ISSN 0027-1349, Moscow University Physics Bulletin, 2024, Vol. 79, Suppl. 2, pp. S774–S783. © Allerton Press, Inc., 2024.

Machine Learning Methods for Statistical Prediction of PM2.5 in Urban Agglomerations with Complex Terrain, Using Grenoble As an Example

A. I. Suslov[1*], M. A. Krinitskiy[1,2], C. Staquet[3], and E. Le Boudec[3]

[1] *Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow, 117997 Russia*
[2] *Moscow Institute of Physics and Technology (National Research University), Dolgoprudny, Moscow oblast, 141700 Russia*
[3] *Université Grenoble Alpes, 1209 Rue de la Piscine, Gières, 38610 France*

Received October 1, 2024; revised October 10, 2024; accepted October 20, 2024

Abstract—In this study, we propose several methods based on machine learning approaches for predicting air pollution levels in cities located in mountain valleys, with Grenoble (France) as a case study. Pollution forecasting is performed using both regression and classification of exceeding threshold levels. We employ a data-driven approach, utilizing various machine-learning models. Based on historical data from 2012 to 2018, collected at several meteorological stations in the Grenoble Valley, multiple machine learning models were trained to predict the daily average concentrations of fine particulate matter PM10 and PM2.5 three days ahead. Days with high PM concentrations exceeding the threshold values set by the World Health Organization (WHO) are of particular interest in our study. It was found that the presence of local meteorological conditions leads to the formation of temperature inversions, which are statistically associated with air pollution levels in this region. Although local meteorological conditions primarily determine the pollution level, the machine learning models considered in our study can be adapted for other cities in valleys by training them on relevant data.

Keywords: machine learning, air pollution, PM2.5, PM10, temperature inversion, air quality

DOI: 10.3103/S0027134924702242

1. INTRODUCTION

Air pollution is a severe problem for most cities in the world. Some of the most health-threatening pollutants are particulate matter (PM) with a diameter of less than 10 μm (PM10) and less than 2.5 μm (PM2.5) [1]. The primary sources of particulate matter are transportation, heating, industry, and waste incineration [2]. Once in the atmosphere, these particles directly affect the respiratory system and provoke cardiovascular diseases [3].

Grenoble is among the most polluted cities in France. It is located in a valley and surrounded by three mountain ridges. Mountain ridges and narrow valleys affect wind directions and form local meteorological conditions. The orography contributes to the distribution of anthropogenic emissions and their following concentration in the valleys [4, 5]. The location of Grenoble, local meteorological conditions, industry, and heating impact air quality negatively.

Given the significant influence of local meteorological conditions on the distribution and concentration of pollutants in the Grenoble Valley, it is crucial to develop accurate predictive models. Forecasting air pollution levels in urban agglomerations with complex terrain is relevant, as operational forecasts allow for timely emission reductions and the minimization of public health risks.

Meteorological conditions play a crucial role in the formation of PM concentrations in the atmosphere [4, 6–8]. This work pays special attention to temperature inversion, which strongly correlates with air pollution levels. Temperature inversion is defined as a layer of the atmosphere with a positive vertical temperature gradient $\partial T/\partial z > 0$. In the presence of an inversion, temperature increases with height, which prevents vertical air mixing and contributes to the accumulation of pollutants [9].

Large-scale meteorological conditions significantly impact temperature inversion formation. The absence of clouds, low wind speed, and the duration of the night period contribute to the intensification

*E-mail: suslov.ai@ocean.ru

of the Earth's surface cooling, thereby increasing the probability of temperature inversion formation. At the same time, the presence of clouds and wind activity can destabilize the inversion layer, causing turbulent mixing of air masses and preventing the formation of a stable temperature gradient.

Studies on the application of machine learning to predict air pollution levels, with particular attention to the effect of temperature inversion on PM concentrations, have been conducted in urban agglomerations with topography similar to Grenoble [6, 10, 11].

The traditional approach to accounting for inversion involves using temperature differences or temperature gradients at different altitudes and, if data are available, analyzing the entire temperature profile [6, 7, 12].

In a study conducted in Tehran [13], an artificial neural network was used to predict PM2.5 concentrations, using the intensity of temperature inversion as a model feature. Inversion is observed up to 70% of days in Tehran, emphasizing the intensity of inversions, measured as the gradient of temperature change with height. The study showed a significant influence of this characteristic on PM2.5 concentrations.

Mlakar and Faganeli Pucer [10] presented an original method of accounting for inversion by applying a clustering algorithm to analyze temperature profiles. The data was divided into 15 categories based on profile characteristics, allowing hidden patterns to be identified and determining features most associated with high PM10 levels. The highest PM concentrations corresponded to winter days with forenoon inversion, low wind speed, and high emissions from transport and heating.

In addition to inversion layer thickness and temperature differences, Zang et al. [12] considered additional parameters to improve PM forecast accuracy: atmospheric optical depth and boundary layer height. Incorporating these parameters in the machine learning model significantly improved forecast accuracy.

Tamas et al. [14] applied clustering based on various meteorological data, including inversion layer thickness. PM concentration was approximated separately for each cluster. Comparison of a simple regression model with a model using preclustered data showed that SOM and K-means algorithms significantly improved the accuracy of pollution peak approximation.

