uncertainty. Bertazzon et al. develop a wind-LUR model [14], which is a variant of LUR model including wind as a relevant meteorological variable, which could alleviate the spatial non-stationarity and spatial autocorrelation problems. In this work, the authors presented an alternative model of two models, i.e., spatially autoregressive model (SAR) which solves spatial non-stationarity and geographically weighted regression model (GWR) that deals with spatial autocorrelation. They substitute SAR and GWR by one single LUR model which is mathematically simpler and outperforms the traditional LUR with an improvement of 10–20% in mean  $R^2$ .

Up to this day, LUR is still showing its powerful ability in air quality prediction. Taking Houston Metropolitan Area, USA, as an example, Zhai et al. prove how the challenge of the spatial scale should be further investigated: The impact of a predictor in a specific distance/radius within a study area does not have the same effect in a different study area (spatial non-stationarity) [19]. The

authors affirm that the need for a clear understanding of the physical–chemical dispersion mechanisms is not always an obligation in every AQM development. This LUR model achieves a mean error rate (MER) under 20% with the best  $R^2 = 0.78$ , the smallest MER = 11.84%, and the lowest root mean square error (RMSE) = 1.43 and outperforms the ordinary kriging by using variables at the optimized spatial scales.

## 2.2. Machine Learning

Machine learning is powerful at predicting unknown values by building models with data, based on computer science and statistical techniques. For example, Basu et al. develop an algorithm that identifies interactions in a system by using iterative random forests that could be applied to many scientific fields [20]. Several studies have used various machine learning algorithms (e.g., neural network, random forest (RF), and regression) to model air quality due to their promising performance for more than a decade. For instance, the artificial neural network (ANN) approach was used back to the year of 2003 for particulate matter (PM<sub>2.5</sub>) prediction [21].

By comparing the advantages and limitations of three neural network algorithms in predicting the air pollutants: multilayer perceptron neural network (MLP), square multilayer perceptron (SMLP), and radial basis function network (RBF), Ordieres et al. conclude that RBF is the best predictor among these three algorithms, outperforming the other two by shorter training times and better stability (the independency of estimation variability on the used training data). In general, ANN is still a good option for air quality prediction that achieves better results than the other classical models, like persistence (which is a simple model supposing that the pollutant concentration level at a specific time corresponds to the value that occurred the same time the day before  $y_t = x_t$ ) and linear regression models, as claimed in [21].

Besides, Xu et al. [22] propose a support vector regression (SVR)-based bi-dimensional exploration framework to predict PM<sub>2.5</sub> in Beijing in 2014. This model considers the time-lagged PM<sub>2.5</sub> time series from surrounding monitoring stations to show how the PM<sub>2.5</sub> concentrations disperse spatially and temporally. This study deploys SVM (support vector machine) from Weka (Waikato Environment for Knowledge Analysis: an easy free tool for machine learning and data mining employment) and finds out that with the increase in the geographical scope and time lag, the prediction errors decreases, but the performance improvement decreases as well. Despite this constraint, the model achieved a good balance between performance and modeling cost.

Jiang et al. report in [23] that regression trees are useful for predicting air quality as well. They detect the outdoor air pollution based on the messages shared on social media, Weibo (Chinese Twitter) and utilize a classifier to distinguish the air pollution levels in Beijing based on the netizens' posts of the city on Weibo. The authors predict the air quality index (AQI) in Beijing using gradient tree boost (GTB), which iteratively builds a regression tree from residuals and outputs weighted sum of the regression trees. By the mean of the multi-additive function of GTB, they develop a successful monitoring and tracking model for air pollutant prediction, using media data classification (to classify whether a posted message by a netizen is a negative or positive one, and get by the end a classification of all the netizens' messages from "excellent" to "serious pollution" categories. In this way, the social media data could be compared to the real measurements of AQI and multi-step filtering to take into account only social media data about outdoor air pollution on Beijing region (discard social media data that do not concern outdoor air pollution, are advertising messages, and those which are by netizens out of the studied region).

Some studies work on improving the performance of an existing model by adding some other relevant variables or proposing an improved version of the model to make it more accurate. By using the results of WRF-Chem (Weather Research and Forecasting (WRF) model coupled with Chemistry [24] as inputs in addition to the air pollutant measurements, Xi et al. [25] design a comprehensive evaluation framework to improve the prediction performance. They test five different machine learning algorithms on four different groups of datasets where each group includes different input features (e.g., pollution observation, weather forecast, wind speed, etc.), and the RF was the best performer with most of the groups. This combination strategy leads to a 3% improvement of the single model which is not incorporated with WRF-Chem data. Furthermore, the authors conclude that the availability of more information increases the possibility to enhance the model accuracy.

RF is an ensemble learning method for classification (and regression) done by the mean of generating classifiers as random trees and then assembling them by an aggregator. Supported by the bootstrap aggregating, RF builds a set of decision trees that contribute to predicting the air quality index (AQI) for Shenyang city, and the aggregating of all these trees results provides the AQI classification [26]. In this paper, RF shows good results when compared to three other algorithms in sensing urban air quality. As a result, this model proves an overall precision of 81% for AQI prediction, and since all the used data are from Internet, it is possible to apply this method to other cities as well. Brokamp et al. report that when combining random

forest with variables that have a significant impact on air pollution as land-use variables (LURF), it becomes possible to cover the limitations of LUR in capturing nonlinear relationships and complex interactions between predictors and the outcome with a small-size training data. This AQM shows better results than LUR with a decrease in a fractional predictive error of at least 5% in most of the studied pollutants elemental components (such as aluminum, copper, and iron) and a cross-validated fractional predictive error less than 30%, with the help of the diverse inputs [27].

### 2.3. Hybrid

In this paper, we chose to give the description of "hybrid technique" to any work that adopts more than one algorithm category (e.g., machine learning, geo-statistic, and land-use regression) to develop an air quality model. This is the case for most of the spatiotemporal AQMs that study not only the spatial aspect of air pollution but the temporal one also, each one by a different method (see below).

Hybrid models usually consist of two or more practical algorithms, which come out with a stronger air quality model, providing better results than using just one single method. For instance, Wilton et al. [28] work with the dispersion model CALINE3 [29] for roadway pollution prediction from the meteorological dispersion model and then performed a LUR model using simultaneous measurements over space for improving spatial concentrations estimates. This hybrid model achieves an improvement in  $R^2$  values for both cities Seattle and LA. In addition, the authors capture a greater amount of the pollutant variation (the near-road gradients) than in the traditional LUR models, since they include roadway lengths and traffic density as predictors.

