Machine Learning Models for Fine Particulate Matter (PM2.5) Prediction: A Case Study in Bac Ninh Province, Vietnam

Hai, P. H.,[1*] Minh, T. H.,[1] Luận, V. N.,[2] Trung, T. V.,[3] Ha, H.T.T.[1] and Long, D. N.[1]
[1]Vietnam Institute of Geodesy and Cartography, Ministry of Natural Resources and Environment, 479 Hoang Quoc Viet Street, Co Nhue 1, Bac Tu Liem District, Hanoi, Vietnam
E-mail: Phamminhhai.vigac@gmail.com,* hoangminh0108@yahoo.com
[2]Pollution Control Department, Ministry of Natural Resources and Environment, 10 Ton That Thuyet, Bac Tu Liem District, Hanoi Vietnam, E-mail: luan19862002@gmail.com
[3]University of Engineering and Technology, Hanoi National University, 144 Xuan Thuy Street, Cau Giay District, Ha Noi, Vietnam, E-mail: trungtv@fimo.edu.vn
*Corresponding Author

DOI: https://doi.org/10.52939/ijg.v19i12.2975

Abstract
The Vietnamese government has established 37automatic ground air quality stations in three northern provinces, including the capital Hanoi, Bac Ninh, and Bac Giang, to provide hourly averaged PM2.5 data. However, these stations are still sparsely and unevenly distributed. This study seeks to develop an approach utilizing advanced machine learning models to forecast PM2.5 air pollution, with a specific focus on Bac Ninh Province in Vietnam. Historical relationship between the independent variable (input variable) and the dependent variable (target variable) to construct a time series model for PM2.5 prediction. The research findings reveal that the AutoARIMA model demonstrates superior performance, exhibiting better accuracy compared to other models (R^2 = 0.81, MSE = 3.5, MAE = 23.24, and RMSE = 0.17). The concentration of PM2.5 dust in Bac Ninh Province reaches a sensitive level that poses a threat to human health (100 micrograms/m^3). Thuan Thanh, a southern district of Bac Ninh Province, registers the highest pollution level in the province, with a dust concentration value of 110 micrograms/m^3. The research methodology is scientifically contributing to raising public awareness about air quality for both individuals and local government stakeholders.

Keywords: Cross-Validation, Fine Dust, Grid Search, Machine Learning, PM2.5

1. Introduction

Air pollution constitutes a global challenge, with an annual toll of 7 million deaths attributed to exposure. The majority of fatalities linked to air pollution are documented in Southeast Asia and the Western Pacific [1]. In Vietnam, major urban centers such as Hanoi and Ho Chi Minh City consistently register unhealthy air pollution levels on the Air Quality Index (AQI), ranging from 150 to 200. This surpasses the World Health Organization's (WHO) annual air quality standard value by a staggering 21.9 times and has seen a significant escalation from 2010 to the present. The primary concern revolves around fine particulate matter measuring less than 2.5 mm (PM2.5), a factor that heightens the risk of cardiovascular and respiratory illnesses [2].

Despite the Vietnamese government's issuance of stringent air pollution control regulations, exemplified by Circular No.10/2021/TT-BTNMT outlining technical provisions for environmental observation and management of environmental quality information and data, air pollution continues to worsen. In the period from 2020 to 2023, the government established 37 automatic ground air quality stations across three northern provinces namely Hanoi, Bac Ninh, and Bac Giang to furnish hourly averaged PM2.5 concentration data [3]. Notably, half of these stations are situated in Bac Ninh province, yet their distribution remains sparse and uneven. In Vietnam, efforts have been made by scientists to create PM2.5 prediction maps.

However, two limitations in the existing ground air quality stations must be addressed to enhance PM2.5 prediction capabilities: the presentation of PM2.5 data as points and the absence of a raster map depicting air pollution; additionally, there is a lack of time series prediction methods to effectively interpolate air pollution [3].