These studies demonstrate various approaches to using temperature inversion data to improve the accuracy of air pollution prediction models. The methods include various inversion features ranging from simple temperature differences at different altitudes to complex clustering methods. This diversity of approaches indicates the potential of using different features characterizing temperature inversion in predicting air pollution levels. Using machine learning methods allows us to automatically take into account complex relationships between the level of atmospheric pollution and temperature inversion parameters.

In the present study, we apply machine learning methods to predict daily average PM2.5 concentrations in the Grenoble Valley for three days. We use data from 2012 to 2018, including PM concentration levels and meteorological variables. We also consider additional factors characterizing meteorological conditions with an anticyclonic blocking effect. In particular, we introduced features characterizing temperature inversion, which strongly correlates with air pollution levels, as shown [4, 5, 8]. The relationship between anticyclonic blocking and temperature inversion was demonstrated in [6, 7]. This study uses temperature differences at various altitudes as temperature inversion. Additionally, we consider various meteorological variables associated with PM10 and PM2.5 pollution episodes to improve forecast quality, such as wind speed, direction, and precipitation.

This approach allows us to study how Grenoble's complex topography and local weather conditions affect air pollution dynamics.

By focusing on temperature inversions, we hope to contribute to a deeper understanding of air quality forecasting in Grenoble and regions with similar topography. Although local meteorological conditions largely govern pollution levels, the machine learning models applied in this study have the potential to be adapted for other cities located in the lowlands, particularly in Chelyabinsk [15] and Krasnoyarsk [16]. The incorporation of temperature inversion data and other meteorological variables in our model could be especially beneficial for these cities, as they share similar topographical features (being located in basins) and face comparable challenges with air pollution, making our approach particularly relevant for predicting PM2.5 levels in these urban environments.

The rest of the paper is organized as follows: The "Problem Formulation and Data Description" section describes the dataset collected for our study, including several data sources, including Les Frenes, Le Versoud, and Chamrousse weather stations. The "Machine Learning Problems and Validation" section outlines the task of predicting air pollution levels in terms of regression and classification. The "Machine Learning Models" section describes several approaches and advanced statistical methods (also known as machine learning methods) used to predict air pollution, as well as the procedure for their validation; the Results section presents the results of our

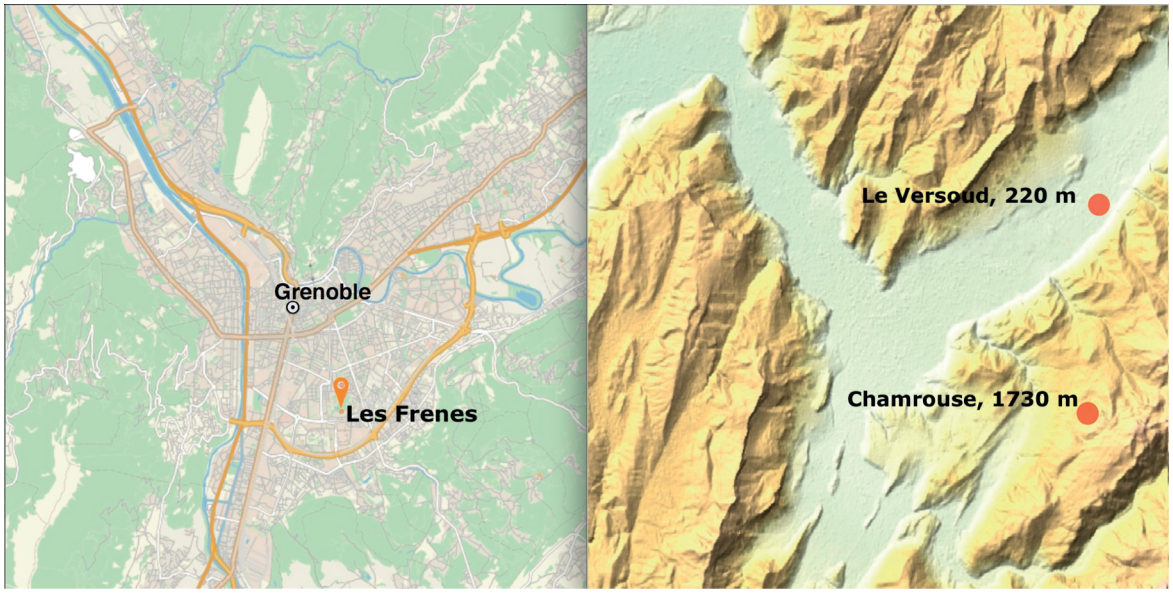

Fig. 1. Location of the main meteorological stations in Grenoble.

applied methods. The Conclusions section provides the conclusions of our work and prospects for further research.

2. PROBLEM FORMULATION AND DATA

In this study, the problem of forecasting the level of atmospheric pollution by fine particles of PM2.5 is solved using two approaches: regression and classification. In the regression approach, the daily average PM concentration level is approximated. In the classification task, we predict whether the concentrations exceed the levels established by the WHO 50 μg/m^3 for PM10 and 25 μg/m^3 for PM2.5.