Most hybrid techniques model the spatial and temporal aspects separately, which allows specific processing for each, and then aggregate the results of both. Zheng et al. try to infer the real-time air quality by defining two separated classifiers: spatial and temporal classifiers. The spatial classifier uses an ANN for modeling the spatial correlation between air qualities of different locations by taking the spatially related features (e.g., the density of POIs and length of highways) as inputs. The temporal classifier uses a linear-chain conditional random field (CRF) to represent the temporal dependency of air quality at a site by taking the temporally-related features (e.g., traffic and meteorology) as predictors. This model outperforms other four standard well-known methods on five Chinese datasets [30]. Another example of using neural network is that Zheng et al. [31] forecast the  $PM_{2.5}$  concentrations in the next 48 h at a monitoring site by aggregating the spatial model (based on ANN) and temporal classifiers (based on

linear regression) with a dynamic aggregator that integrates the spatial and temporal results in a way that uses the meteorological data as a reference of accordance. For every station, meteorological information will be taken into account such as wind speed, wind direction, weather state (foggy/sunny/etc.), etc., to determine a weight for each classifier. Also, the authors create a predictor that detects the sudden sharp changes in the air. They verify the model performance using 43 cities in China, and the results outperformed all the other baseline models such as autoregressive moving average (ARMA), linear regression, and regression tree. The accuracy was 75% in the first six hours, and it remains good even when sudden changes in air quality occur.

To study each of the spatial variability and temporal variability independently and make use of the spatiotemporal variables, Li et al. [32] deal with more than one effect of air pollution (e.g., spatial effect, temporal effect, local effect, etc.) to develop a spatiotemporal model to predict nitrogen oxides at a high spatiotemporal resolution, incorporating nonlinear and spatial effects. The authors create a constrained nonlinear mixed-effect model with ensemble learning. Their approach integrates nonlinear relationships, fixed and random effects from the predictors by expressing the spatiotemporal variability of concentrations with mixed-effect models. Then, they perform an ensemble learning of all these models and carry out a constrained optimization that copes with the constraint of locations with large temporally incomplete data, by the mean of minimizing the difference between the concentrations to adjust its corresponding prediction output. In addition, Li et al. utilize the dispersion model CALINE4 to estimate the mean (temporal average) of air pollutant on roads. This approach reduces variance and enhances the reliability of prediction, by improving the results from initial mixed effects (that do not incorporate ensemble learning and the proposed constrained optimization) with  $R^2$  values equal to 0.85 and 0.86 for nitrogen dioxide ( $\text{NO}_2$ ) and nitrogen oxides ( $\text{NO}_x$ ), respectively.

#### 2.4. Others

In this section, we focus on all remaining AQM techniques that are important but less frequently used than the categories discussed above.

The first example is geo-statistical techniques, which incorporate statistics to analyze the spatiotemporal variation of the air pollutants. Fontes et al. [33] perform interpolation of air quality for the urban sensitive area of region Oporto/Asprela and showed that inverse distance weighting (IDW) is a better interpolator than kriging for this studied region. Ramos develop in [34] a technique by combining IDW and kriging with a well-selected set of relevant

variables. The authors show that the hybrid model outperforms the use of each method separately. Geo-statistical methods could be better than other techniques as Rivera-González et al. [35] show by selecting ordinary kriging as the best performer among the six tested methods. Ordinary kriging, in this case, is powerful not only due to its excellent performance on air pollutants concentrations prediction, but also because it computes the corresponding standard error (estimation variance) of the prediction.

Another useful technique is the regression method based on SRS (satellite remote sensing) as applied by Guo et al. in the work of [36], where they predict ground-level  $\text{PM}_{2.5}$  depending only on PARASOL level 2 AOD modeled by four different empirical models: the linear regression model, the quadratic regression model, the power regression model, and the logarithmic regression model. All of them show reasonably good results but underestimate the  $\text{PM}_{2.5}$  concentrations compared to the ground-level  $\text{PM}_{2.5}$  concentrations.

The stochastic models represent a mean of air quality analysis as well. Sun et al. [37] use the uncommon hidden Markov models (HMMs) and try to represent the hidden layer by a non-Gaussian distribution. In the interest of improving HMM, the authors implement three different emission distribution functions: log-normal, gamma, and generalized extreme value (GEV) to predict  $\text{PM}_{2.5}$  concentrations. Consequently, the true prediction rate could be improved by these three non-Gaussian distributions to 150%, and more importantly, false alarms (that alert when the air quality exceeds a defined index of pollution) could be reduced by 78%. Yet another statistical model that gives good results in adjusting the raw model bias [38] is Kalman filter (KF) predictor approach. KF serves as a bias adjustment tool, increasing the value of  $R^2$  from 0.43 for the raw model forecasts to 0.90 for the KF bias-adjusted forecasts at more than 90% of the studied sites. This model can make the correlation coefficients with measurements higher, besides the methodology is easily adapted for real-time applications.

Some air quality modeling studies calibrate the AQMs just before carrying out the prediction to enhance the performance, and it is done by applying some conditions while developing the model. This calibration's efficiency is validated afterward by one of the parameters mentioned later in this Sect. 4. For example, Li et al. [32] impose a degree of freedom (10) for the explanatory variables to decrease the model's overfitting and get better results, while the studies [17, 18] select only the predictors with a  $p$  value greater than 0.1. Brokamp et al. [27] perform the same selection for their inputs, in addition to a parameter of variance inflation factors (VIFs) that should be less than three to keep a variable as predictor in the model to improve the model's  $R^2$ . In several of the reviewed LUR

AQMs in this manuscript, they adopt a stepwise approach to determine the most relevant inputs to consider in the model (e.g., [15, 18, 39, 40], etc.) and to get rid of all of the unnecessary predictors.

### 3. Dataset Analysis

Finding available datasets is the first step when selecting an appropriate technique for building AQMs. One needs to consider the input datasets to the model as well as the intent of building this model (either it is finding the air pollution contribution for health studies [5] or determining the air quality in green spaces [41], etc.). In this section, we describe commonly used datasets in the reviewed articles in sections above, with indicating when possible, references to free and paid datasets that could be utilized to build an AQM.

#### 3.1. Accessibility

Data availability is one of the most critical factors for building a good air quality model [22, 25, 42] since incorporating more information (inputs) usually can improve the accuracy of AQMs. In this paper, we classify the datasets in the reviewed articles by their accessibility into three categories: free (publicly available), paid, and unauthorized (restricted/difficult to obtain).

##### 3.1.1. Free, paid, and unauthorized datasets

Many studies work with free datasets that are available online. In the case of Houston Metropolitan Area Texas, USA, Zhai et al. use publicly available datasets (pollutant concentrations, land use/cover, road network, and census data) and they provide links in their article for accessing these data [16]. Yu et al. [26] predict AQIs for Shenyang city, China, using open data by getting traffic and road information from Baidu and Google maps. With the available air quality data of Ciudad Juarez and El Paso-Mexico, Ordieres et al. determine the PM<sub>2.5</sub> concentrations for the remaining 16 h of the day [21]. Lin et al. extract geographic features from OpenStreetMap, a crowdsourced world map, for building the air quality prediction model [43] and the forecasting model [44]. Leveraging open-source data enables the models to be generalizable to other study areas. Furthermore, sometimes data are available even for a large number of cities as in [22]; the authors predict air pollution for 190 Chinese cities based on open data. The case of Vienna, Austria [45], is a similar case as well, where the volunteered geographic information (VGI) serves in generating land-use patterns, without any remote sensing techniques or official data. However, some data are

claimed to be free, but we could not access them due to invalid links [35, 39] (checked 14/03/2018).

