Predicting PM_{10} from $PM_{2.5}$ and NO_2 Levels in Metropolitan Lima, Peru: A Regularized Regression Approach

Jorge Chavarri Centeno, Hilter Lanche Cerron, Joel Rivera Matias, Juan Jesús Soria-Quijaite, Nemias Saboya
Escuela Profesional de Ingeniería de Sistemas, Facultad de Ingeniería y Arquitectura,
Universidad Peruana Unión, Lima, Perú

***Abstract*—Suspended particle pollution (PM_{10}) represents a serious public health problem in Metropolitan Lima, since its high levels are related to an increase in respiratory and heart diseases. This work aims to anticipate PM_{10} concentrations in seven districts of Lima using a multiple linear regression (OLS) model with the Ridge and Lasso regularization techniques. $PM_{2.5}$ and NO_2 were selected as predictors because they were the only pollutant variables available in the SENAMHI monitoring dataset; hourly records collected between 2015 and 2024 were used, yielding 169,308 observations after removing missing values and outliers. The data were separated into 80% for training and 20% for validation, applying cross-validation to determine the most appropriate regularization parameter. The final model reached a coefficient of determination (R^2) of 0.23 and a mean squared error (MSE) of 521 in the test set. The findings indicate that $PM_{2.5}$ is the main factor that predicts PM_{10}, while NO_2 exerts a secondary impact. Although the predictive capacity of the models is limited, the use of both Ridge and Lasso helped stabilize the coefficients and minimize overfitting. Given this limited explanatory power, the model should be regarded as an initial, interpretable baseline for PM_{10} monitoring in Metropolitan Lima rather than an operational forecasting tool. It is suggested to include meteorological variables and investigate non-linear models in future studies to enhance the accuracy and usefulness of the model in air quality management.**

Keywords*—Multiple linear regression; particulate matter of 10 micrometers or less (PM_{10}); particulate matter of 2.5 micrometers or less ($PM_{2.5}$); nitrogen dioxide (NO_2); ridge regularization; Lasso regularization

I. INTRODUCTION

Over the past few decades, rapid industrialization, increased fossil fuel consumption, population growth, urbanization, and other human-induced changes have dramatically increased air pollution throughout the world [1]. Air pollution is one of the main environmental and public health challenges worldwide, especially in large urban centers. This phenomenon is mainly caused by a combination of industrial emissions, pollution from transport, waste burning, and other domestic activities. Its effects are manifested in the deterioration of the environment and have direct impacts on human health, causing diseases such as lung cancer, asthma, chronic bronchitis, and heart disease [2]. Among the most studied atmospheric pollutants is particulate matter, which is a complex mixture of solid particles and aerosols composed of microscopic droplets of liquid, dry solid fragments, and solid cores with liquid coatings [3]. Some of these particles have a diameter of less than or equal to 10 micrometers (PM_{10}), which, due to its small size, can soak deep into the respiratory system and even reach the bloodstream, generating a series of adverse effects on human health [4]. In addition, recent studies have shown a significant correlation between critical levels of air pollution and subjective satisfaction with life [5]. Epidemiological studies of the last decade have found a positive association between the concentration of PM_{10} and the increase in total, respiratory, and cardiovascular mortality as a consequence, even at exposure levels below the air quality standards established by the United States Environmental Protection Agency [6],[7],[8]. In addition, PMs have been found to reach the brain both through the olfactory pathway and through blood circulation, which can have serious implications for neurological health, such as neurodegenerative diseases [9]. Another study also found that the contaminant PM_{10} is considered a determining factor in the pathophysiology and aging of the skin, as well as in the appearance of dermatological diseases [10]. The WHO establishes a threshold of 45 $\mu g/m^3$ for 24-hour exposure and 15 $\mu g/m^3$ for annual exposure, limits that have been widely exceeded in some urban regions [11].

Various studies in Latin American and Asian cities have shown the impact of PM_{10} on public health. In Temuco, Chile, a positive association was found between PM_{10} concentrations and mortality due to respiratory and cardiovascular causes [12]. In La Paz, even without exceeding national limits, PM_{10} was related to an increase in medical consultations for respiratory conditions and acute symptoms such as cough and shortness of breath [13]. Likewise, in the Seoul metro, high concentrations of PM_{10} exceeded background urban values and were associated with more episodes of respiratory diseases among workers and frequent users [14]. In Iran, prolonged exposure to high levels of PM_{10} was associated with unacceptable risks of cancer and other adverse public health effects, far exceeding recommended values [15]. Research carried out in Poland has examined the correlation between air pollution and the appearance of respiratory diseases such as COVID-19. Although recent research did not find a systematic correlation between infection rates and exposure to high levels of $PM_{2.5}$ and PM_{10}, it was noted that the periods of low levels of these particles were linked to a reduced incidence of the disease [8]. A study carried out in the city of Andimshk also highlights the serious impacts of prolonged exposure to PM_{10}, with 39.31% of short-term deaths associated with air pollution. These findings underscore the importance of preventive strategies, such as wearing masks and reducing outdoor activities on days with high concentrations of particles, to protect public health from the adverse effects of air pollution [16],[17]. On the other hand, a more recent study on the spatio-temporal variability of particulate matter (PM_{10} and $PM_{2.5}$) and ozone (O3) in Barranquilla, Colombia, found average PM_{10} concentrations of 46.4 μg/m^3, 51.4 μg/m^3, and 39.7 μg/m^3, depending on the season. These concentrations were higher in the south of the city, due to local sources such as traffic and industrial activities [18]. In Metropolitan Lima and Callao, a DIGESA study identified PM_{10} as the main pollutant, registering in districts such as Carabayllo concentrations 333% higher than the Peruvian annual standard. At the same time, in coastal areas the levels were lower. This scenario highlights the need for specific control strategies to reduce the burden of PM_{10} and protect public health in critical areas [19].

The measurement and prediction of PM_{10} concentrations have become central to pollution prevention and control [20]. However, most predictive studies on PM_{10} remain confined to specific geographical contexts and rarely establish a transparent, reproducible baseline before adopting more complex modeling strategies; for Metropolitan Lima, no prior study has evaluated regularized linear regression as an interpretable reference model using only routinely monitored pollutant variables. This gap limits local authorities' ability to understand which variables drive PM_{10} variability before investing in more data-intensive approaches.

Ridge and Lasso regression were selected over more advanced algorithms, such as Random Forest or gradient boosting, for three reasons: they produce interpretable coefficients valuable for environmental policy design; with only two predictors ($PM_{2.5}$ and NO_2) available, the added complexity of nonlinear or ensemble methods is not justified and risks overfitting; and regularization directly addresses multicollinearity and coefficient instability while preserving the transparency needed for a benchmarkable baseline.