2.1. Exploratory Data Analysis

We consider the observation station Les Frenes [17], located at a remote distance from intensive sources of pollution (production, transport interchanges, etc.). Thus, this station's pollution level depends mainly on local and large-scale meteorological conditions.

The study is based on seven years of data on meteorological variables (Table 1), with an emphasis on temperature inversions (temperature differences at different altitudes). The temperature inversion intensity is taken into account as the temperature difference between the Chamrousse (**1730** m) and Le Versoud (**220** m) stations. Figure 1 presents Grenoble's map and the location of meteorological stations used in the current study.

To account for the relationship between anticyclonic blocking phenomena and temperature inversion, we introduce converted pressure:

$$P_0 = P - 1013.25 \text{ hPa},$$

P_0 is the difference between the measured pressure P and the normal sea-level pressure 1013.25 hPa. A high value of $P_0 > 0$ indicates an anticyclonic regime. Additionally, we use geopotential height (Φ), eastern (U), and northern (V) components of synoptic wind at the 500 hPa levels to describe the state of the atmosphere.

The year is divided into winter and summer periods, depending on the heating season. Table 2 describes these periods. Exploratory data analysis was conducted to classify days as polluted and nonpolluted based on WHO criteria, categorized by temperature inversion and high atmospheric pressure. The results of this preliminary study are summarized in Table 3.

The data in Table 3 show that only 15% of PM2.5 measurements exceed the WHO threshold values. Most PM level exceedance (99%, 298 out of 300 cases) occur during winter, with more than 30% (108 out of 298) cases under a temperature inversion. Winter average PM2.5 concentrations are approximately twice as high as summer ones.

Figure 2 shows that days characterized by high PM concentrations are located in the "heavy tails" of the target variable distributions, PM10 and PM2.5.

Figure 3 presents a time series of meteorological variables, where high pollution levels (marked in red) often coincide with anticyclones (increased pressure) and the presence of temperature inversion (positive

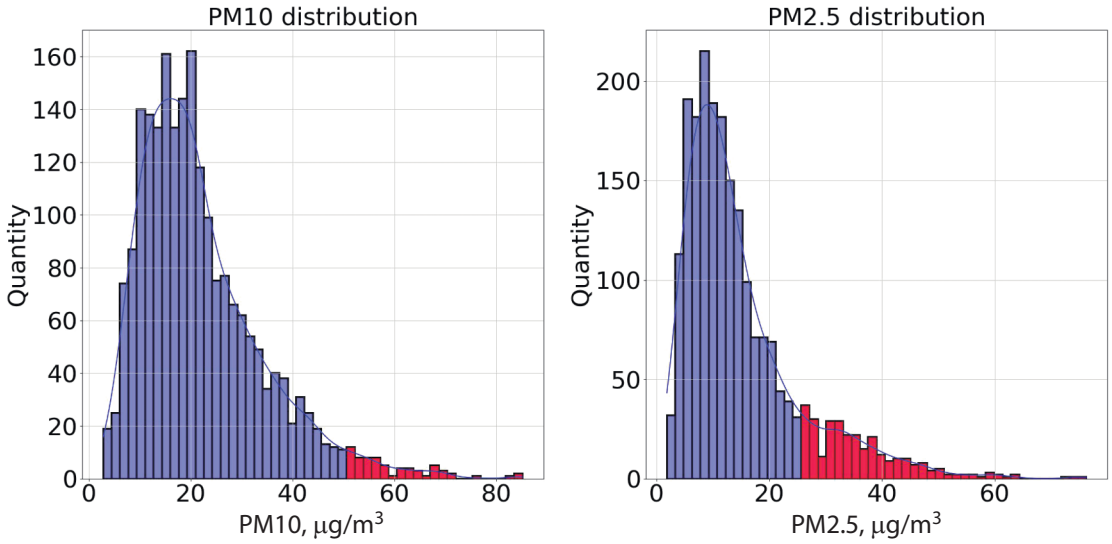

Fig. 2. Distribution of PM10 and PM2.5 concentrations. Episodes where PM concentration exceeds the WHO threshold are highlighted in red.