In recent times, discussions on air pollution forecasts have become widespread, particularly with a focus on the predictive capabilities of machine learning (ML) models. ML, being an interdisciplinary field that encompasses statistics, data science, and computing, has garnered significant interest across various domains of study [4]. What sets our study apart is the utilization of meteorological time series data sources and ML models for learning and forecasting spatiotemporal PM2.5 concentrations. Traditional approaches have seen the widespread application of Artificial Neural Networks (ANN) [5] and [6], Support Vector Machine (SVM) [7], and Random Forest (RF) [8] [9] and [10] for facilitating dust size retrieval. However, the intricacies of complex datasets necessitate the use of sophisticated machine learning models [11]. The selection of the most suitable ML model in our study hinges on the availability of historical data and the relationship between the training and test variables [12]. Employing complex ML models offers several advantages, including the ability to comprehend and execute various tasks based on experience in searching for optimal parameters and appropriate ML models. This involves utilizing a weighted average of past observations, which is well-suited for long-term time series data and evident seasonal patterns, as well as accommodating sparse data series and time series with discontinuous variation [13].

Hence, the primary goal of this study is to introduce a novel approach employing advanced machine learning (ML) models to forecast spatiotemporal PM2.5 air pollution, focusing on the case of Bac Ninh province in Vietnam. The key contributions of this research include:

1. The proposal of a methodology for preliminary input data analysis, addressing scenarios with no data, values of 0, and -9999 for PM2.5 input values.
2. Investigation of ML models categorized into three types: Level time series, Trend time series, and Seasonal time series, with the aim of selecting the most suitable ML model for the input dataset.
3. Implementation of cross-validation as a method to train, test, and validate ML models for PM2.5 forecasting.
4. Establishment of short-term PM2.5 forecast maps within the study area.

2. Study Are and Data Source

2.1 Study Area

Situated north of the Hanoi capital, Bac Ninh spans 822.71 square kilometers and is home to a population of 1,488,250 (as of 2022) [14] (Figure 1). The topography features a plain terrain that slopes from north to south, with hill and mountain areas reaching approximately 400 meters above sea level. Presently, Bac Ninh grapples with pollution stemming from rapid urbanization, expansive industrial zones, and craft villages. Following Hanoi capital, Bac Ninh province holds the second position in the air pollution index table, boasting an average AQI of 171. Pollution primarily results from anthropogenic activities such as transportation, industrial emissions, and trade villages within the province, adversely impacting the health of residents and hindering economic development.

2.2 Ground Based Measurements

The study acquired hourly PM2.5 data from ground air quality stations covering the period from January 1, 2022, to June 30, 2023. This data was sourced from the National Environmental Observation Center's website (https://cem.gov.vn/), the division of the Vietnam Ministry of Natural Resources and Environment (MONRE). The spatial distribution of PM2.5 monitoring stations across the province is uneven, with a concentration in the western region. During data preprocessing, abnormal PM2.5 values exceeding 300 micrograms/m^3 were excluded in the analysis. Additionally, observations of less than 12 hours a day were also excluded to ensure data quality. The temporal resolution of the datasets used in this study is 1 hour, enabling the calculation of daily average PM2.5 concentrations for input into ML modeling. The time stamp is linked to short-term forecasts at intervals of 1, 2, 4, 8, 16, 24, and 48 hours.

3. Methodology

3.1 The research Flow

The study aims to enhance the uniformity of input datasets and identify the optimal model for forecasting PM2.5, leading to the creation of forecast maps for the study area. Initial data analysis was conducted to ensure input dataset consistency. Subsequently, three ML models, namely Level time-series, Trend time-series, and Seasonal-time series, were employed to determine the most suitable model.