Air quality modeling datasets are not always reachable due to the state of property (data rights), only the members of an organization (e.g., laboratory or university) can have access to the data. Ramos et al. [34] work on the Canadian city Calgary air quality, the dataset of this city has information which are: public (traffic volume data, census of population, and industrial point information), dedicated to the members of the University of Waterloo only (road traffic and land use), and information that they got from the National Surveillance Air Pollution Surveillance of Canada (air quality data, wind speed, and direction information). In the case of Beijing and Shanghai real-time air quality prediction, Zheng et al. use GPS trajectories from a large number of taxis that they gather themselves [30]. In some regions, even the pollutant concentrations are not available. Fontes and Barros held a campaign to measure the pollutant concentrations for the urbanized region of Asprela in Oporto by themselves, and these measures are not published nor publicly shared [33].

The last data type is the paid data. For instance, TeleAtlas is a company that provides digital maps information like road network data [28, 46]. These paid data help in improving AQM performance by incorporating it with the other existing data. Yang et al. combine SRS images which are from paid sources with ground-based measurements, and their model works well for the case of NO<sub>2</sub> pollutant [39].

We sum up useful links to free and paid data in Table 1.

#### 3.2. The availability and quality of data

To date, the lack of data hampers the development of air quality models as it directly influences the selection of significant variables in explaining the air pollution variation. For example, the geographical data of the studied area are necessary variables in the model because they describe the topology of the region (e.g., the high vegetation covering areas would have totally different pollutant concentration value from the roadway or tar roofed building) [17, 18, 39]. Also, traffic-related information [47], meteorological data [16], the number of the network monitoring stations (known for being the main obstacle because of its scarcity in almost all the AQMs works) [15, 18, 19, 30, 33, 34, 39], and other information [32, 35, 42] are required in modeling air quality. Lack of data can lead to inaccurate results. For instance, when the number of monitor stations is low, interpolation methods would not perform well due to the small number of measurements in contrast to when monitoring stations are dense [44, 48–50]. Some AQMs use the surrogate or supplemental variables to cover data scarcity. For instance,

**Table 1** Useful references to datasets (checked 03/04/2018)

Data type	References	Data	References/links
Paid (software)	[17]	Traffic, Road network	TeleAtlas: <a href="http://www.tele-mart.com">http://www.tele-mart.com</a>
	[36, 38]	PM <sub>2.5</sub> ground level	MODIS: <a href="https://modis.gsfc.nasa.gov">https://modis.gsfc.nasa.gov</a>
	[46]	Road network	TeleAtlas
		Traffic	ESRI: <a href="https://www.esri.com/fr-fr/home">https://www.esri.com/fr-fr/home</a>
	[13]	All dataset	ARC/INFO
	[32]	Traffic density	ESRI, ArcGIS: <a href="https://www.esri.com/fr-fr/arcgis/products/arcgis-pro">https://www.esri.com/fr-fr/arcgis/products/arcgis-pro</a>
		Distance to roadway	ESRI
Free		Population density	ArcGIS
	[15]	Land-use data	USGS: <a href="https://www.usgs.gov">https://www.usgs.gov</a>
		Pollutant measurements	EPA: <a href="https://www.epa.gov/outdoor-air-quality-data">https://www.epa.gov/outdoor-air-quality-data</a>
		Population density	U.S. Census Bureau: <a href="http://www.census.gov/main/www/access.html">http://www.census.gov/main/www/access.html</a>
		Road network	ESRI: <a href="http://www.openstreetmap.org/#map=5/51.500/-0.100">http://www.openstreetmap.org/#map=5/51.500/-0.100</a>
	[26]	Weather data	<a href="https://rp5.ru/Météo_Monde:Weather%20for%20243%20Countries%20of%20the%20World">https://rp5.ru/Météo_Monde:Weather for 243 Countries of the World</a>
		POI (point of interest)	Google maps: <a href="https://www.google.com/maps">https://www.google.com/maps</a>
		Traffic and Road data	Baidu maps: <a href="https://map.baidu.com">https://map.baidu.com</a>
	[34]	Hourly data of PM <sub>2.5</sub>	NAPS CANADA: <a href="http://maps-cartes.ec.gc.ca/rmspa-naps/data.aspx">http://maps-cartes.ec.gc.ca/rmspa-naps/data.aspx</a>
		Meteorological measurements	National Climatic Data and Information Archive of Environment Canada: <a href="http://climate.weather.gc.ca/index_e.html">http://climate.weather.gc.ca/index_e.html</a>
	Smog info	Environnement et Changement climatique Canada: <a href="http://www.ec.gc.ca/infosmog/default.asp?lang%80=%80En&amp;n=669E620B-1">http://www.ec.gc.ca/infosmog/default.asp?lang%80=%80En&amp;n=669E620B-1</a>	
	Population density	Statistics Canada: <a href="http://www12.statcan.ca/censusrecensement/2006/ref/dict/geo021-eng.cfm">http://www12.statcan.ca/censusrecensement/2006/ref/dict/geo021-eng.cfm</a>	
[43]	Geographic data	OSM: <a href="https://www.openstreetmap.org/">https://www.openstreetmap.org/</a>	
[47]	Vehicles data	<a href="https://pubs.acs.org">https://pubs.acs.org</a>	

including a significant indicator of air pollution as wind information [18, 47] shows good improvement on the traditional model and considers that meteorological variable is the missing and needed information to enhance the model. Surrogate variables help in many cases to cover the lack of data by estimating real variables from other inputs like satellite data. Due to the lack of geographic variables, Su et al. [17] use ETM + remote sensing data to cover the lack of datasets and obtain more accurate prediction result with data of greenness or soil brightness. SRS (satellite remote sensing) data show the ability in explaining the spatial variation of NO<sub>2</sub> in Pearl River Delta region China by replacing the missing industrial, geographic, and socioeconomic data [39]. Images can be an alternative to other essential data like when Yu et al. try to obtain the road length and traffic congestion status by studying and inducing it from images of public map service providers [26]. However, sometimes even surrogate variables are still in need of adjustment when they are coarse [39]; otherwise, it would be useless to work with.

Meteorology variables (such as wind speed, wind direction, and temperature) are also indicators of air

quality. Their significant role has been discussed in several studies, knowing that meteorological elements are key factors that affect air pollution behaviors such as pollutants' emissions, transport, and transformation. For instance, Seo et al. [51] examine the influence of meteorology on long-term air quality measurements and observed changes in PM<sub>10</sub> and O<sub>3</sub> that were related to the meteorology trends. They induce that the long-term increase over the decade of 2002–2012 in wind speed leads to an improvement in air quality (in addition to the applied emission control policies), causing a ventilation of pollutants.

Moreover, Kamińska et al. [52] analyze air pollution effects using meteorological conditions along with traffic information of Wroclaw city in Poland and found out that the meteorological parameters such as wind speed are the most important impacts in modeling PM<sub>2.5</sub>, besides traffic flow for NO<sub>x</sub>. In another study of air quality interpolation, meteorological data are integrated as predictors in addition to a set of information, and Le et al. [53] use deep learning model to estimate and predict air pollution over Seoul city. The authors test the model performance with different

input combinations (with only air pollution data, air pollution and meteorological data, air pollution and traffic volume data, air pollution and vehicles average speed data, air pollution and external air pollution data, and air pollution and all related factors). The best RMSE for interpolating and forecasting is the one of air pollution and meteorological data inputs, even better than the one with all factors included, deducing that meteorological parameters have the most significant role compared to the remaining ones.