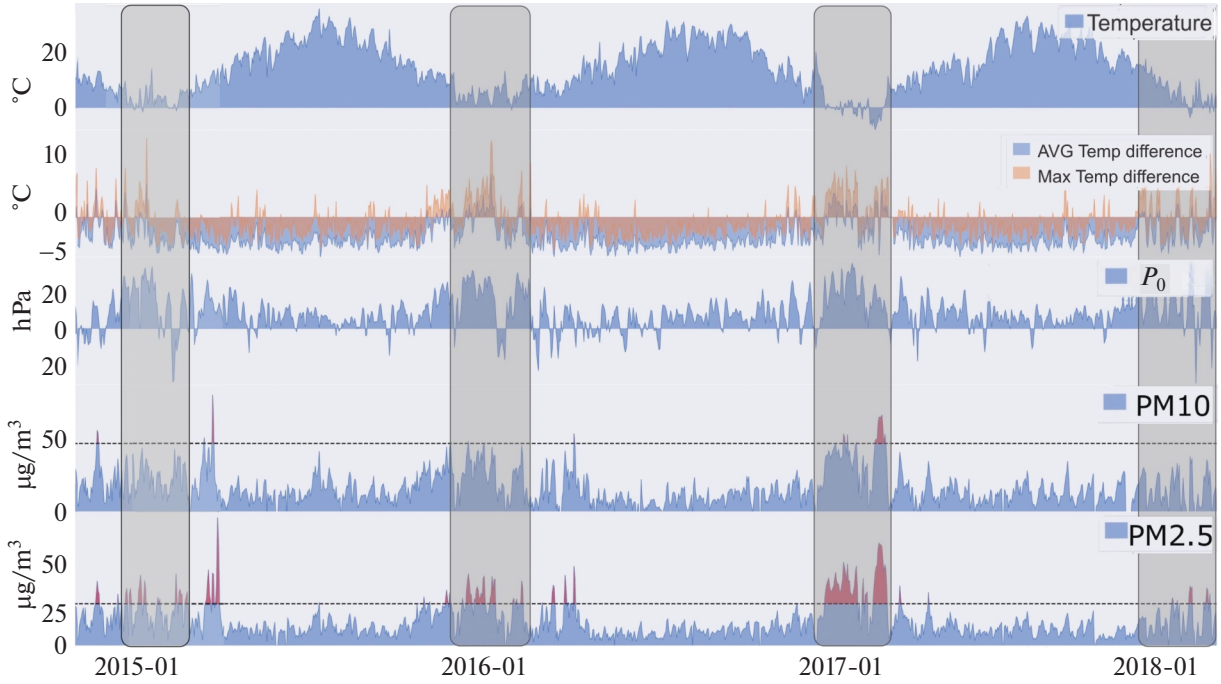

Fig. 3. Time series of meteorological variables. Values for PM10 and PM2.5 exceeding the pollution threshold set by WHO are shown in red. The dashed line indicates the threshold set by WHO.

temperature difference), predominantly in the winter period.

In ML, positional encoding is utilized to analyze cyclical variables such as days of the week or months, improving the recognition of recurring patterns and the interpretation of time cycles by ML models. For days of the week, the following formula is used:

$$\sin_{\text{day}} = \sin\left(2\pi \times \frac{\text{date}}{7}\right),$$

$$\cos_{\text{day}} = \cos\left(2\pi \times \frac{\text{date}}{7}\right).$$

Cyclic variables for months are calculated in the same manner:

$$\sin_{\text{month}} = \sin\left(2\pi \times \frac{\text{date}}{12}\right),$$

$$\cos_{\text{month}} = \cos\left(2\pi \times \frac{\text{date}}{12}\right).$$

Table 1. Meteorological variables and their units of measurement

Variable	Unit of measurement
Concentration of PM10	$\mu g/m^3$
Concentration of PM2.5	$\mu g/m^3$
T—temperature at versoud	°C
δT—temperature difference between Chamrousse and Versoud	°C
P_0—difference between measured pressure and sea level pressure	hPa
PCPN—precipitation	mm
U—eastern component of synoptic wind at 500 hPa level	m/s
V—northern component of synoptic wind at 500 hPa level	m/s
Φ—geopotential height at 500 hPa level	m^2/s^2

Table 2. Heating seasons and their date ranges

Heating season	Date range
Summer period	15.04–14.10
Winter period	15.10–14.04

According to Table 1, each day in the dataset is characterized by temporary variables (month, day of the week, season) and meteorological variables: PM10 and PM2.5 concentrations, temperature in Versoud (T), temperature difference between Chamrousse and Versoud ($\boldsymbol{\delta T}$), converted pressure (P_0), precipitation (PCPN), eastern (U) and northern (V) wind components at the 500 hPa level, and geopotential height at the same level (Φ).

3. MACHINE LEARNING PROBLEMS AND VALIDATION

3.1. Regression

Generally, the regression problem is formulated as an approximation of a continuous target variable. In the present study, the target variable is the predicted value of the daily average concentration of PM2.5 for each of the three days in the future from the events under consideration. In the case of the regression formulation of the problem, the following metrics are used to assess the quality of the model:

Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |P_i - M_i| .$$

Mean Square Error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (P_i - M_i)^2 .$$

Root Mean Square Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (P_i - M_i)^2},$$

where P is the model estimate of PM10 or PM2.5 concentration, and M is the measured reference value of the corresponding concentration.