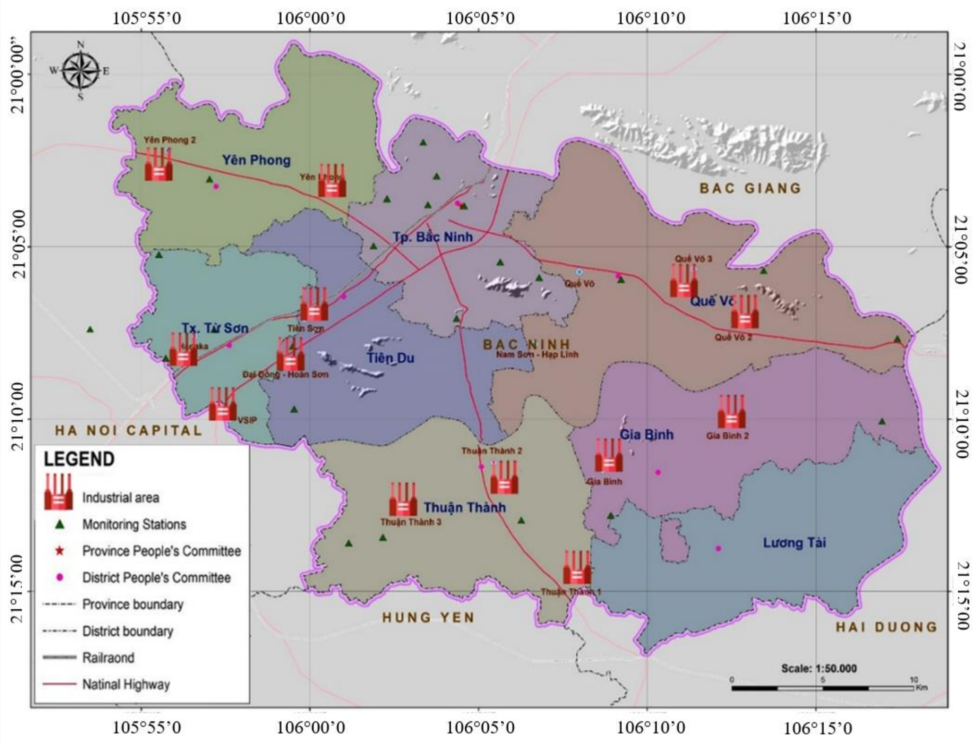


Figure 1: Bac Ninh province, Viet Nam

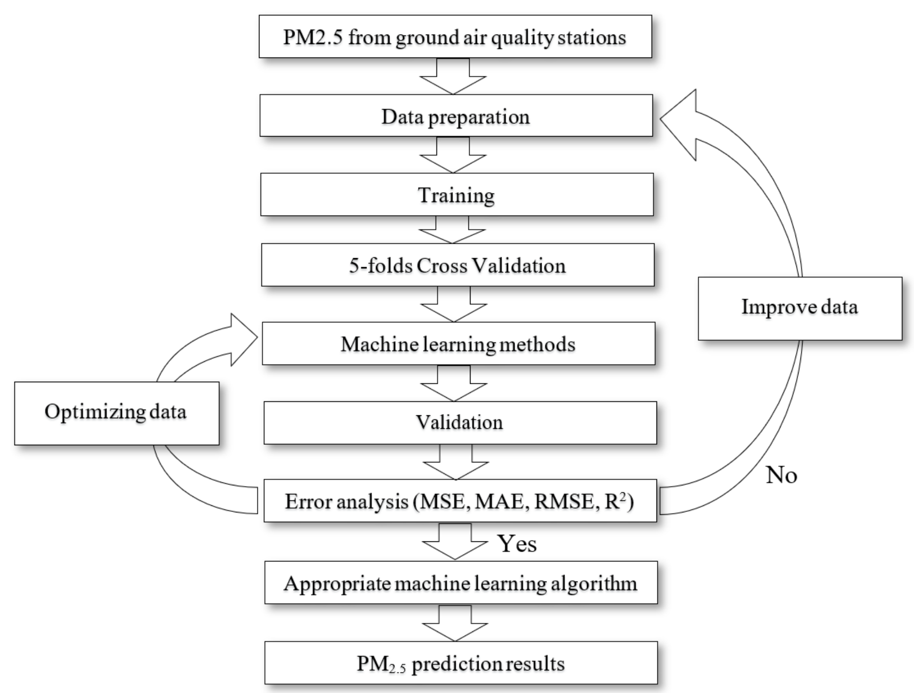

PM2.5 from ground air quality stations
Data preparation
Training
5-folds Cross Validation
Machine learning methods
Validation
Error analysis (MSE, MAE, RMSE, R^2)
Yes
Appropriate machine learning algorithm
PM$_{2.5}$ prediction results
Optimizing data
Improve data
No

Figure 2: Study workflow

Cross-validation was employed for the training, testing, and validation of the ML models in PM2.5 forecasting. Ultimately, PM2.5 forecast maps at a 1:50,000 scale was generated. Figure 2 provides a visual representation of the PM2.5 retrieval process through the ML models.