By dint of having a relevant impact on air quality, the meteorological air quality influencers are now even studied in a more detailed way. Xie et al. [54] investigate the role of different wind field types on different pollutants in Pearl River Delta region. The authors find out that  $PM_{2.5}$ ,  $PM_{10}$  and  $NO_2$  spatial distributions depend on the wind field patterns. The air quality was changing following the characteristics of each type of wind fields.

Therefore, the role of meteorology parameters in studying and analyzing (e.g., [55]) the air quality is proved to be very important, as well as in controlling and improving the latter (e.g., [56]).

#### 4. AQMs' Validation

There are several validation methods to evaluate the performance of an AQM. In this section, we choose to discuss

popular ones, with identifying the used metrics in the discussed work in the second chapter (Table 2).

Every developed system, service, or model needs to be validated to see whether it responds to the intended objective. The validation step has a relevant role in evaluating the AQM's performance that could not be in no case ideal for these two reasons [57]:

- Observations express single realizations from an infinite ensemble of cases under the same conditions, while air quality models estimate ensemble means.
- Different sources are responsible for the model predictions' uncertainties, like random turbulence of the atmospheric layer, input data errors, or uncertainties in model physics [58–61].

The efficient way of evaluating air quality models is to carry out a statistical performance analysis, which needs to call out for specific measures. The reason for building an air quality model defines the parameters of validation to use, based on the studied case circumstances and conditions. When the AQM is for assessing the health state, the validation parameter to use should look for the most correlated causes (model inputs) with the health deterioration, which is different from when evaluating whether a prediction is good or not; here, we look for the most precise model possible.

Different applications/technologies for AQMs require different validation parameters to evaluate the performance in various ways, and it is by reason that there is not a

**Table 2** Examples of used validation parameters in air pollution

References	Validation parameters
[17]	Variance inflation factors (VIFs), Cook's distances, Moran's I statistic, Chow's test, normalized mean bias (NMB), normalized mean error (NME), and cross-validation (CV)
[40]	The best fitness and the highest adjusted $R^2$
[18]	Leave-one-out cross-validation (LOOCV), $R^2$ , and $R^2$
[39]	$R^2$ , RMSE, LOOCV, and regional cross-validation (RCV)
[15]	CV, RMSE, and mean absolute percentage error (MAPE)
[19]	CV, $R^2$ , mean error rate (MER), and RMSE
[27]	LOOCV
[21]	RMSE, $R^2$ , and mean absolute error (MAE)
[23]	Correlation with real measurements
[26]	Precision, recall, $F$ -score, relative absolute error (RAE), and CV
[28]	CV and $R^2$
[32]	LOOCV
[30]	Recall and precision
[38]	$R^2$ , RMSE, NME, mean bias (MB), NMB, and R
[42]	CV, RMSE, and R
[34]	LOOCV, RMSE, and $R^2$
[35]	CV and RMSE
[72]	$R^2$



general measure suitable to all cases under any conditions [62].

#### 4.1. The Most Used Parameters and Reasons

We do not give an exhaustive description of all the possible parameters of air quality evaluation in this review, but we introduce the most common ones in the studies discussed earlier:

Root mean square error (RMSE): Adopted as the decisive criteria of the model performance in many air pollution studies, RMSE expresses the difference between values predicted by a model and the values actually observed by the following formula in (1):

$$\text{RMSE} = \sqrt{\sum_{i=1}^n \frac{(X_{\text{obs},i} - X_{\text{model},i})^2}{n}} \quad (1)$$

$R^2$ , denoted by  $R^2$  or  $r^2$ , is the coefficient of determination representing the proportion of the variance in the dependent variable that is predictable from the independent variables, summed up by the explained variation/total variation.  $R$ , correlation coefficient denoted as  $R$  or  $r$ , is used in many works to measure the correlation among the variables (input datasets) and between the observed and modeled values as well, to see how much they correlate. It varies between  $+1$  and  $-1$ , where  $1$  is a total positive linear correlation as in (2):

$$r = \frac{\sum_{i=1}^n (x_{\text{obs},i} - x_{\text{predic},i}) \cdot (y_{\text{obs},i} - y_{\text{predic},i})}{\sqrt{\sum_{i=1}^n (x_{\text{obs},i} - x_{\text{predic},i})^2 \cdot \sum_{i=1}^n (y_{\text{obs},i} - y_{\text{predic},i})^2}} \quad (2)$$

$p$  value is the correlation between variables could be evaluated by  $p$  value, which we compare to the significance level (usually represented by  $\alpha = 0.05$ ). In case  $p > \alpha$ , then the correlation is different from 0. If not, then it is impossible to conclude that the correlation is different from 0.

Cross-validation is one of the most common techniques for assessing the variation of model's prediction performance. By dividing the dataset randomly into  $X$  folds, the model would be refitted  $X$  times with the set of each fold removed in turn from the training set, knowing that part of the data is for fitting the different models and the remainder data for measuring the predictive performance of the model by the validation errors (that could be done by one of the discussed measures above). By the end, the model with the best performance is adopted [63]. Cross-validation is a good solution for detecting and preventing overfitting problem as well.

Table 2 presents the utilized parameters of validation in the previously discussed articles.

AQMs can also be evaluated with respect to spatial or temporal coverage of the validation data. For example, LUR models are mostly used for spatial analysis rather than temporal dimension. The spatial scale can differ from each other, e.g., it can be a city [14, 16, 18], a number of cities [15, 38], or even a whole state [19]. Besides, some studies that deal with spatiotemporal prediction can have diverse spatial and temporal scales. For instance, biweekly predictions for different sub-counties of California [32], daily prediction for Mexico border region [21], 74 cities in china [25], Montreal city [34], and Mexico city [35], hourly prediction for [22, 26, 36, 42], and even a real-time prediction in [30].

## 5. AQMs Recommendation Based on Inputs

The purpose of air quality modeling is not only getting the most accurate prediction/forecasting results of air pollution but also detecting the main contributors to the air quality. There are some techniques of AQM studying the cause-and-effect aspect between air pollution and environment, helping to draw inferences about the air pollution prime causes. The two cause-and-effect models we encountered (which give an explanation of the existing effects by finding the causes) are LUR and random forests, which provide the access to the most significant contributors on air quality. In this section, we recommend the AQM to adopt following the available inputs, in other words, the possible techniques to employ depending on the available variables in the dataset.

### 5.1. Recommendations for Building AQMs

Based on several works on air quality modeling, we build a set of recommendations that help when developing an AQM, to choose suitable inputs for a certain method and vice versa.