The general objective of this study is to develop a predictive model of PM_{10} using multiple linear regression regularized by Ridge and Lasso, from $PM_{2.5}$ and NO_2 data collected in seven districts of Metropolitan Lima between 2015 and 2024 [21]. Specifically, it aims to: 1) quantify the relative contribution of $PM_{2.5}$ and NO_2 to PM_{10}; 2) compare the predictive performance of OLS, Ridge, and Lasso under a common cross-validation framework; and 3) establish an interpretable baseline for future, more sophisticated approaches. These findings will support more rigorous air-quality monitoring and evidence-based environmental management in Metropolitan Lima.

II. Related Work

In recent years, various studies have used Machine Learning models for the prediction of atmospheric pollutants, seeking to anticipate dangerous concentrations at specific points. These investigations have focused mainly on Multiple Linear Regression algorithms and their regularized variants, Ridge and Lasso, on a set of data extracted from different cities.

One of these investigations was carried out in India, where a predictive model based on Machine Learning was developed to estimate air quality (AQI) from the concentrations of various pollutants, comparing the aforementioned models and their results of RMSE and MSE metrics, it was concluded that Standard Multiple Linear Regression (OLS) obtained better performance, implying that regularization does not always improve accuracy when the variables do not have high correlation [22].

On the other hand, recent research conducted in Junín, Peru, also proposed a predictive Machine Learning model to estimate air quality behavior using environmental records such as temperature, humidity, and pressure. In this case, PM2.5 AQI, corresponding to the air quality index based on the concentration of fine particles, was predicted. Linear Regression, Ridge, Lasso, and Elastic Net models were applied. The results showed that Ridge reached more stable values in RMSE and R2, so the authors concluded that the integration of environmental variables in regularized regression models improves the predictive capacity and accuracy of the model [23].

Likewise, another study carried out in Lima Este developed a Linear Regression model to analyze the relationship between PM_{10} and $PM_{2.5}$ concentrations. The results showed a highly significant relationship between both variables and relevant coefficients, which evidences a good fit for the model and the ability to estimate $PM_{2.5}$ levels from PM_{10} [24].

The studies carried out confirm that Multiple Linear Regression models and their regularized variants offer positive results in predicting air quality. However, most of this type of research has been carried out in specific geographical areas, which limits the generalization of its results. More recently, other authors have moved beyond linear formulations toward ensemble and deep-learning architectures, seeking to capture nonlinear relationships that regression-based models cannot represent.

One such study, conducted in Jiangxi Province, China, developed a hybrid model that integrated Random Forest and XGBoost to estimate high-resolution daily $PM_{2.5}$ concentrations from satellite reflectance, meteorological, topographic, and vegetation data. The hybrid approach substantially reduced prediction error compared with a standalone Random Forest baseline, showing that combining both algorithms captures nonlinear relationships that a single ensemble method cannot fully represent [25].

Likewise, a study carried out in Tarnów, Poland, compared three ensemble supervised learning algorithms, Decision Tree (CART), Random Forest, and Cubist Rule, to predict PM_{10} concentrations and estimate long-term pollution trends after meteorological normalisation. Random Forest and Cubist achieved noticeably higher accuracy than the simpler Decision Tree approach, and air temperature and boundary layer height were consistently identified as the strongest predictors of PM_{10} variability across all three algorithms [26].

In a related effort, research conducted in Macao proposed a stacking ensemble learning model that used LSTM and XGBoost as base learners and LightGBM as the meta-learner to predict $PM_{2.5}$ concentrations from historical pollutant and meteorological data collected between 2015 and 2023. Across several evaluation metrics, this combined architecture surpassed each of its constituent learners, LSTM, CNN, Random Forest, and XGBoost, when applied individually, illustrating the benefit of combining heterogeneous learners for particulate matter forecasting [27].

Additionally, a study developed for Shanghai and Kaohsiung introduced a dual-channel deep learning model that fused surveillance images with atmospheric, meteorological, and temporal data to forecast $PM_{2.5}$ and PM_{10} concentrations. The proposed VGG16–LSTM architecture achieved a strong fit for both pollutants on the Shanghai dataset and retained this level of accuracy when subsequently validated on two independent stations in Kaohsiung, Taiwan, confirming its adaptability across different urban contexts [28].

Finally, a comparative study conducted in Abu Dhabi, United Arab Emirates, evaluated Decision Tree, Random Forest, Support Vector Regression, Convolutional Neural Networks, LSTM, and Facebook Prophet for forecasting $PM_{2.5}$ and PM_{10} over multiple horizons (1–2 hours, 1 day, and 1 week) using five years of data from six monitoring stations. Support Vector Regression performed best for $PM_{2.5}$ across both short-term horizons, Convolutional Neural Networks led PM_{10} forecasts only at the 1-hour horizon (with SVR regaining the lead at 2 hours), and Facebook Prophet consistently outperformed the rest at longer horizons (1 day and 1 week) for both pollutants [29].

TABLE I. COMPARISON OF RECENT STUDIES ON PARTICULATE MATTER MODELLING

Study	**Predictors**	**Algorithm(s)**	**Key finding**
Soria et al. [24]	PM_{10}, monitoring setting	OLS regression	Coef. 0.57–0.67 by setting
Xie et al. [17]	Meteorology, NO_x, SO_2, CO, VOCs	PCA + XGBoost	Accurate multi-pollutant
Olarte et al. [23]	Temperature, humidity, pressure	OLS, Ridge, Lasso, EN	$R^2 = 0.148$, RMSE = 25.36
Tang et al. [25]	Reflectance, meteorology, DEM, NDVI	RF + XGBoost (hybrid)	$R^2 = 0.82$
Gora & Rzeszutek [26]	Meteorology, ERA5, trend, calendar	CART, RF, Cubist	$R^2 \approx 0.88$–0.89
Tian et al. [27]	Historical pollutants, meteorology	Stacking (LSTM+XGB+LGBM)	Outperforms base learners
Wu et al. [28]	Pollutants, meteorology, images	Dual-channel VGG16–LSTM	$R^2 = 0.95/0.90$
Abuouelezz et al. [29]	Historical pollutants, meteorology	DT, RF, SVR, CNN, LSTM, Prophet	Varies by horizon

Table I shows that nonlinear ensembles and deep-learning models generally exploit a broader range of meteorological, spatial, temporal, and visual predictors than linear approaches. Their reported results demonstrate the value of capturing nonlinear and spatio-temporal relationships, but they also require more heterogeneous data and substantially greater model complexity. In this sense, the present study proposes the implementation of a Multiple Linear Regression and Ridge model that allows predicting PM_{10} from $PM_{2.5}$ and NO_2, applied in Metropolitan Lima, home to over 10.29 million inhabitants (30.2% of the national population), a figure equivalent to the combined population of the country's six other most populous regions [30]. Its contribution is therefore not superior predictive accuracy, since OLS, Ridge, and Lasso explain only about 23% of PM_{10} variability. Rather, the comparison establishes the performance attainable with a parsimonious, interpretable model for this densely populated area, and identifies the inclusion of meteorological variables and nonlinear learners as the clearest direction for subsequent improvement.