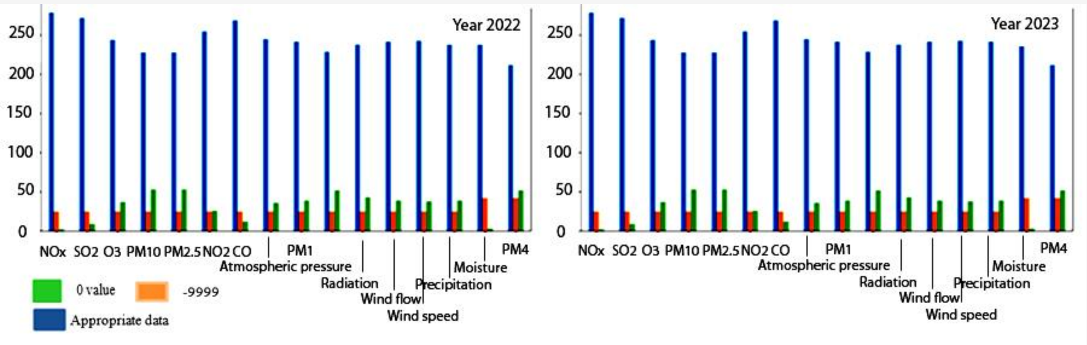


Figure 3: Status of input data

3.2 Preliminary Data Analysis

Data plays a crucial role in the effectiveness of machine learning methods, significantly influencing the outcomes derived for ML models. Therefore, when implementing a machine learning model, the key priority is to curate a suitable dataset that facilitates effective model learning. It is imperative that the data accurately represents new cases to ensure generalizability. In this study, supervised learning methods were employed, necessitating the careful definition and preparation of training data for both input and output datasets. The timestamp used in the study is set at 1 hour. However, the collected datasets contained records with 0, null, and -9999 values (indicating errors). The graphical representation illustrates the distribution of attribute values (the column), with -9999 values accounting for 10.89% and 8.31% in 2022 and 2023, respectively. Additionally, null data corresponds to 8.86% and 16.25% for the same years (Figure 3). Data preparation is an essential preliminary procedure aimed at rendering the dataset more suitable for machine learning [12]. In this study, three interpolation methods were proposed to address 0, null, and -9999 values, outlined as follows:

- Replacement of missing data with the daily average of measurements at one station (comprising 24 values per day). If the average value is unavailable, the corresponding data point is discarded.
- In cases where a daily average is not accessible, the study substitutes missing data with the monthly average at the respective station.
- In case a monthly average is also unavailable, the study resorts to replacing missing data with the annual average at the designated station

In this study, a supervised learning method was employed, leveraging a labeled dataset for training to classify data and predict outcomes. The training sample is defined as an input vector x_i, representing the value at time i, where i ranges from 1 to N as defines in Equation 1.

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_N^T \end{bmatrix} = \begin{bmatrix} x_{0,1} & x_{0,2} & \cdots & x_{p-1,1} \\ x_{0,2} & x_{1,2} & \cdots & x_{p-1,2} \\ & \cdots & & \\ x_{0,N} & x_{1,N} & \cdots & x_{p-1,N} \end{bmatrix}$$

Equation 1

Where:

T represents the forecast for the next 1, 2, 4, 8, 16, 24, and 48 hours.

x_i is row of the matrix X.

x_p is column of the matrix X with $p = 0, 1, 2, . . ., p – 1$

N represents the number of observations prepared for sampling and cross-validation, utilizing a dataset spanning from January 1, 2022, to June 30, 2023.

3.3 Machine Learning Models

Time series forecasting holds significant importance within the realms of statistics and machine learning. This nomenclature is aptly chosen because these models are specifically designed for datasets with temporal elements. The foundation of time series forecasting rests on the assumption that past patterns will recur in the future. Consequently, this study involves modeling the historical relationship between the independent variable (input variable) and the dependent variable (target variable) to construct a time series model. Considering the influence of objective factors, the time series analyzed in this study exhibit distinct characteristics: 1) Long observation time series; 2) Dispersed data resulting from the uneven distribution of ground observation stations; 3) Seasonal variation; 4) Intermittent volatility; and 5) Univariate time series. [15] recommends the use of machine learning models aligned with the aforementioned characteristics for effective forecasting. Accordingly, this study proposes the most suitable machine learning models for the specified time series models.

Level Time Series: AutoRegressive Integrated Moving Average model (AutoARIMA), GARCH model, Simple Moving Average model, Exponential Smoothing models (ETS) including Simple Exponential Smoothing model (SES), Simple Smooth Optimized model (SSO), Seasonal Exponential Smoothing model (SASO), Vector Exponential Smoothing model (VectorETS), Holt's method (HOLT) models.

Intermittent Time Series: Random Walk model (RWD), Croston model including Croston Classic model (CC) and Croston Optimized model (CO), Intermittent Multiple Aggregation Algorithm model (iMAPA), Theta including Models included AutoTheta model (AutoTheta), Standard Theta model (STheta), Optimized Theta model (OTM), and Dynamic Standard Theta model (DST) model.