3.2. Classification

Classification determines whether an event belongs to one of the predefined categories, e.g., in the context of this study, whether a polluted episode will occur based on WHO criteria:

$$\mathrm{Output} = \begin{cases} 1, & \text{if } \langle PM2.5\rangle_{24h} > 25\ \mu g\,/\,m^3 \\ 0, & \text{otherwise.} \end{cases}$$

The data analysis from the Les Frenes weather station showed a significant imbalance: the percentage of days exceeding the WHO pollution threshold was 3.5% for PM10 and 15% for PM2.5. An unbalanced sample makes it very difficult to predict elements of an underrepresented class. In the case of imbalanced classification, the most straightforward quality assessment metrics, such as Accuracy (proportion of correct answers), are uninformative. Most quality assessment metrics can be calculated using the values of the so-called confusion matrix [18]. In this matrix, True Positive (TP) and True Negative (TN) indicate correctly classified objects of each class, and False Positive (FP) and False Negative (FN) indicate type I and type II errors, respectively.

$$\mathrm{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Precision is the relation of correctly classified positive values to actual positives, while **Recall** is the relation of actual positive values correctly identified. In the Results section, we use Recall and F1 score, the harmonic mean between Precision and Recall:

$$\mathrm{Precision} = \frac{TP}{TP + FP},$$

$$\mathrm{Recall} = \frac{TP}{TP + FN},$$

$$\mathrm{F1\ Score} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

Table 3. Meteorological characteristics and pollution levels by season

Season	PM2.5		Total days	High PM2.5 level	Temp. inversion	High pressure	All three factors combined
Winter	20	6.25	1015	298	161	789	108
Summer	9.74	5.8	974	2	1	821	0

Table 4. Results for approximation of daily averaged PM2.5 concentrations using increments of meteorological variables and Gaussian noise

Model	1 day MAE		1 day RMSE		2 days MAE		2 days RMSE		3 days MAE		3 days RMSE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
ΔCatBoost (ΔMeteo + GN)	**3.34**	0.54	**4.97**	0.68	**4.74**	0.58	6.92	0.88	**5.47**	0.62	7.84	0.86
ΔCatBoost (ΔMeteo)	3.38	0.56	5.03	0.68	4.75	0.65	**6.91**	0.97	5.48	0.70	**7.83**	0.98
CatBoost abs. values	3.77	0.96	5.53	1.38	4.91	1.15	7.13	1.68	5.50	1.03	7.92	1.52
Persistence	4.10	0.75	5.93	0.93	5.85	0.92	8.30	1.20	6.58	0.92	9.30	1.11

3.3. Cross-Validation

Cross-validation (CV) is a technique for evaluating the generalizability of a machine-learning model on previously unseen data to detect overfitting or prediction bias. In CV, the data is divided into k parts: one part is used for testing, and the other $k-1$ parts are used for training. This process is repeated with each test part. The final results are averaged to estimate the model's overall performance of the ML algorithm. The study applies five-fold cross-validation. Model evaluation includes analysis of mean values (Mean) and standard deviations (STD) of quality assessment metrics for classification and regression problems.

4. MACHINE LEARNING MODELS

The present study implemented air pollution level prediction using regression and classification approaches. Heuristic models, including constant and inertial models, were used to evaluate the baseline quality, assuming their performance would be lower than any ML model in this study. Model descriptions are provided below.

The baseline models, particularly the inertial forecast model, were used to estimate the quality of more complex models.

The result of the inertial forecast model (Persistence) is the value of PM2.5 pollution levels known from the dataset from the previous day.

The following algorithms demonstrated the best performance:

Catboost: A gradient-boosting algorithm developed by Yandex used in regression and classification [19].

Balanced Random Forest (BRF): A modification of a random forest that can handle unbalanced data [20].

4.1. Regression Approach

4.1.1. Absolute values approximation. The target variable, PM concentration, has a skewed distribution where pollution episodes are located in the "heavy tail" of distribution (2). In the regression approach, the approximation of absolute values shows relatively low accuracy because extreme pollution events are located in the "tail" of the distribution, and their approximation presents significant challenges for statistical modeling.

4.2. Improving Forecast Accuracy Techniques

4.2.1. Target variable increments approximation. To enhance the accuracy of forecasting in the regression task and adequately predict the pollution peaks, instead of using the absolute values of PM10 and PM2.5 concentrations, we analyze the changes in PM10 and PM2.5 concentrations relative to the previous day:

$$\Delta PM_{d-1} = PM_{d-1} - PM_{d-2},$$

$$\Delta PM_{d0} = PM_{d0} - PM_{d-1},$$
$$\Delta PM_{d+1} = PM_{d+1} - PM_{d0},$$
$$\Delta PM_{d+2} = PM_{d+2} - PM_{d+1},$$
$$\Delta PM_{d+3} = PM_{d+3} - PM_{d+2}.$$

Here, ΔPM_{d-1} and ΔPM_{d0} denote differences in air pollution levels for the previous and current day, respectively. ΔPM_{d+1}, ΔPM_{d+2}, and ΔPM_{d+3} represent increments of pollution levels for the next day $(d+1)$, two days ahead $(d+2)$, and three days ahead $(d+3)$. PM_{d-1}, PM_{d0}, PM_{d+1}, PM_{d+2}, and PM_{d+3} are PM concentrations for the previous day $(d-1)$, current day $(d0)$, next day $(d+1)$, two days ahead $(d+2)$, and three days ahead $(d+3)$, respectively.

$$PM_{d+1} = PM_{d0} + \Delta PM_{d+1},$$
$$PM_{d+2} = PM_{d0} + \Delta PM_{d+1} + \Delta PM_{d+2},$$
$$PM_{d+3} = PM_{d0} + \Delta PM_{d+1} + \Delta PM_{d+2} + \Delta PM_{d+3}.$$

4.2.2. Meteorological variables increments. In order to improve the efficiency of pollution peak forecasting, we use both the value of meteorological variables and their changes over time. To improve the approximation accuracy, increments of meteorological variables were added to the training data set of the MO model, namely:

$$\Delta T_{(d-1)}, \Delta T_{(d0)}, \Delta \delta T_{(d-1)},$$
$$\Delta \delta T_{(d0)}, \Delta P_{0(d-1)}, \Delta P_{0(d0)},$$
$$\Delta U_{(d-1)}, \Delta U_{(d0)}, \Delta V_{(d-1)}, \Delta V_{(d0)}, \Delta \Phi_{(d-1)}, \Delta \Phi_{(d0)}.$$

Increments of meteorological variables were used in both regression and classification tasks.