III. METHODOLOGY

A. Data

The data for this study come from the National Meteorological and Hydrological Service of Peru (SENAMHI) through its Network of Automatic Air Quality Monitoring Stations in Metropolitan Lima, which provides validated hourly measurements of PM_{10}, $PM_{2.5}$, and NO_2 in the period 2015–2024. Each record includes station information: name, latitude, longitude, altitude, department, province, district, and ubigeo, along with the date, time, and concentrations of the three pollutants: PM_{10}, $PM_{2.5}$, and NO_2.Of these, $PM_{2.5}$ and NO_2 were used as predictor variables because they are the variables available in the dataset most directly related to the formation of PM_{10}; $PM_{2.5}$ as it is a granulometric fraction contained within PM_{10}, and NO_2 as it acts as a tracer of vehicular emissions. It should also be noted that the dataset does not include meteorological variables such as temperature, humidity, or precipitation; consequently, these variables were not incorporated into the model despite their known influence on PM_{10} concentrations.The seven selected stations, located in the districts of Carabayllo, Campo de Marte, San Martín de Porres, San Juan de Lurigancho, Santa Anita, San Borja, and Villa María del Triunfo (see Fig. 1), allow a representative analysis of the spatial and temporal variability of air quality in Metropolitan Lima.

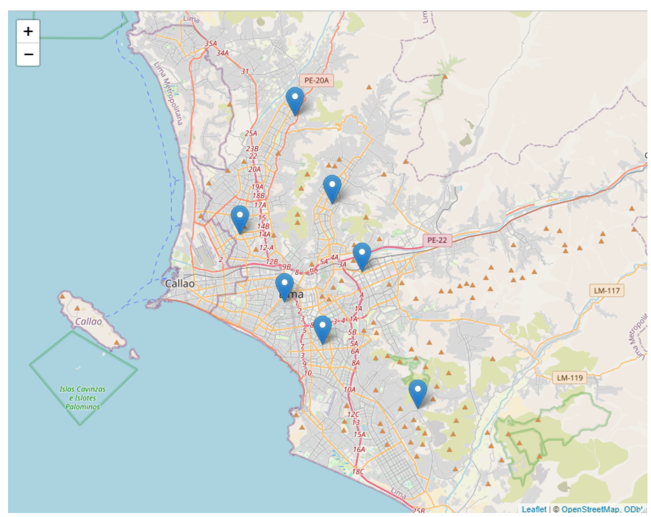

Fig. 1. Location of the monitoring stations in Metropolitan Lima.

B. Model Implementation

1) Multiple Linear Regression: The linear regression model is a statistical technique that allows predicting the value of a dependent variable as a result of the relationship between this variable Y and a set of independent variables $X_2, X_3, \ldots, X_k$. In its population form, it is expressed as [24], [31]:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \quad (i = 1, ..., n), \quad (1)$$

where, β_1 is the intercept, β_j (for $j = 2, \ldots, k$) are the partial coefficients indicating the mean change in Y per unit of X_j, and u_i represents the random error term. By grouping the n observations, the system can be written in matrix form [24], [31]:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & X_{21} & X_{31} & \dots & X_{k1} \\ 1 & X_{22} & X_{32} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \dots & X_{kn} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}}_{\mathbf{u}} \implies \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{2}$$

Under the classical assumptions (linearity in parameters, $E[u_i] = 0$, homoscedasticity, independence, and absence of perfect multicollinearity), the $\boldsymbol{\beta}$ parameters are estimated using the Ordinary Least Squares (OLS) method as

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathsf{T}}\mathbf{Y},$$

ensuring that the estimators are the best linear unbiased (BLUE). In addition, the goodness of fit is evaluated with R^2, and t and F tests are performed to verify the significance of the coefficients where Y is the dependent variable, X_i are the independent variables, β_i are the coefficients of the model, and ϵ is the error term.

In this study, Y corresponds to PM_{10} levels, while X_1 and X_2 correspond to $PM_{2.5}$ and NO_2, respectively. The model will be adjusted using the ordinary least squares method (OLS).

2) Ridge regression model: Ridge regression (Ridge) is an extension of the least squares technique that allows addressing the problem of multicollinearity between predictor variables, common in models with highly correlated variables. This method introduces a penalty on the magnitude of the coefficients, controlled by a λ parameter, which gives rise to biased estimators, but with less variance, thus facilitating better generalization and reducing overfitting. Unlike the Lasso model, Ridge decreases all the coefficients of the correlated predictors, but does not bring them exactly to zero, thus allowing for the use of the information of all the variables included in the model. Mathematically, the Ridge regression seeks to minimize the sum of the squared residuals and a proportional penalty to the square of the norm of the coefficients, as expressed in the following Eq. (3):

$$\hat{\beta}^{ridge} = \text{argmin}_{\beta}\left\{\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\right\} \tag{3}$$

This formulation allows for more stable and robust estimates, especially in contexts where there is multicollinearity or a large number of predictor variables [32], [31].

3) Lasso regression model: Lasso regression is an extension of the least squares (OLS) model that incorporates a penalty on the sum of the absolute values of the coefficients. Unlike the Ridge model, Lasso can reduce some coefficients to exactly zero, which allows the elimination of variables that do not contribute to the model. This is a feature that is useful when working with a large number of variables. Mathematically, the Lasso model seeks to minimize the sum of the squared errors together with a penalty applied to large coefficients, as expressed in the following Eq. (4):

$$\hat{\beta}^{lasso} = \text{argmin}_{\beta}\left\{\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\} \tag{4}$$

This formulation favors simpler and more generalized models, avoiding overfitting, and improving interpretability by selecting only the most relevant variables [32], [31].

C. Data Pre-Processing

Data pre-processing was performed. Since the data had empty records and outliers, as shown in Fig. 2, that could affect the quality of the predictive model; the rows that contained null data in at least one of the variables PM_{10}, $PM_{2.5}$, or NO_2 were cut. Likewise, only complete records were kept to ensure that the model had consistent and representative data for the analysis. Subsequently, outliers were identified and removed by applying Tukey's criterion based on the interquartile range (IQR), considering as an outlier any observation outside the interval $[Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR]$, where Q_1 and Q_3 correspond to the 30th and 70th percentiles of each variable, respectively, and $IQR = Q_3 - Q_1$. This filtering was applied sequentially to the three variables. In this way, representative and adequate data for predictive analysis were guaranteed.