Seasonal Time Series: Simple Moving Average (SMA) model, Naïve model including Seasonal Naïve and Naïve models.

3.3.1 Models for level time series

The AutoRegressive Integrated Moving Average (AutoARIMA) model [16] was utilized as an Automatic forecast model, employing a long-term time series to predict future trends in PM2.5 dust concentration. The effective use of this model entails determining appropriate data, configuring parameters, and computing forecasts [15]. The model operates based on the assumption of a stationary series and constant error variance. It utilizes past values of the forecasted series, incorporating both auto-regression and moving average components. Given that many time series tend to exhibit upward or downward trends over time, obtaining a stationary series can be challenging. To address this, the series is transformed into a stationary state through differencing. To tackle these issues, the ARIMA model incorporates three sub-models: *p*, *d*, and *q*. Here, *p* represents the autoregressive part of the model, signifying the number of lagged series used for future predictions [17]. The parameter *d* indicates how many differences are required to render the series stationary. The AutoARIMA (*p*, *d*, *q*) expressed in Equation 2.

$$\varphi(B)(1-B^d)Y(t)=\delta+\vartheta(B)\varepsilon(t)$$

Equation 2

Where:

$\varphi(B)$, $\vartheta(B)$: polynomials of p, q, respectively.
Δ : a constant.
B : an operator.
$Y(t)$: a variable.
$\varepsilon(t)$: a noise at time t

Subsequently, the study employed the Generalized Autoregressive Conditional Heteroskedasticity Process (GARCH) model, initially formulated by Bollerslev [18]. This statistical model finds application in time-series data where the variance error exhibits serial autocorrelation. Within the GARCH model in Equation 3, the conditional variance is expressed as a linear function involving the square root of past observed values and previously calculated conditional variances. Notably, the model has gained prominence due to its ability to fit datasets more effectively than other models, with its parsimonious parameterization. Over the years, GARCH series have proven increasingly effective in adjusting level time series data, as they incorporate a second moment to gauge time variation.

$$\varepsilon_t \mid \psi_{t-1} \square N(0,h_t)$$

Equation 3

Where:

h_t : the conditional variance.
ψ_{t-1} : information at time t_{-1}
N : the conditional distribution

Then the GARCH [19] is determined from Equation 4.

$$h_t=\alpha_0+\sum_{t=1}^{q}\alpha_t\varepsilon_{t-1}^2+\sum_{j=1}^{p}\beta_j h_{t-1}$$

Equation 4

with $\alpha_0 > 0$, $\beta_j \geq 0$, for i=1, 2…, q; j=1, 2,…, p

The input dataset exhibits a time series format with a discernible trend fluctuating over time. To address this trend, Exponential Smoothing (ETS) models were employed [10] and [20]. Notably, ETS models represent specific instances of AutoARIMA, being non-stationary in contrast to the stationary nature of ARIMA models. In this study, various ETS models were explored, including Simple Exponential Smoothing (SES), Simple Smooth Optimized (SSO), Seasonal Exponential Smoothing (SASO), Vector Exponential Smoothing (VectorETS), and Holt's (HOLT).

SES operates by employing a weighted average of past observations, with weights diminishing exponentially. This model assigns varying levels of influence and importance to values at different times, with those closer to the forecasted time receiving higher weight than those further in the past. SES is particularly suitable for forecasting data with a subtle trend but lacks clear direction [15].

The forecast *t+1* is the estimation of average level at time *t* as expressed in Equation 5.

$$\hat{L}_t = \alpha(y_t - \hat{S}_{t-m}) + (1-\alpha)\hat{L}_{t-1}$$

Equation 5

Where:

$\hat{L}_{t-1}$: level forecast in period *t-1*

y_t : observed value at in period *t*

$\hat{S}_{t-m}$: seasonal effect

α : smoothing constant ($0<\alpha<1$)

Aishwarya [21] proposed an enhancement to SES. The SSO model was crafted by optimizing the initial level and trend of SES to minimize the Mean Squared Error (MSE) error function. The SSO model is widely recognized in the realm of time series modeling, appreciated for its intuitive functionality and adeptness in capturing seasonality. The equation for the SSO model can be derived effortlessly in just a few steps, as expressed in Equations 6 to 9.