4.2.3. Artificial data augmentation. Data augmentation with random noise involves adding Gaussian noise to the original dataset, enhancing the model's ability to generalize by simulating variations in the data. This technique is particularly useful for creating additional training examples in cases where the dataset is limited, helping to prevent overfitting and improve the robustness of machine learning models.

To increase the training sample size, the classical augmentation technique is used. We add similar examples to existing data with the addition of random Gaussian noise. This technique has been used for both regression and classification problems. For each meteorological variable $\mathbf{x}_i$, the augmented value $\mathbf{x}'_i$ is calculated by adding Gaussian noise:

$$\mathbf{x}'_i = \mathbf{x}_i + \epsilon, \quad \text{where} \quad \epsilon \sim \mathcal{N}\left(0, \frac{\sigma_{\mathbf{x}_i}}{\mathbf{c}}\right),$$

where $\mathbf{x}_i$ is the original value of the meteorological variable, and ϵ represents Gaussian noise with Mean $= 0$ and standard deviation $\frac{\sigma_{\mathbf{x}_i}}{\mathbf{c}}$. $\sigma_{\mathbf{x}_i}$ is the standard deviation of $\mathbf{x}_i$, and $\mathbf{c}$ is a noise scaling factor. In our study, $\mathbf{c} = 10$.

For variables like precipitation (PCPN), negative values are not physically meaningful. Therefore, the augmentation process ensured that the augmented value remained nonnegative:

$$\mathbf{x}'_i = \begin{cases} \mathbf{x}_i + \epsilon, & \text{if} \quad \mathbf{x}_i > 0 \\ 0, & \text{if} \quad \mathbf{x}_i \leq 0. \end{cases}$$

4.3. Classification Approach

To address unbalanced data, techniques such as upsampling of the minority class were used. Specifically, data from the winter period characterized by the highest concentration of PM2.5, 01.11-31.03 were extracted and augmented to expand the training dataset. This process involves creating noisy versions of the original winter data by adding Gaussian noise, as described in the previous subsection. This augmentation is repeated twice (the number is determined empirically). This allows for a significant increase in the representation of days with high PM2.5 concentrations. A condition is imposed to prevent data leakage during the augmentation process so that the augmented data in each fold are generated only from indices not included in the test dataset at a given iteration of that fold. This approach prevents the model from accessing information from the test set during training, thereby preserving the integrity of the cross-validation process. The inclusion of these extended datasets, along with the original training data, significantly improved the robustness and generalizability of the model and allowed machine learning models to be trained more efficiently on unbalanced data. Another approach is to employ a weighted loss function. In this method, a multiplicative weight is assigned to each predicted value. A higher weight is assigned to values with significant errors, while a lower weight is assigned to those with lower errors. By default, all items belonging to each class have the same weight, typically set to 1. In this study, identifying the majority class (PM levels exceeding the WHO threshold) is more important than the minority class. Statistically, upsampling and the use of a weighted loss function are equivalent.

4.4. Regression Results

4.4.1. Target variable increments approximation. Hereinafter, abbreviations related to the problem formulation and model input data will be used: CatBoost abs. values—regression problem in the formulation of forecasting absolute PM values,

Table 5. Improvement in regression quality metrics depending on problem formulation and feature description of the events; averaged over three days

Model comparison	MAE error reduction (average %)	RMSE error reduction (average %)
ΔCatBoost (ΔMeteo + GN) vs CatBoost abs. values	5.14%	4.69%

Table 6. F1-score metric results for classification

Model	1 day F1		2 days F1		3 days F1	
	Mean	Std	Mean	Std	Mean	Std
CatBoost binary (ΔMeteo GN)	**0.77**	0.07	**0.64**	0.10	**0.59**	0.12
CatBoost (ΔMeteo GN)	**0.76**	0.09	**0.67**	0.10	**0.62**	0.10
BRF (ΔMeteo GN)	**0.74**	0.07	**0.67**	0.09	**0.63**	0.08
Persistence	0.75	0.06	0.62	0.11	0.55	0.14

Table 7. Recall results for classification

Model	1 day Recall		2 days Recall		3 days Recall	
	Mean	Std	Mean	Std	Mean	Std
CatBoost binary (ΔMeteo GN)	**0.74**	0.07	**0.62**	0.11	**0.57**	0.15
CatBoost (ΔMeteo GN)	**0.85**	0.07	**0.78**	0.10	**0.73**	0.10
BRF (ΔMeteo GN)	**0.88**	0.04	**0.82**	0.07	**0.78**	0.10
Persistence	0.74	0.06	0.62	0.11	0.55	0.14