We recommend using traffic indicators with LUR along with pollutants' concentrations, due to the good performance it shows in various studies. For LUR models, in most cases the traffic-related predictors are the most influencing input data among the given variables, in addition to the land-use variable-related information [39]. In the case of Ross et al.'s study [16], only the traffic information accounts for over 54% of the variation. Besides, Lee et al. find that the length of major roads, urban green areas, semi-natural, and forested areas are the most significant predictors [18]. Even when using three different models of LUR as in [15], where the inputs period's measurements are different and the used variables too (a model with 28 counties for the period of (1999–2001), the second for the same period but only for 9 counties, and the

last one for 2000 winter for 28 counties), traffic explains the greatest part of variance (37–44%) in all the models, followed by the population density indicator. Moreover, when predicting air pollution in Los Angeles [17], Su et al. use several variables as inputs of the LUR model, and since the studied area is near roadway, evidently the impact of local traffic is the most significant one, ignoring all the remaining contributors. Taking the seasonal criteria into consideration, traffic is identified to be the most significant feature in summer and population density in winter for  $\text{NO}_2$  also [18]. In addition, Moore et al. prove that traffic volume is one of the top affecting factors along with industrial and government areas [46]. Zhai et al. report that traffic is still the stronger factor along with land-use variables and compared to the population distribution and distance to the coast, and this is due to the great urbanization and intensive road traffic in Houston, USA [19].

Random forest is a good choice for predicting air quality when having urban sensing data, such as point of interests (POI), surrogate data from public map providers, and other relevant information as discussed in [26]. Another case when to use this technique is with land-use indicators, Brokamp et al. take land-use indicators as predictors into random forests. Random forest on air quality modeling could be inferred with amazingly high accuracy from these datasets [27].

The social media data can provide useful information for estimating air quality since citizens always like to express their opinions about air pollution in their city through social media. Machine learning is recommended in this case, for example gradient tree boosting (GTB) [23] that solves classification and regression problems. The benefit of social media data is that they help to get the air pollution level at unmonitored locations especially in large cities as in the work of [64], which all utilize machine learning to measure the air pollution.

Fontes et al. do not include in [33] predictors like meteorological data, traffic information, distance to the ocean, and industrial emissions, the only available inputs they have are the pollutants concentrations. In this situation, kriging and IDW techniques can estimate the air pollutants by interpolating the measurements from the known monitoring stations. So, even in the situation where we have only pollutants' concentrations, the air pollution prediction is still possible by the mean of kriging and IDW methods. Other works prove the same thing and were able to perform air pollution estimation with achieving good results [65–68].

Satellite remote sensing data can predict the air quality index; hence, in [42] Guo et al. are able to give reasonably good results by performing regression algorithms using satellite data. This type of data covers the limitation of ground-level monitoring in spatial coverage and resolution

and gives promising results when performed by regression models [69–71].

$\text{PM}_{2.5}$  and  $\text{NO}_x$  are often selected for studying air quality (the most examined pollutants in the reviewed articles), because  $\text{PM}_{2.5}$  is one of the most dangerous pollutants on the human health and environment and  $\text{NO}_x$  is a good tracer of traffic-related pollution. We advise to use machine-learning-based AQMs for analyzing  $\text{PM}_{2.5}$  and  $\text{PM}_{10}$ , while LUR (e.g., [19, 39, 40]) and hybrid models (e.g., [28, 31, 32]) are usually used to model  $\text{PM}_{2.5}$  and  $\text{NO}_x$  concentrations. To deal with traffic-related pollutants such as gases of  $\text{CO}_2$  [72] and  $\text{NO}_x$  [47] and particles such as particle-bound polycyclic aromatic hydrocarbons (PB-PAH), particle number count (PNC) [47], and UFP (ultrafine particles) [73], it is preferable to utilize mathematical models to predict these pollutants' concentrations in roads.

## 6. Case study

In this section, we present a real case study of  $\text{PM}_{10}$  estimation in Northern France region. Based on the given recommendations in Sect. 5, we adopt IDW, kriging, and nearest neighbor since we only have the pollution data and compare their performance estimating  $\text{PM}_{10}$  concentrations.

### 6.1. Study area and data

The study area is North France region which has 6 million inhabitants and a population density of 189 inhabitants/ $\text{km}^2$ , on January 1, 2014. It is the third most populous region in France and the second most densely populated region in metropolitan France, after Ile-de-France. Covering an area of 32,000  $\text{km}^2$ , that represents 5.7% of the surface area of metropolitan France, the North France region is bordered on the North by the North Sea for a distance of 45 km and on the West by the Channel for a distance of 120 km. The region is exposed to a temperate, oceanic climate. It has cool, wet winters, and mild summers. It gathers many industrial and agricultural activities, fishing ports, passenger transport, and significant roads and sea traffic. It is located in the center of northern Europe and the Paris–Brussels–London triangle [74].

The inputs data are  $\text{PM}_{10}$  concentrations and the coordinates of the 12 sites that provide the  $\text{PM}_{10}$  measurement (blue dots in Fig. 1). The  $\text{PM}_{10}$  observations are measured every 15 minutes from January 1, 2013, to December 31, 2013. These measurements were provided by ATMO Hauts-de-France [75] and are not available online. The 12 sites were classified in a previous thesis to three categories [76], following their position characteristics as:



**Fig. 1** The positions of measured  $PM_{10}$  by 12 stations in North France region

- Continental stations representing the stations localized in the urban area which are: Campagne-lès-Boulonnais (RU1), Saint-Omer (SO1), Béthune Stade (BE2), Armentières (MO1), Lille Fives (MC5), and Valenciennes Acacias (VA1).
- Coastal stations far from the industry zone we have in Dunkerque city which are: Calais Berthelot (CA8), Calais Parmentier (CA9), and Malo-les-Bains (DK4).
- Coastal stations near the industries of Dunkerque city represented by Gravelines PC-Drire (DKG), Mardyck (DKC), and Saint-Pol-sur-Mer (DK7).

Due to the malfunctions occurred in the measuring sensors, the original  $PM_{10}$  observations contain some outliers (values that are bigger than the possible actual values of  $PM_{10}$  concentrations) and negative values. Therefore, we did some data preprocessing to remove the outliers and negative values from the inputs. To assess the performance of the three methods in predicting  $PM_{10}$  concentrations over the whole presented region, we computed for each RMSE and  $R^2$  by LOOCV (explained in Sect. 4).

## 6.2. Results and discussion

Table 3 presents the RMSE and  $R^2$  results of the three methods. We observe that IDW gives the smallest RMSE

of  $9.45 \mu\text{g}/\text{m}^3$  and the highest  $R^2$  of 67%, followed by kriging and nearest neighbor. However, there is no big difference between IDW and kriging, but a strong gap compared to nearest neighbor, even if this latter takes in account in its estimation process just the nearest site or the average of the surrounding sites, while the others consider all the sites' information. This is because the nearest neighbor does not consider any spatial variance and the number of observed sites is limited and they are far from each other. Thus, according to the discussion in the "Dataset analysis" section (Sect. 3), when having small number of monitoring sites, which are not equally distributed over the space, advanced (e.g., kriging) and simple techniques (nearest neighbor) lead to comparable results. Also, the study region is exposed to industrial emission sources (in Dunkirk city) plus meteorological phenomena (at the coastal part), which makes it more difficult for these methods to estimate the  $PM_{10}$  concentrations correctly.