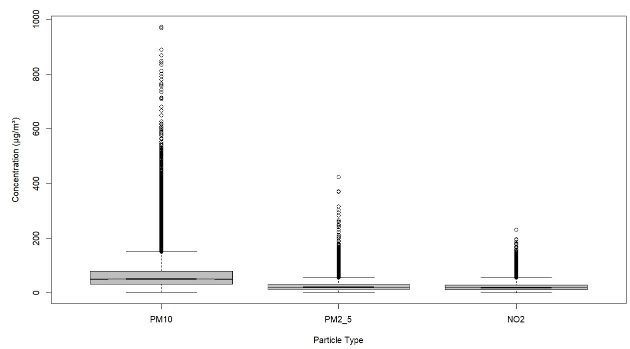

Fig. 2. Initial distribution of PM_{10}, $PM_{2.5}$ and NO_2.

D. Exploratory Data Analysis (EDA)

Subsequently, an Exploratory Data Analysis (EDA) was performed, a fundamental step due to the large volume of information. The EDA allowed for the identification of patterns, correlations, and possible anomalies in the data related to the concentrations of PM_{10}, $PM_{2.5}$, and NO_2. In addition, visualization and descriptive statistics tools were used to better understand the distribution and behavior of the variables studied, using box diagrams and histograms.

E. Data Splitting

A division of the data was carried out: the data were divided into two complementary subsets: 80% of the observations integrated the training set, used to adjust the parameters of the regression models, and the remaining 20% constituted the test set, used to evaluate the predictive performance of the final solutions.

F. Cross-Validation Procedure

To select the optimal regularization parameter λ for the Ridge and Lasso models, k-fold cross-validation was applied on the training set ($k = 10$). In this procedure, the training data are randomly partitioned into k equally sized folds; the model is fitted on $k-1$ folds and evaluated on the remaining fold, and this process is repeated k times so that each fold serves once as the validation set. The mean squared error (MSE) is averaged across the k iterations for each candidate value of λ, producing the cross-validation curve shown in Fig. 6 and 8.

Two values of λ were considered for each model: λ_{min}, the value that minimizes the cross-validated MSE, and λ_{1se}, the largest value of λ within one standard error of λ_{min}. The latter was selected to fit the final models, following the one-standard-error rule, which favors a more parsimonious model with a comparable predictive performance and reduces the risk of overfitting to the training set. A fixed random seed (`set.seed(1235)`) was used to ensure the reproducibility of the fold partitioning.

G. Model Evaluation

Finally, performance metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R^2) were used to evaluate the accuracy and robustness of the model.

H. Software Implementation

All statistical analyses and predictive modeling were performed in the R language (version 4.5.1, 2025-06-13), using the packages detailed in Table II.

TABLE II. R PACKAGES USED FOR MODEL IMPLEMENTATION

Package	Version	Use
dplyr	1.2.1	Data manipulation
tidyr	1.3.2	Data transformation
readr	2.2.0	Dataset loading
tibble	3.3.1	Data structures
ggplot2	4.0.3	Results visualization
scales	1.4.0	Logarithmic scales in plots
corrplot	0.95	Correlation matrix
Hmisc	5.2-6	Correlation computation
reshape2	1.4.5	Data preparation for EDA
corrr	0.4.5	Correlation analysis
scatterplot3d	0.3-45	3D model visualization
skimr	2.2.2	Exploratory data analysis
DataExplorer	0.9.0	Exploratory data analysis
glmnet	5.0	Ridge and Lasso model fitting
nortest	1.0-4	Normality test

IV. RESULT

A. Data Cleaning

The process began with the cleaning of the dataset, first eliminating the records with null values and rows, which reduced the dataset to 209,121 rows. Subsequently, outliers were removed by applying Tukey's criterion based on the interquartile range (IQR), resulting in a final dataset of 169,308 observations, equivalent to the removal of 19.04% of the data as outliers. Fig. 3 shows the new distribution of PM_{10}, $PM_{2.5}$, and NO_2 concentrations, where a significant reduction in outliers is observed, which represents a more accurate and adequate dataset for analysis.

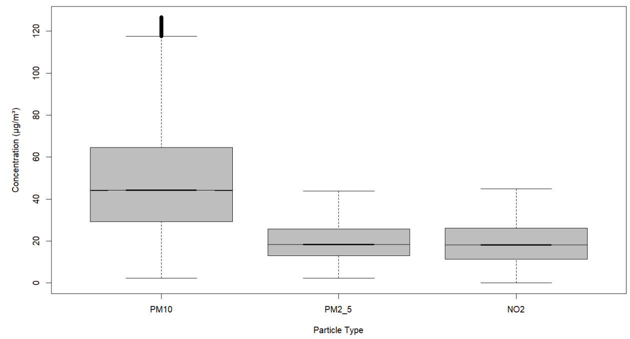

Fig. 3. Final distribution of PM_{10}, $PM_{2.5}$ and NO_2 concentrations.

B. Statistical Analysis

The statistical results presented in Table III show that the concentrations of atmospheric pollutants PM_{10}, $PM_{2.5}$, and NO_2 evidence considerable variability in their levels throughout the analyzed period. In the case of PM_{10}, the values range between 2.31 and 126.74 μg/m^3, with a median of 44.18 μg/m^3 and a mean of 49.44 μg/m^3, indicating the presence of some high values but, in general, a moderate central tendency. For $PM_{2.5}$, concentrations vary between 2.30 and 43.94 μg/m^3, with a median of 18.40 μg/m^3 and a mean of 19.99 μg/m^3, also reflecting a slight asymmetry toward higher values. As for NO_2, the levels recorded range from 0.00 to 44.90 ppb, with a median of 18.20 ppb and a mean of 19.30 ppb, showing similar behavior in terms of dispersion and central tendency. In all cases, the interquartile range reveals that most of the data is concentrated at relatively low values, although there is a proportion of measurements that reach significantly higher levels. These results allow us to appreciate the diversity and fluctuation of pollutants in the air, highlighting the importance of continuing to monitor these variables to better understand their impact on environmental quality and population health.