Table 8. δT impact on daily average PM2.5 concentration forecast quality

Metrics Improvement, %	1 day		2 days		3 days	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Improvement	0.82%	1.01%	1%	0.83%	0.97%	0.78%

ΔCatBoost—regression problem in the formulation of PM increments approximation ΔPM, ΔMeteo—the usage of meteorological variables increments in the model input data, GN—introduction of Gaussian noise in the model input data, CatBoost binary—the binary transformation of the regression results:

$$\text{Output} = \begin{cases} 1, & \text{if} \quad \langle \text{PM2.5} \rangle_{24h} > 25\ \mu\text{g}/\,\text{m}^3 \\ 0, & \text{otherwise.} \end{cases}$$

In this study, the daily averaged PM2.5 value increments were approximated between consecutive days. The CatBoost model showed the best result using increments of meteorological variables and Gaussian noise augmentation. Figure 4 shows the approximation of PM2.5 concentration level increments for the CatBoost model.

4.4.2. Regression quality metrics improvement. Tables 4 and 5 show that approximating PM increments between consecutive days improved the quality of all models in the presence of high pollution levels. The MAE and RMSE metrics decreased about 5%. However, there is a lag between the approximated and actual values over time. The model cannot accurately predict a sudden increase in PM concentration after more than a day and is unable to adequately approximate the initial value of the changed PM concentration.

4.4.3. Significance of temperature inversion features. The study evaluated the impact of the

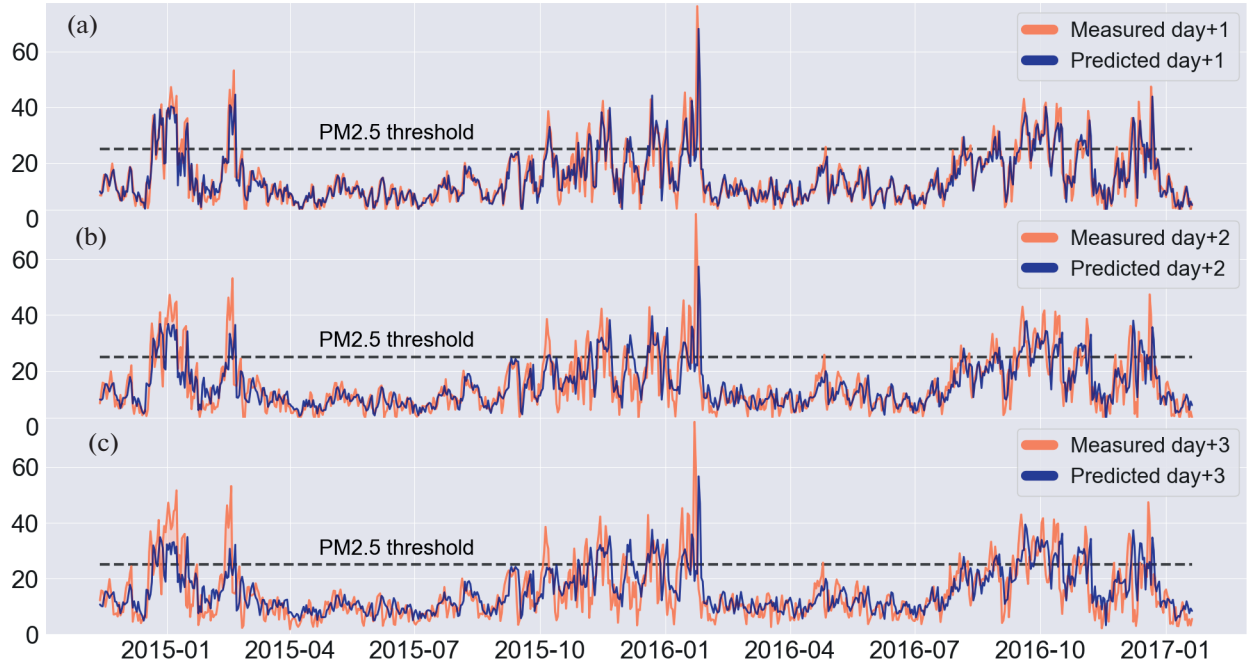

Fig. 4. Typical behavior of the CatBoost model in the task of predicting increments of PM2.5 pollution level for (a) one day ahead, (b) two days ahead, and (c) three days ahead. The orange line shows the measured values and the blue line indicates the approximated values.

temperature difference (δT) feature on the daily averaged PM2.5 concentration prediction quality. In order to evaluate the impact of the temperature difference variable, a CatBoost model was trained with the parameter δT in the model input data. Then, the temperature difference was eliminated from the dataset, and the same model was re-trained without this parameter. The results are presented in Table 8.

4.5. Classification Results

In the classification problem, two primary performance metrics were used: F1-score and Recall. These metrics were calculated for predictions made one, two, and three days ahead. The results for both metrics are shown in Tables 6 and 7.