$$\hat{L}_t = \alpha(y_t - \hat{S}_{t-m}) + (1-\alpha)\hat{L}_{t-1}$$

Equation 6

$$\hat{L}_t = \alpha(y_t - \hat{S}_{t-m}) + \hat{L}_{t-1} - \alpha)\hat{L}_{t-1}$$

Equation 7

$$\hat{L}_t = \hat{L}_{t-1} + \alpha(y_t - \hat{S}_{t-m}) - \alpha)\hat{L}_{t-1}$$

Equation 8

$$\hat{L}_t = \hat{L}_{t-1} + \alpha(y_t - \hat{S}_{t-m} - \hat{L}_{t-1})$$

Equation 9

Where:

$\hat{S}_{t-m} - \hat{L}_{t-1}$ the forecast at time *t* and time *t-1.*

Similarly, the smoothing formulation of the equation for the Seasonal Exponential Smoothing model (SASO) is presented in Equation 10.

$$\hat{S}_t = \delta(y_t - \hat{L}_t) + (1-\delta)\hat{S}_{t-m}$$

Equation 10

Where:

β: a smoothing constant ($0< \beta <1$).

VectorETS is considered to be a good model for forecasting by using Akaike Information Criterion (AIC) [15] as presented in Equation 11.

$$AIC = -2log[L]+2k$$

Equation 11

Where:

L: the likelihood of the model.

k: the total number of variances.

In addressing multiple variances, the study adopts a multivariate approach, representing them in vector form as presented in Equation 12.

$$y_t = \begin{bmatrix} y_{1,t} & y_{2,t} & \ldots & y_{n,t} \end{bmatrix}^T$$

Equation 12

Where:

n: number of records.

VectorETS can be calculated from Equations 13 to 16:

$$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1+\varepsilon_t)$$

Equation 13

$$\ell_t = (\ell_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$$

Equation 14

$$b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$$

Equation 15

$$s_t = s_{t-m}(1+\gamma\varepsilon_t)$$

Equation 16

Where:

$\alpha = 0.1908$, $\beta = 0.0392$, and $\gamma = 0.0002$.

Holt's model, termed Triple Exponential, employs a simple moving average with equal weighting for past observations. The ETS function is utilized to allocate exponentially declining weights over time [15]. In this scenario, the time series exhibits a trend, leading the forecast to predict the trend for the upcoming period (t+1) as presented in Equation 17.

$$T_t = \beta(F_t - F_{t-1}) + (1-\beta)T_{t-1}$$

Equation 17

Where:

β: a smoothing constant ($0< \beta<1$).

Holt Winter's model is an extension of Holt's method. The forecast for time (*t+1*) is the sum of the trend adjusted by a seasonality index for (*t+1*). The trend relationships mirror those in Holt's model, with the distinction that calculations are grounded in de-seasonalized data [15]. The equation for Holt Winter is commonly formulated as Equation 18.

$$s_t = \gamma^*(y_t - \ell_t) + (1-\gamma^*)s_{t-m}$$

Equation 18

Where:

$\gamma^* = \gamma(1-\alpha)$ with 0< γ<*1*, which translates to 0< γ<*1*- α

3.3.2 Models for intermittent time series

Croston's is a forecasting method specifically valuable for intermittent demand time series [5][22] and [23]. Croston Classic (CC) dissects the dataset into inter-demand intervals and models them using Simple Exponential Smoothing with a predefined parameter. The equation for the CC model is represented as Equation 19.

$$\hat{T}_{t+1} = \hat{T}_t + \beta(T_t - \hat{T}_t)$$

Equation 19

Where:

$T_t = t_n - t_{n-1}$: interval time between t_n *and* t_{n-1}.

t_n : Present time.

t_{n-1} : Previous time.

$\hat{T}_t$: Forecast inter-demand intervals in time *t*.

α : Smoothing parameter, value 0-1.