TABLE III. DESCRIPTIVE STATISTICS OF PM_{10}, $PM_{2.5}$ AND NO_2 ON THE ANALYZED PERIOD

Particle	Descriptive Statistics			
	Minimum	Mean	Median	Maximum
PM_{10} (μg/m^3)	2.31	49.44	44.18	126.74
$PM_{2.5}$ (μg/m^3)	2.30	19.99	18.40	43.94
NO_2 (ppb)	0.00	19.3	18.2	44.9

C. Correlation Matrix

The correlation matrix presented in Fig. 4 shows the linear relationship between the concentrations of PM_{10}, $PM_{2.5}$, and NO_2 particles. In this matrix, a moderate positive correlation can be observed between PM_{10} and $PM_{2.5}$ ($r = 0.48$), indicating that both variables tend to increase together; however, their relationship is not strong enough to consider them redundant. On the other hand, there is a correlation between PM_{10} and NO_2 ($r = 0.17$), as well as between $PM_{2.5}$ and NO_2 ($r = 0.23$). Both correlations are low, suggesting that NO_2 presents a relatively independent behavior with respect to the fractions of particulate matter analyzed. These results indicate that there is no significant multicollinearity between the variables analyzed, allowing their use in predictive models. In addition, the relative independence of NO_2 with respect to PM_{10} and $PM_{2.5}$ may be associated with differences in emission sources or atmospheric processes that affect each pollutant.

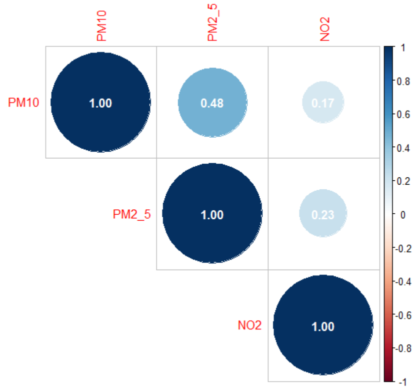

Fig. 4. Correlation matrix of PM_{10}, $PM_{2.5}$ and NO_2 particles.

D. Multiple Linear Regression Equation

In this study, a multiple linear regression model was built using the Ordinary Least Squares method (OLS) to predict PM_{10} concentrations (dependent variable) based on $PM_{2.5}$ and NO_2 concentrations (independent variables). The general form of the model is represented in equation (5):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{5}$$

where, Y represents the concentration of PM_{10} (dependent variable), X_1 corresponds to the concentration of $PM_{2.5}$ and X_2 to the concentration of NO_2 (both independent variables); β_0 is the intercept, β_1 and β_2 are the regression coefficients associated with each independent variable, and ε is the error term. The model was adjusted using the training set (80%). As a result, the following equation (6) was obtained:

$$Y_{PM_{10}} = 19.237093 + 1.344208 \cdot PM_{2.5} + 0.172471 \cdot NO_2 \tag{6}$$

In the model, it is observed that the intercept (19.237093 μg/m^3) corresponds to the base level of PM_{10} when $PM_{2.5}$ and NO_2 are zero. The $PM_{2.5}$ coefficient (1.344208) indicates that, keeping NO_2 constant, each increase of 1 μg/m^3 in $PM_{2.5}$ raises the prediction of PM_{10} by 1.344208 μg/m^3. For its part, the NO_2 coefficient (0.172471) reflects an increase of 0.172471 μg/m^3 in PM_{10} for each additional unit of NO_2, with fixed $PM_{2.5}$. The value of $R^2 = 0.2335$ shows that approximately 23.4% of the variability in PM_{10} is explained by these two variables.

Fig. 5 shows how, along the $PM_{2.5}$ axis, each increase of 1 μg/m^3 raises PM_{10} by 1.344 μg/m^3, while the effect of NO_2 is milder (0.172 μg/m^3 per unit). The low corner of the surface ($PM_{2.5}$ = 5, NO_2 = 0) predicts $PM_{10} \approx 25$ μg/m^3 and the high corner ($PM_{2.5}$ = 45, NO_2 = 45) approximates $PM_{10} \approx 85$ μg/m^3, confirming that $PM_{2.5}$ has a stronger impact on PM_{10} concentrations.All model coefficients were statistically significant ($p < 0.001$), the intercept ($t = 122.74$), $PM_{2.5}$ ($t = 212.65$), and NO_2 ($t = 30.64$), confirming that both variables contribute significantly to the prediction of PM_{10}. The 95% confidence intervals support this precision, the $PM_{2.5}$ coefficient ranges between 1.332 and 1.357 μg/m^3, and the NO_2 coefficient ranges between 0.161 and 0.183 μg/m^3. Since the two intervals do not overlap, this confirms that $PM_{2.5}$ constitutes the predictor of greatest relative importance for estimating PM_{10}.

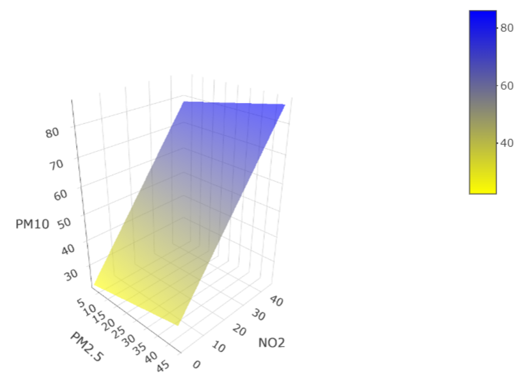

Fig. 5. Multiple linear regression PM_{10}, $PM_{2.5}$ and NO_2.

E. Optimization of the model using the Ridge technique

The adjustment of the Ridge model was carried out by dividing the data into training (80 %) and test (20 %) sets, using ridge regression to control overfitting and improve the stability of the coefficients. To select the optimal regularization parameter λ, cross-validation was applied. Fig. 6 shows how the mean squared error (MSE) varies when modifying λ, highlighting the minimum value found ($\lambda_{\min} = 1.258057$) and the value suggested by the rule of a standard deviation ($\lambda_{1se} = 3.500618$). This last value of λ was used to adjust the final model, and it was observed that the coefficients estimated by Ridge (for example, 1.1844 for $PM_{2.5}$ and 0.1805 for NO_2) are lower than those of the classic linear regression model, evidencing the effect of the penalty.

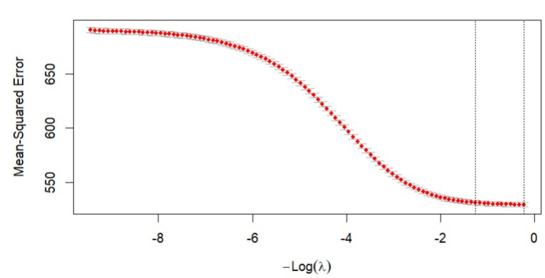

Fig. 6. Regularization parameter Lambda – Ridge model.

Fig. 7 below shows how the coefficients of the predictor variables $PM_{2.5}$ and NO_2 change as the regularization parameter λ increases. This graph shows the effect of the penalty applied by Ridge, where larger coefficients tend to decrease progressively, while smaller coefficients stabilize. This behavior results in more robust models, minimizing the risk of overfitting.