## 7. Discussion and Conclusion

Air pollution problems could be alleviated by building urban parks that have a remarkable impact on reducing air pollutants [77]. However, setting up of trees and urban parks in dense cities is constrained by the size and location,

**Table 3** Nearest neighbor, IDW, and kriging results in estimating 2013 PM<sub>10</sub> of Northern France

Interpolation method	RMSE $\mu\text{g}/\text{m}^3$	$R^2$
Nearest neighbor	11.88	0.49
IDW	9.45	0.67
Kriging	9.68	0.66

like for large dense, polluted cities a big relative urban park is needed. Lam et al. suggest a possible way of improving the urban environment, by cleaning the air and reducing the noise with a good design and arrangement of green spaces before their establishment [41]. Liu et al. conduct a case study where they addressed the outcome of green space changes on air pollution and microclimates, via structural equation modeling. The authors indicate that the changing pattern of green space areas has a great influence in diminishing air pollution, making rainfall patterns smaller and cooling temperatures [78]. In addition, the project of [79] helps people who aim at studying air pollution or mitigating the human impact on the planet based on AoT (Array of Things), by giving a detailed explanation of how the Internet of Things (IoT) could serve as an instrument for research and development across many disciplines, including air pollution.

However, without knowing the current state of air pollution, we would not know how to act toward poor air quality. Therefore, air quality modeling becomes a necessity for air quality analysis. AQMs have been applied successfully for studying and analyzing air quality as well as its contributions. The models could be developed based on various possible techniques and dataset. Although many limitations would affect the model's performances (e.g., lack or low quality of datasets), there are some alternatives and efficient ways to build a more accurate model (e.g., surrogate variables or hybridizing techniques). Since there exist various possible options for building AQMs, this review could serve as a first manual for selecting datasets and techniques, which covers the essential elements of AQMs: existing techniques, dataset types, and validation methods, with emphasizing the limitations and strengths of AQMs. Modeling air quality could be carried out via a multitude of available methods but going back to the purpose of creating an AQM helps and determines the techniques to utilize. Through this paper, we present a sort of guideline to anyone interested in elaborating an AQM by detailing every element of this latter.

**Acknowledgements** This work was supported by the scholarship of Excellence from the National Center for Scientific and Technical Research (CNRST) of Morocco and Université du Littoral Côte d'Opale of Dunkerque, France. We would like to thank Atmo Hauts-de-France for providing us the measurements used in the study case.

### Compliance with Ethical Standards

**Conflict of interest** The authors declare that they have no conflict of interest.

### References

- [1] M. Kampa and E. Castanas, Human health effects of air pollution. *Environ. Pollut.* **151**(2) (2008) 362-367.
- [2] B. Karimi and S Samadi, Mortality and hospitalizations due to cardiovascular and respiratory diseases associated with air pollution in Iran: a systematic review and meta-analysis. *Atmos. Environ.* **198** (2019) 438-447.
- [3] N. Künzli, M. Jerrett, W.J. Mack, B. Beckerman, L. Labree, F. Gillil, D. Thomas, J. Peters and H.N. Hodis, Ambient air pollution and atherosclerosis in Los Angeles. *Environ Health Perspect.* **113** (2005) 201-206.
- [4] D.E. Schraufnagel, J.R. Balmes, C.T. Cowl, S. De Matteis, S.H. Jung, K. Mortimer and G.D. Thurston, Air pollution and non-communicable diseases: a review by the Forum of International Respiratory Societies' Environmental Committee, Part 2: air pollution and organ systems. *Chest*, **155**(2) (2019) 417-426.
- [5] H. Chen, J. Kwong, R. Copes, K. Tu, A. van Donkelaar, P. Hystad, P. Villeneuve, R. Martin, B. Murray, B. Jessiman and A. Kopp, Living near major roads and the incidence of dementia, Parkinson's disease and multiple sclerosis in Ontario, Canada: population-based study. In *ISEE conference abstracts* (2016).
- [6] J.G. Miller, J.S. Gillette, E.M. Manczak, K. Kircanski and I.H. Gotlib, Fine particle air pollution and physiological reactivity to social stress in adolescence: the moderating role of anxiety and depression. *Psychosom. Med.* **81** (2019) 641-648.
- [7] F. Vadillo-Ortega, A. Osornio-Vargas, M.A. Buxton, B.N. Sánchez, L. Rojas-Bracho, M. Viveros-Alcaráz, M. Castillo-Castrejon, J. Beltrán-Montoya, D.G. Brown and M.S. O'Neill, Air pollution, inflammation and preterm birth: a potential mechanistic link. *Med. Hypotheses*, **82**(2) (2014) pp. 219-224.
- [8] N. Hudda and S.A. Fruin, International airport impacts to air quality: size and related properties of large increases in ultrafine particle number concentrations. *Environ. Sci. Technol.* **50**(7) (2016) 3362-3370.
- [9] P.H. Ryan and G.K. LeMasters, A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation Toxicol.* **19**(sup1) (2007) 127-133.
- [10] G. Hoek, R. Beelen, K. De Hoogh, D. Vienneau, J. Gulliver, P. Fischer and D Briggs, A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **42**(33) (2008) 7561-7578.
- [11] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur and A. Baklanov, Real-time air quality forecasting, part I: History, techniques, and current status. *Atmos. Environ.* **60** (2012) 632-655.
- [12] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur and A. Baklanov, Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmos. Environ.* **60** (2012) 656-676.