Based on Tables 6 and 7, one can conclude that the choice of machine learning model depends on the quality metrics. For the Accuracy metric, the CatBoost model is the most effective. For crisis events prediction where the cost of error is high, the Balanced Random Forest model is preferred, but the probability of type I error increases with the number of correctly identified polluted episodes.

5. CONCLUSIONS

In this study, based on meteorological data from previous days, the average PM2.5 daily concentrations were predicted for three days ahead. The CatBoost algorithm showed the highest efficiency for approximating pollution peaks. For imbalanced classification, CatBoost and balanced random forest performed the best. Changing the problem formulation from predicting absolute PM concentration values to the approximation of their increments between consecutive days improved the forecast accuracy. Accounting for meteorological variable increments also improved the quality of the forecast in both classification and regression formulation. Augmenting the training sample with Gaussian noise improved the forecast accuracy in both approaches. Including the temperature difference parameter in the input data of ML models improved the approximation accuracy by only 1%, partially supporting already published studies. Presumably, the temperature (T) and converted pressure (P_0) features strongly correlate with temperature differences. Using a weighted loss function in the classification task allowed us to classify polluted episodes more precisely. The number of correctly classified polluted days increased as the probability of type I errors increased.

FUNDING

This work was supported by state agreement no. FMWE-2024-0017.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Yi. Du, X. Xu, M. Chu, Ya. Guo, and J. Wang, J. Thorac. Dis. **8**, E8 (2015). https://doi.org/10.3978/j.issn.2072-1439.2015.11.37
2. F. Karagulian, C. A. Belis, C. F. C. Dora, A. M. Prüss-Ustün, S. Bonjour, H. Adair-Rohani, and M. Amann, Atmos. Environ. **120**, 475 (2015). https://doi.org/10.1016/j.atmosenv.2015.08.087
3. C. A. Pope and D. W. Dockery, J. Air Waste Manage. Assoc. **56**, 709 (2006). https://doi.org/10.1080/10473289.2006.10464485
4. Ya. Largeron and Ch. Staquet, Atmos. Environ. **135**, 92 (2016). https://doi.org/10.1016/j.atmosenv.2016.03.045
5. A. S. Panicker, G. Pandithurai, P. D. Safai, and T. V. Prabha, Q. J. R. Meteorol. Soc. **142**, 2968 (2016). https://doi.org/10.1002/qj.2878
6. E. Isaev, B. Ajikeev, U. Shamyrkanov, K.-U. Kalnur, K. Maisalbek, and R. C. Sidle, Aerosol Air Qual. Res. **22**, 210336 (2022). https://doi.org/10.4209/aaqr.210336
7. M. Pasic, H. Hadziahmetovic, I. Ahmovic, and M. Pasic, Sustainability **15**, 11230 (2023). https://doi.org/10.3390/su151411230
8. G. D. Silcox, K. E. Kelly, E. T. Crosman, C. D. Whiteman, and B. L. Allen, Atmos. Environ. **46**, 17 (2012). https://doi.org/10.1016/j.atmosenv.2011.10.041
9. J. H. Seinfeld and S. N. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, 2nd ed. (Wiley, Hoboken, NJ, 2006).
10. P. Mlakar and J. Faganeli Pucer, Atmosphere **14**, 481 (2023). https://doi.org/10.3390/atmos14030481
11. T. T. Trinh, T. T. Trinh, T. T. Le, T. D. H. Nguyen, and B. M. Tu, Environ. Geochem. Health **41**, 929 (2019). https://doi.org/10.1007/s10653-018-0190-0
12. Z. Zang, W. Wang, W. You, Yi. Li, F. Ye, and Ch. Wang, Sci. Total Environ. **575**, 1219 (2017). https://doi.org/10.1016/j.scitotenv.2016.09.186
13. R. A. Bahari, R. A. Abbaspour, and P. Pahlavani, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XL–2/W3**, 73 (2014). https://doi.org/10.5194/isprsarchives-xl-2-w3-73-2014
14. W. Tamas, L. Paulik, and Z. Nejedly, Atmos. Pollut. Res. **7** (1), 96 (2016).
15. V. V. Zavoruev, O. V. Sokolova, E. N. Zavorueva, and O. E. Yakubailik, Atmos. Oceanic Opt. **36**, 663 (2023). https://doi.org/10.1134/s1024856023060246
16. T. G. Krupnova, O. V. Rakova, K. A. Bondarenko, A. F. Saifullin, D. A. Popova, S. Potgieter-Vermaak, and R. H. M. Godoi, Int. J. Environ. Res. Public Health **18**, 12354 (2021). https://doi.org/10.3390/ijerph182312354
17. Les Frenes station. https://aqicn.org/city/france/rhonealpes/isere/grenoble-les-frenes/. Cited August 26, 2024.
18. Confusion matrix. https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5. Cited August 26, 2024.
19. A. V. Dorogush, V. Ershov, and A. Gulin, arXiv Preprint (2018). https://doi.org/10.48550/arXiv.1810.11363
20. Balanced random forest algorithm. https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.BalancedRandomForestClassifier.html. Cited August 26, 2024.

AI tools may have been used in the translation or editing of this article.