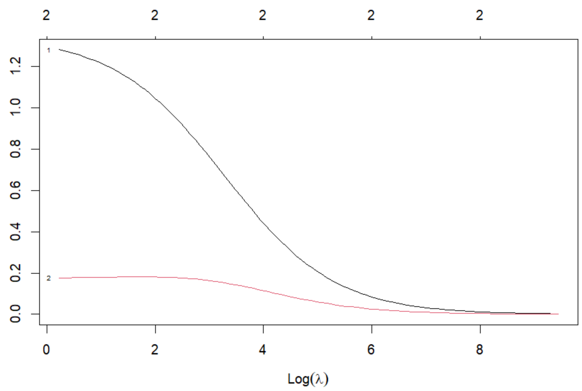

Fig. 7. Evolution of the coefficients of the ridge model as a function of the regularization parameter λ.

After adjusting the final model using the value λ_{1se} (3.500618) recommended by the cross-validation, the following equation (7) was adjusted to predict PM_{10}:

$$Y_{PM_{10}} = 22.2783 + 1.1844 \cdot PM_{2.5} + 0.1805 \cdot NO_2 \quad (7)$$

The performance of the Ridge model, evaluated by means of the mean squared error (MSE) and the coefficient of determination (R^2), was similar in both training and testing (MSE $\approx$ 531 and 521; $R^2 \approx 0.23$), indicating a good generalization but a limited explanatory capacity, since only about 23% of the variability of PM_{10} is explained. This regularization reduced the magnitude of the coefficients without eliminating them, which made it possible to obtain a more stable and less noise-sensitive model.

F. Optimization of the Model Using the Lasso Technique

The adjustment of the Lasso model was carried out in a similar way to the Ridge model, with the division of the data into training (80 %) and test (20 %) sets. In this case, Lasso regression was applied in order to control overfitting and improve the interpretability of the model. To determine the optimal regularization parameter λ, cross-validation was also used. Fig. 8 shows how the mean squared error (MSE) varies as the value of λ changes. It is observed that the MSE decreases for low penalty values and then tends to stabilize, highlighting the minimum value found ($\lambda_{\min} = 0.06871855$) and the value suggested by the rule of one standard deviation ($\lambda_{1se} = 1.119941$). This last value of λ was used to adjust the final model, prioritizing lower complexity without loss of precision in the results. In addition, it was observed that the coefficients estimated by Lasso allowed the coefficients to be progressively reduced compared to Ridge and OLS, which evidences the ability to simplify the model while maintaining a good predictive capacity.

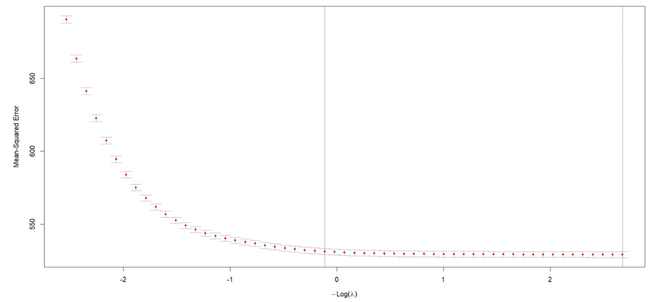

Fig. 8. Regularization parameter Lambda – Lasso model.

Fig. 9 shows how the coefficients of the predictor variables ($PM_{2.5}$ and NO_2) change as the regularization parameter λ increases. It is observed that as the penalty increases, the coefficients decrease; some of these tend to be values close to zero. This behavior reflects the main characteristic of the Lasso model, which makes a selection of variables, reducing the influence of the less relevant ones.

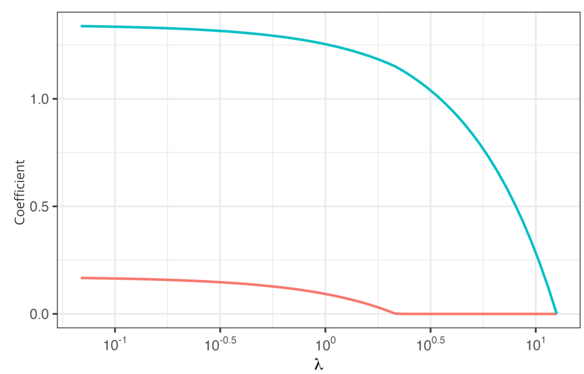

Fig. 9. Evolution of the coefficients of the Lasso model as a function of the regularization parameter λ.

After adjusting the final Lasso model using the λ_{1se} value (1.119941) recommended by the cross-validation, the following Eq. (8) was adjusted to predict PM_{10}:

$$Y_{PM_{10}} = 22.9766 + 1.2434 \cdot PM_{2.5} + 0.0831 \cdot NO_2 \quad (8)$$

The performance of the Lasso model, evaluated by means of the mean squared error (MSE) and the determination coefficient (R^2), was similar in both training and testing (MSE $\approx$ 531 and 521; $R^2 \approx 0.23$), indicating a good generalization, but a limited explanatory capacity, since only about 23% of the variability of PM_{10} is explained. Unlike Ridge, the model penalty reduced the coefficients toward near-zero values, reflecting the simplicity and selection of important variables.

G. Benchmarking of Performance Between OLS, Ridge and Lasso

The statistical results presented in Table IV show a very similar performance among the three models evaluated, considering performance metrics such as the MAE, RMSE, MSE, and the coefficient of determination R^2. The OLS model

TABLE IV. Comparison of Performance Metrics Between the Models

Model	Performance Metrics			
	MAE	RMSE	MSE	R^2
OLS	18.34225	22.77559	518.7273	0.2365
Ridge	18.39823	22.82010	520.7569	0.2335
Lasso	18.42202	22.81856	520.6867	0.2336

obtained slightly lower errors, while Ridge and Lasso presented close metrics. Together, the models explain about 23% of the variability of PM_{10}.This relatively low R^2 value is mainly explained by the simplicity of the model, which uses only two predictor variables, and by the absence of factors with recognized influence on PM_{10}, such as meteorological variables and specific emission sources.

Compared with results reported in previous studies, the performance obtained here is modest yet consistent with expectations for a parsimonious linear model using only two pollutant predictors. Olarte et al. [23], applying Ridge and Lasso regression to predict $PM_{2.5}$ AQI in Junín using meteorological variables, reported a considerably lower R^2 of 0.148 and an RMSE of 25.36, suggesting that the present model's R^2 of approximately 0.23 already compares favorably within the family of regularized linear approaches. In contrast, studies that incorporated richer predictor sets and nonlinear algorithms achieved substantially higher accuracy: Tang et al. [25], using a hybrid Random Forest–XGBoost model with satellite reflectance and meteorological data, reported an R^2 of 0.82, while Wu et al. [28], employing a dual-channel deep learning architecture that fused surveillance images with atmospheric and meteorological data, reached R^2 values of 0.95 and 0.90 for $PM_{2.5}$ and PM_{10}, respectively. This contrast reinforces that the performance ceiling of the present model is constrained primarily by the limited predictor set available, rather than by the choice between OLS, Ridge, and Lasso, and it further supports the recommendation to incorporate meteorological and nonlinear modeling components in future work.