- [13] D.J. Briggs, S. Collins, P. Elliott, P. Fischer, S. Kingham, E. Lebrecht, K. Pryl, H. Van Reeuwijk, K. Smallbone and A. Van Der Veen, Mapping urban air pollution using GIS: a regression-based approach. *Int. J. Geogr. Inf. Sci.* **11**(7) (1997) 699-718.
- [14] S. Bertazzon, M. Johnson, K. Eccles and G.G. Kaplan, Accounting for spatial effects in land use regression for urban air pollution modeling. *Spatial Spatio-Temporal Epidemiol.* **14** (2015) 9-21.
- [15] Z. Ross, M. Jerrett, K. Ito, B. Tempalski and G.D. Thurston, A land use regression for predicting fine particulate matter concentrations in the New York City region. *Atmospheric Environ.* **41**(11) (2007) 2255-2269.
- [16] Z. Ross, P.B. English, R. Scaif, R. Gunier, S. Smorodinsky, S. Wall and M. Jerrett, Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. *J. Exposure Sci. Environ. Epidemiol.* **16**(2) (2006) 106.
- [17] J.G. Su, M. Jerrett, B. Beckerman, M. Wilhelm, J.K. Ghosh and B. Ritz, Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy. *Environ. Res.* **109**(6) (2009) 657-670.
- [18] J.H. Lee, C.F. Wu, G. Hoek, K. de Hoogh, R. Beelen, B. Brunekreef and C.C. Chan, Land use regression models for estimating individual NO<sub>x</sub> and NO<sub>2</sub> exposures in a metropolis with a high density of traffic roads and population. *Sci. Total Environ.* **472** (2014) 1163-1171.
- [19] L. Zhai, B. Zou, X. Fang, Y. Luo, N. Wan and S. Li, Land use regression modeling of PM<sub>2.5</sub> concentrations at optimized spatial scales. *Atmosphere* **8**(1) (2016) 1.
- [20] S. Basu, K. Kumbier, J.B. Brown and B. Yu, Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci.* (2018) 201711236.
- [21] J.B. Ordieres, E.P. Vergara, R.S. Capuz and R.E. Salazar, Neural network prediction model for fine particulate matter (PM<sub>2.5</sub>) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). *Environ. Model. Softw.* **20**(5) (2005) 547-559.
- [22] W. Xu, C. Cheng, D. Guo, X. Chen, H. Yuan, R. Yang and Y. Liu, PM<sub>2.5</sub> Air Quality Index Prediction Using an Ensemble Learning Model. In *International Conference on Web-Age Information Management* (2014) (pp. 119-129). Springer, Cham.
- [23] W. Jiang, Y. Wang, M.H. Tsou and X. Fu, Using social media to detect outdoor air pollution and monitor air quality index (AQI): a geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter). *PLoS One* **10**(10) (2015) e0141185.
- [24] G.A. Grell, S.E. Peckham, R. Schmitz, S.A. McKeen, G. Frost, W.C. Skamarock and B. Eder, Fully coupled "online" chemistry within the WRF model. *Atmos. Environ.* **39**(37) (2005) 6957-6975.
- [25] X. Xi, Z. Wei, R. Xiaoguang, W. Yijie, B. Xinxin, Y. Wenjun and D. Jin, A comprehensive evaluation of air pollution prediction improvement by a machine learning method. In *2015 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)* (2015) (pp. 176-181). IEEE.
- [26] R. Yu, Y. Yang, L. Yang, G. Han and O.A. Move, Raq—a random forest approach for predicting air quality in urban sensing systems. *Sensors* **16**(1) (2016) 86.
- [27] C. Brokamp, R. Jandarov, M.B. Rao, G. LeMasters and P. Ryan, Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmos. Environ.* **151** (2017) 1-11.
- [28] D. Wilton, A. Szpiro, T. Gould and T. Larson, Improving spatial concentration estimates for nitrogen oxides using a hybrid meteorological dispersion/land use regression model in Los Angeles, CA and Seattle, WA. *Sci. Total Environ.* **408**(5) (2010) 1120-1130.
- [29] P.E. Benson, A review of the development and application of the CALINE3 and 4 models. *Atmos. Environ. Part B. Urban Atmos.* **26**(3) (1992) 379-390.
- [30] Y. Zheng, F. Liu and H.P. Hsieh, U-Air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (2013) (pp. 1436-1444). ACM.
- [31] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang and T. Li, Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (2015) (pp. 2267-2276). ACM.
- [32] L. Li, F. Lurmann, R. Habre, R. Urman, E. Rappaport, B. Ritz, J.C. Chen, F.D. Gilliland and J. Wu, Constrained mixed-effect models with ensemble learning for prediction of nitrogen oxides concentrations at high spatiotemporal resolution. *Environ. Sci. Technol.* **51**(17) (2017) 9920-9929.
- [33] T. Fontes and N. Barros, Interpolation of air quality monitoring data in an urban sensitive area (2010).
- [34] Y. Ramos, B. St-Onge, J.P. Blanchet and A. Smargiassi, Spatio-temporal models to estimate daily concentrations of fine particulate matter in Montreal: Kriging with external drift and inverse distance-weighted approaches. *J. Exposure Sci. Environ. Epidemiol.* **26**(4) (2016), 405.
- [35] L.O. Rivera-González, Z. Zhang, B.N. Sánchez, K. Zhang, D.G. Brown, L. Rojas-Bracho, A. Osornio-Vargas, F. Vellido-Ortega and M.S. O'Neill, An assessment of air pollutant exposure methods in Mexico City, Mexico. *J. Air Waste Manag. Assoc.* **65**(5) (2015) 581-591.
- [36] Y. Guo, N. Feng, S.A. Christopher, P. Kang, F.B. Zhan and S. Hong, Satellite remote sensing of fine particulate matter (PM<sub>2.5</sub>) air quality over Beijing using MODIS. *Int. J. Remote Sens.* **35**(17) (2014) 6522-6544.
- [37] W. Sun, H. Zhang, A. Palazoglu, A. Singh, W. Zhang and S. Liu, Prediction of 24-hour-average PM<sub>2.5</sub> concentrations using a hidden Markov model with different emission distributions in Northern California. *Sci. Total Environ.* **443** (2013) 93-103.
- [38] D. Kang, R. Mathur and S. Trivikrama Rao, Assessment of bias-adjusted PM<sub>2.5</sub> air quality forecasts over the continental United States during 2007. *Geosci. Model Dev.* **3**(1) (2010) 309-320.
- [39] X. Yang, Y. Zheng, G. Geng, H. Liu, H. Man, Z. Lv, K. He and K. de Hoogh, Development of PM<sub>2.5</sub> and NO<sub>2</sub> models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. *Environ. Pollut.* **226** (2017) pp.143-153.
- [40] C. Liu, B.H. Henderson, D. Wang, X. Yang and Z.R. Peng, A land use regression application into assessing spatial variation of intra-urban fine particulate matter (PM<sub>2.5</sub>) and nitrogen dioxide (NO<sub>2</sub>) concentrations in City of Shanghai, China. *Sci. Total Environ.* **565** (2016) 607-615.
- [41] K.C. Lam, S.L. Ng, W.C. Hui and P.K. Chan, Environmental quality of urban parks and open spaces in Hong Kong. *Environmental Monit. Assess.* **111**(1-3) (2005) 55-73.
- [42] H. Guo, T. Cheng, X. Gu, H. Chen, Y. Wang, F. Zheng and K. Xiang, Comparison of four ground-level PM<sub>2.5</sub> estimation models using parasol aerosol optical depth data from China. *Int. J. Environ. Res. Public Health*, **13**(2) (2016), 180.
- [43] Y. Lin, Y.Y. Chiang, F. Pan, D. Stripelis, J.L. Ambite, S.P. Eckel and R. Habre, Mining public datasets for modeling intra-city PM<sub>2.5</sub> concentrations at a fine spatial resolution. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems* (2017) (p. 25). ACM.
- [44] Y. Lin, N. Mago, Y. Gao, Y. Li, Y.-Y. Chiang, C. Shahabi and J.L. Ambite, Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning. In *Proceedings of the*