V. Discussion

Our comparative analysis of the OLS, Ridge, and Lasso models reveals important findings about their performance and usefulness for the prediction of PM_{10} concentrations. The regularization techniques applied made it possible to stabilize the coefficients of the model, showing how the penalty controls overfitting by reducing the larger values and adjusting the smaller ones. This behavior, observed in the trajectories of the coefficients, demonstrates the ability of regularized models to handle data with collinearities between the predictor variables and generate more robust estimates.

The nearly identical performance between the regularized models and OLS is explained by the low correlation between $PM_{2.5}$ and NO_2 (r = 0.23) and the small number of predictors. Since Ridge and Lasso primarily correct instability caused by multicollinearity or a high number of variables.

The selection of the optimal parameter λ was carried out through cross-validation in both cases, ensuring a balance between the fit of the model and its ability to generalize. The recommended value was key to building a final model with penalized coefficients that reflect a more controlled impact of the predictor variables. However, although the adjusted models explain 23% of the variability in PM_{10} concentrations, the graphs obtained suggest that there are additional factors that were not considered and that could significantly influence the quality of the adjustment.

This highlights the importance of exploring other relevant variables, such as meteorological factors (temperature, wind speed, and precipitation), as well as incorporating detailed data on specific emission sources. The integration of these aspects in future analyses could significantly improve the explanatory and predictive capacity of the models.The evidence suggests that the lack of explanatory variables is the most determining factor, PM_{10} includes not only the fine fraction ($PM_{2.5}$) but also a coarse fraction (particles between 2.5 and 10 μm) originating from sources unrelated to NO_2 or $PM_{2.5}$. In addition, the current performance of the models suggests the opportunity to experiment with non-linear techniques, such as polynomial regression, decision trees, or neural networks, to capture more complex and non-obvious relationships in the current dataset.

On the other hand, the analysis highlights the relevance of having high-quality data and broader coverage in the monitoring stations. A richer and more representative database would allow us to characterize the variability of PM_{10} more accurately in Metropolitan Lima, benefiting not only the construction of more robust predictive models but also the design of more effective environmental policies.

VI. Conclusion

Finally, although the adjusted models have limited explanatory capacity ($R^2 \approx 0.23$), they provide an initial, interpretable tool for exploring the relationships between PM_{10} concentrations and its main determinants in Metropolitan Lima. Given this limited predictive power, the models should not be regarded as suitable for operational forecasting or as a standalone basis for regulatory decisions; rather, they represent a transparent starting point that can inform future, more robust modeling efforts. Future research should incorporate meteorological variables (e.g., wind speed, temperature, and precipitation), traffic-related information, land-use characteristics, and advanced machine learning algorithms, such as ensemble methods and deep learning, to improve predictive accuracy and better capture the nonlinear and spatio-temporal dynamics of PM_{10}. Complementing these modeling efforts with emission mitigation strategies, awareness campaigns, and expansion of urban infrastructure could contribute to reducing PM_{10} exposure and improving the respiratory health of the population in Metropolitan Lima.

VII. Declaration of the Use of AI

To support the writing and linguistic coherence of this paper, the authors made use of a generative artificial intelligence tool (ChatGPT and Gemini). Its application was at all times supervised by the authors, who reviewed, adjusted, and validated the generated content to ensure its technical and academic rigor. Therefore, the full responsibility for the content, ideas, and conclusions presented in this manuscript lies exclusively with the authors.

REFERENCES

[1] B. Choubin, M. Abdolshahnejad, E. Moradi, X. Querol, A. Mosavi, S. Shamshirband, and P. Ghamisi, "Spatial hazard assessment of the PM10 using machine learning models in Barcelona, Spain," *Science of The Total Environment*, vol. 701, p. 134474, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969719344651

[2] M. Mostaghim, A. Imam, A. Fallatah, A. R. B. Lomer, M. Maleki, J. Wang, I. D. Stewart, and N. Moradpour, "Seasonal outdoor PM10 changes based on the spatial local climate zone distribution," *Urban Climate*, vol. 58, p. 102148, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212095524003456

[3] M. Maleki, M. Asadi, S. Z. Naghadehi, A. Khosravi, J. Wang, I. D. Stewart, and M. Shakeryari, "Detecting local climate zone change and its effects on PM10 distribution using fuzzy machine learning in Tehran, Iran," *Urban Climate*, vol. 49, p. 101506, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212095523001001

[4] D. C. Joshi, P. Negi, S. Devi, H. Lohani, R. Kumar, M. Gupta, and L. C. Ming, "Fine particulate matter (PM2.5, PM10): A silent catalyst for chronic lung diseases in India; a comprehensive review," *Environmental Challenges*, p. 101215, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667010025001349

[5] Y. Mohammadi, O. Zandi, M. Nasseri, and Y. Rashidi, "Spatiotemporal modeling of PM10 via committee method with in-situ and large scale information: Coupling of machine learning and statistical methods," *Urban Climate*, vol. 49, p. 101494, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212095523000883

[6] V. H. Borja-Aburto, D. P. Loomis, S. I. Bangdiwala, C. M. Shy, and R. A. Rascón-Pacheco, "Ozone, suspended particulates, and daily mortality in Mexico City," *American Journal of Epidemiology*, vol. 145, no. 3, pp. 258–268, 1997.

[7] C. Girotti, L. Fernando Kowalski, T. Silva, E. Correia, A. R. Prata Shimomura, F. Akira Kurokawa, and A. Lopes, "Air pollution dynamics: The role of meteorological factors in PM10 concentration patterns across urban areas," *City and Environment Interactions*, vol. 25, p. 100184, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590252024000448

[8] M. Czwojdzińska, M. Terpińska, A. Kuźniarski, S. Płaczkowska, and A. Piwowar, "Exposure to PM2.5 and PM10 and COVID-19 infection rates and mortality: A one-year observational study in Poland," *Biomedical Journal*, vol. 44, no. 6, Supplement 1, pp. S25–S36, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2319417021001669

[9] J. Lee, P. D. Weerasinghe-Mudiyanselage, B. Kim, S. Kang, J.-S. Kim, and C. Moon, "Particulate matter exposure and neurodegenerative diseases: A comprehensive update on toxicity and mechanisms," *Ecotoxicology and Environmental Safety*, vol. 266, p. 115565, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0147651323010692

[10] J. Santamaria, Y. Gilaberte, L. Prudkin, and J. Piquero-Casals, "[artículo traducido] la contaminación, un factor relevante del exposoma en el envejecimiento cutáneo y el papel de la fotoprotección multibeneficio," *Actas Dermo-Sifiliográficas*, vol. 116, no. 6, pp. T611–T620, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001731025002212

[11] World Health Organization, "WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide," World Health Organization, Geneva, Tech. Rep., 2021. [Online]. Available: https://iris.who.int/handle/10665/345329

[12] P. S. Ha, C. V. R., and P. M. Gb., "Impacto de la contaminación del aire por PM10 sobre la mortalidad diaria en Temuco," *Revista Médica de Chile*, vol. 134, pp. 754–761, 2006.

[13] P. Aldunate, O. Paz, and K. Halvorsen, "Los efectos de la contaminación atmosférica por PM10 sobre la salud en la ciudad de La Paz - Bolivia (3650 m.s.n.m.)," *Acta Nova*, vol. 3, pp. 422 – 442, 06 2006. [Online]. Available: http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S1683-07892006000100019&nrm=iso

[14] D. Roy, J. Kim, M. Lee, S. Kim, and J. Park, "PM10-bound microplastics and trace metals: A public health insight from the Korean subway and indoor environments," *Journal of Hazardous Materials*, vol. 477, p. 135156, 2024.

[15] A. R. S. Abadi, N. H. Hamzeh, D. G. Kaskaoutis, J.-F. Vuillaume, K. A. Shukurov, and M. Gharibzadeh, "Spatio-temporal distribution of PM2.5 and PM10 concentrations and assessment of public health risk in the three most polluted provinces of Iran," *Sustainability*, vol. 17, no. 1, 2025. [Online]. Available: https://www.mdpi.com/2071-1050/17/1/44

[16] N. Seihei, M. Farhadi, A. Takdastan, P. Asban, F. Kiani, and M. J. Mohammadi, "Short-term and long-term effects of exposure to PM10," *Clinical Epidemiology and Global Health*, vol. 27, p. 101611, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213398424001076

[17] L. Xie, J. He, R. Lei, M. Fan, and H. Huang, "Accurate and efficient prediction of atmospheric PM1, PM2.5, PM10, and O3 concentrations using a customized software package based on a machine-learning algorithm," *Chemosphere*, vol. 368, p. 143752, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045653524026535

[18] A. L. Duarte, I. L. Schneider, P. Artaxo, and M. L. Oliveira, "Spatiotemporal assessment of particulate matter (PM10 and PM2.5) and ozone in a Caribbean urban coastal city," *Geoscience Frontiers*, vol. 13, no. 1, p. 101168, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1674987121000323

[19] DIGESA, "2do estudio de saturación de Lima Metropolitana y el Callao 2011 (en línea)," http://www.digesa.minsa.gob.pe/depa/informes_tecnicos/Estudio%20de%20Saturacion%202012.pdf, 2012, consultado 3 febrero del 2015.

[20] X. Ma, H. Liu, and Z. Peng, "Improving PM2.5 and PM10 predictions in China from WRF-chem through a deep learning method: Multiscale depth-separable UNet," *Environmental Pollution*, vol. 364, p. 125344, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026974912402061X.

[21] Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI), "Monitoreo de los contaminantes del aire en Lima Metropolitana," https://datosabiertos.gob.pe/dataset/monitoreo-aire-lima, 2025, [Online; accessed 16-June-2025].

[22] R. Buvana, S. A. A. Sathya, C. Rajshekhar, J. Cm, and D. K. Jashna, "Air pollution prediction with machine learning," in *2022 International Conference on Computational Science and Technology (ICCST)*. Malaysia: IEEE, Nov. 2022, pp. 1–4.

[23] E. Olarte, J. Gutierrez, G. Roque, J. Soria, H. Fernandez, J. E. Perez Carpio, and O. Poma Porras, "Predictive model with machine learning for environmental variables and PM2.5 in Huachac, Junín, Perú," *Atmosphere*, vol. 16, p. 323, 03 2025.

[24] J. J. Soria, A. O. Cardenas, and L. S. Peña, "Linear regression with PM2.5 and PM10 concentration for air quality in East Lima, Peru," in *Artificial Intelligence Algorithm Design for Systems*, R. Silhavy and P. Silhavy, Eds. Cham: Springer Nature Switzerland, 2024, pp. 519–533.

[25] Y. Tang, J. Deng, X. Cui, Z. Liu, L. Yang, S. Zhang, and Y. Liang, "High-resolution spatial prediction of daily average PM2.5 concentrations in Jiangxi Province via a hybrid model integrating random forest and XGBoost," *Atmosphere*, vol. 16, no. 12, p. 1317, 2025.

[26] K. Gora and M. Rzeszutek, "Comparison of selected ensemble supervised learning algorithms used for meteorological normalisation of particulate matter (PM10)," *Sustainability*, vol. 17, no. 12, p. 5274, 2025.

[27] H. Tian, H. Kong, and C. Wong, "A novel stacking ensemble learning approach for predicting PM2.5 levels in dense urban environments using meteorological variables: A case study in Macau," *Applied Sciences*, vol. 14, no. 12, p. 5062, 2024.

[28] Y. Wu, X. Wang, M. Wang, X. Liu, and S. Zhu, "Time-series forecasting of PM2.5 and PM10 concentrations based on the integration of surveillance images," *Sensors*, vol. 25, no. 1, p. 95, 2025.

[29] W. Abuouelezz, N. Ali, Z. Aung, A. Altunaiji, S. B. Shah, and D. Gliddon, "Exploring PM2.5 and PM10 ML forecasting models: A comparative study in the UAE," *Scientific Reports*, vol. 15, p. 9797, 2025.

[30] Instituto Nacional de Estadística e Informática (INEI), "Al 2024 la población peruana proyectada alcanza los 34 millones de habitantes," https://www.gob.pe/institucion/inei/noticias/987317-al-2024-la-poblacion-peruana-proyectada-alcanza-los-34-millones-de-habitantes,

2024, documento fuente: Situación de la Población Peruana, 2024: Una mirada de la diversidad étnica. Accessed: August 2026.

[31] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[32] L. A. Geraldo-Campos, J. J. Soria, and T. Pando-Ezcurra, "Machine learning for credit risk in the reactive Peru program: A comparison of the lasso and ridge regression models," *Economies*, vol. 10, no. 8, 2022. [Online]. Available: https://www.mdpi.com/2227-7099/10/8/188