- 26th ACM SIGSPATIAL international conference on advances in geographic information systems (2018) (pp. 359-368).
- [45] J. Jokar Arsanjani, M. Helbich, M. Bakillah, J. Hagenauer and A. Zipf, Toward mapping land-use patterns from volunteered geographic information. *Int. J. Geogr. Inf. Sci.* **27**(12) (2013) 2264-2278.
- [46] D.K. Moore, M. Jerrett, W.J. Mack and N. Künzli, A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA. *J. Environ. Monit.* **9**(3) (2007) 246-252.
- [47] L. Li, J. Wu, N. Hudda, C. Sioutas, S.A. Fruin and R.J. Delfino, Modeling the concentrations of on-road air pollutants in southern California. *Environ. Sci. Technol.* **47**(16) (2013) 9291-9299.
- [48] S.Y. Kim, L. Sheppard and H. Kim, Health effects of long-term air pollution: influence of exposure prediction methods. *Epidemiology* (2009) 442-450.
- [49] Z. Ross, K. Ito, S. Johnson, M. Yee, G. Pezeshki, J.E. Clougherty, D. Savitz and T. Matte, Spatial and temporal estimation of air pollutants in New York City: exposure assignment for use in a birth outcomes study. *Environ. Health*, **12**(1) (2013) 51.
- [50] A.A. Szpiro, C.J. Paciorek and L. Sheppard, Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology (Cambridge, Mass.)* **22**(5) (2011) 680.
- [51] J. Seo, D.S.R. Park, J.Y. Kim, D. Youn, Y.B. Lim and Y. Kim, Effects of meteorology and emissions on urban air quality: a quantitative statistical approach to long-term records (1999–2016) in Seoul, South Korea. *Atmos. Chem. Phys.* **18**(21) (2018) 16121-16137.
- [52] J.A. Kamińska, The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław. *J. Environ. Manag.* **217** (2018) 164-174.
- [53] V.D. Le, T.C. Bui and S.K. Cha, Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. arXiv preprint [arXiv:1911.12919](https://arxiv.org/abs/1911.12919) (2019).
- [54] J. Xie, Z. Liao, X. Fang, X. Fang, Y. Wang, Y. Zhang and B. Wang, The characteristics of hourly wind field and its impacts on air quality in the Pearl River Delta region during 2013–2017. *Atmospheric Res.* **227** (2019) 112-124.
- [55] E. Radzka, The Effect of Meteorological Conditions on Air Pollution in Siedlce. *J. Ecol. Eng.* **21**(1) (2020) 97-104.
- [56] P. Wang, H. Guo, J. Hu, S.H. Kota, Q. Ying and H. Zhang, Responses of PM<sub>2.5</sub> and O<sub>3</sub> concentrations to changes of meteorology and emissions in China. *Sci. Total Environ.* **662** (2019) 297-306.
- [57] J.S. Irwin, Statistical evaluation of centreline concentration estimates by atmospheric dispersion models. *Int. J. Environ. Pollut.* **14**(1-6) (2000) 28-38.
- [58] D.G. Fox, Uncertainty in air quality modeling: a summary of the AMS workshop on quantifying and communicating model uncertainty, Woods Hole, Mass., September 1982. *Bull. Am. Meteorol. Soc.* **65**(1) (1984) 27-36.
- [59] R.A. Anthes, Y.-H. Kuo, E.-Y. Hsie, S. Low-Nam and T.W. Bettge, Estimation of skill and uncertainty in regional numerical models. *Quart. J. R. Meteorol. Soc.* **115** (1989) 763–806.
- [60] S.R. Hanna, D.G. Strimaitis and J.C. Chang, Hazard Response Modeling Uncertainty (A Quantitative Method). Volume 2. Evaluation of Commonly Used Hazardous Gas Dispersion Models. SIGMA RESEARCH CORP WESTFORD MA (1993).
- [61] M.B. Beck, J.R. Ravetz, L.A. Mulkey and T.O. Barnwell, On the problem of model validation for predictive exposure assessments. *Stoch Hydrol. Hydraul.* **11** (1997) 229–254.
- [62] J.C. Chang and S.R. Hanna, Air quality model performance evaluation. *Meteorol. Atmos. Phys.* **87**(1-3) (2004) 167-196.
- [63] Y. Zhang and Y. Yang, Cross-validation for selecting a model selection procedure. *J. Econom.* **187**(1) (2015) 95-112.
- [64] S. Mei, H. Li, J. Fan, X. Zhu and C.R. Dyer, Inferring air pollution by sniffing social media. In *Proceedings of the 2014 IEEE/ACM international conference on advances in social networks analysis and mining* (2014) (pp. 534-539). IEEE Press.
- [65] A. Kumar, I. Gupta, J. Brandt, R. Kumar, A.K. Dikshit and R.S. Patil, Air quality mapping using GIS and economic evaluation of health impact for Mumbai city, India. *J. Air Waste Manag. Assoc.* **66**(5) (2016) 470-481.
- [66] W.A. Hassan, Produce an analytical map for the distribution of air pollution by toxic gases in Baghdad city by geographic information system. *J. Al-Nahrain Univ. Sci.* **21**(2) (2018) 81-87.
- [67] A. Kumar, R.S. Patil, A.K. Dikshit and R. Kumar, Air quality assessment using interpolation technique. *Environment Asia*, **9**(2) (2016) 140-149.
- [68] M.H. Ehrampoush, S. Jamshidi, M.J. Zare Sakhvidi and M. Miri, A comparison on function of Kriging and inverse distance weighting models in PM<sub>10</sub> zoning in urban area. *J. Environ. Health Sustain. Dev.* **2**(4) (2017) 379-387.
- [69] C. Lin, Y. Li, Z. Yuan, A.K. Lau, C. Li and J.C. Fung, Using satellite remote sensing data to estimate the high-resolution distribution of ground-level PM<sub>2.5</sub>. *Remote Sens. Environ.* **156** (2015) 117-128.
- [70] H.J. Lee, R.B. Chatfield and A.W. Strawa, Enhancing the applicability of satellite remote sensing for PM<sub>2.5</sub> estimation using MODIS deep blue AOD and land use regression in California, United States. *Environ. Sci. Technol.* **50**(12) (2016) 6546-6555.
- [71] X. Meng, Q. Fu, Z. Ma, L. Chen, B. Zou, Y. Zhang and W. Xue et al., Estimating ground-level PM<sub>10</sub> in a Chinese city by combining satellite data, meteorological information and a land use regression model. *Environ. Pollut.* **208** (2016) 177-184.
- [72] S.A. Fruin, N. Hudda, C. Sioutas and R.J. Delfino, Predictive model for vehicle air exchange rates based on a large, representative sample. *Environ. Sci. Technol.* **45**(8) (2011) 3569-3575.
- [73] N. Hudda, S.P. Eckel, L.D. Knibbs, C. Sioutas, R.J. Delfino and S.A. Fruin, Linking in-vehicle ultrafine particle exposures to on-road concentrations. *Atmos. Environ.* **59** (2012) 578-586.
- [74] [https://hautsdefrance.cci.fr/wp-content/uploads/sites/6/2016/04/Atlas\\_NPCP\\_Edition2016.pdf](https://hautsdefrance.cci.fr/wp-content/uploads/sites/6/2016/04/Atlas_NPCP_Edition2016.pdf).
- [75] <https://www.atmo-hdf.fr>.
- [76] C. Gengembre, Analyse dynamique, en champ proche et à résolution temporelle fine, de l'aérosol submicronique en situation urbaine sous influence industrielle. (Doctoral dissertation). Retrieved from <http://www.theses.fr/> with ID 2018DUNK0489 (2018).
- [77] D. Nowak and G. Heisler, Air quality effects of urban trees and parks. Research Series Monograph. Ashburn, VA: National Recreation and Parks Association Research Series Monograph. 44 (2010) 1-44.
- [78] H.L. Liu and Y.S. Shen, The impact of green space changes on air pollution and microclimates: a case study of the Taipei Metropolitan area. *Sustainability* **6**(12) (2014) 8827-8855.
- [79] C.E. Catlett, P.H. Beckman, R. Sankaran and K.K. Galvin, Array of things: a scientific research instrument in the public way: platform design and early lessons learned. In *Proceedings of the 2nd international workshop on science of smart city operations and platforms engineering* (2017) (pp. 26-33). ACM.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.