B : Smoothing parameter for intervals, 0<β<1

The Intermittent Multiple Aggregation Algorithm (iMAPA) is akin to CC but operates with larger time intervals (weekly to monthly or quarterly). It employs a traditional time series forecasting method to predict the aggregated data, enhancing the accuracy of the forecasting process [23]. Consequently, iMAPA can be computed through a three-step procedure, which involves aggregating time series using non-overlapping means of length k. The temporally aggregated time series is denoted with a superscript [k]. Given a time series with Tt and t = 1..., n, the formula for iMAPA can be expressed as Equation 20.

$$T_i^{[k]} = k^{-1} \sum_{t=1+(i-1)k}^{ik} T_t$$

Equation 20

In addition to Croston's method, the Theta forecast method is also widely recognized. It involves modifying the local curvature through a coefficient "Theta" (θ) and includes variations such as AutoTheta, Standard Theta (STheta), Optimized Theta (OTM), and Dynamic Optimized Theta (DOTM) models. AutoTheta, a univariate forecasting method, decomposes the original data into two or more lines, referred to as Theta lines, extrapolates them using forecast models, and combines them to yield the final forecasts. [24] presented AutoTheta in a very intuitive and simple formula, as expressed in Equation 21.

$$Z_t(\theta) = \theta y_t + (1-\theta)(\hat{\alpha} + \hat{B}_t^{'})$$

Equation 21

Where:

y_t : The original time series with *t = 1, ..., n*.

$\hat{\alpha}$ and $\hat{B}_t^{'}$: Least square estimators.

To retain the long-observed data in this study, the research focused in the *θ > 1*, which will be optimized. Thus, the decomposition for the OTM [19] is given by Equation 22.

$$Y_t = (1-\frac{1}{\theta})(\hat{\alpha} + \hat{\beta}t) + \frac{1}{\theta} Z_t(\theta)$$

Equation 22

with *θ* > 1 and the forecast for k steps of time *t* are expressed as Equation 23.

$$\hat{Y}_{t+k|t} = (1-\frac{1}{\theta})\left[\hat{\alpha} + \hat{\beta}(t+k)\right] + \frac{1}{\theta}\hat{Z}_{t+k|t}(\theta)$$

Equation 23

Where:

$\hat{Z}_{t+k|t}$ is extrapolated theta lines

The Theta method mentioned earlier employs two Theta lines; however, forecasting the original time series can involve using more Theta lines by optimizing the parameters θ to minimize forecast errors. The modification of AutoTheta is OTM [19], which incorporates unequal weights in the recompositing procedure for the final forecasts. The formula for OTM can be articulated by combining Theta lines, as presented in Equation 24.

$$Y_t = wZ_t(\theta_t) + (1-w)Z_t(\theta_2)$$

Equation 24

Where:

w is calculated from Equation 25.

$$w = \frac{\theta_2 - 1}{\theta_2 - \theta_1}$$

Equation 25

OTM relies on *θ* parameters, and the optimal value of *θ₂* with *θ* > 1 is determined by *w = 1/θ*. The t parameters are updated based on historical observed data. In this context, OTM can be represented by Equation 26.

$$\bar{Y}_t = \frac{1}{t}\left[(t-1)\bar{Y}_{t-1} + Y_t\right]$$

Equation 26

where: *t=1, 2,…,n,* $\alpha \in [0,1]$ is the smoothing parameter with *θ>1* . If *l*$_0$, *α*, and *θ* parameters are calculated by minimizing the sum of squared error, The forecast at time *t* is calculated by the DOTM [12] as presented in Equation 27:

$$(\hat{l}_0, \hat{\alpha}, \hat{\theta}) = \arg_{l_0,\alpha,\theta} \min \sum_{t=i}^{n} (Y_t - \mu_t)^2$$

Equation 27

3.3.3 Models for seasonal time series

In the context of a simple forecast model applied to seasonal time series, our study employed the Simple Moving Average (SMA) model, as well as Naïve models, which encompass Seasonal Naïve and Naïve models. The Naïve model provides a method for transforming a non-stationary time series into a stationary one by computing the differences between observations. This differencing technique stabilizes the dataset, reducing the impact of trend and seasonality. The implementation of the Naïve model can be expressed through Equation 2.

$$\hat{y} = y_{t-1}$$

Equation 28

Where ε_t indicates noise, ε_t indicates the variability, the higher ε_t the more rapidly the values will change. The RDW model for the original series can be determined from Equation 29.

$$y_t = y_{t-1} + \varepsilon_t$$

Equation 29

Vice versa, Naïve and Simple Moving Average (SMA) model can be considered as opposite ideas in level time series [25]. SMA uses the mean for a small value of the time series as expressed in Equation 30.

$$\hat{y}_t = \frac{1}{m} \sum_{j=1}^{m} y_{t-j}$$

Equation 30

Where:

m is observation times.

Similar to the Naïve model, Seasonal Naïve relies on the most recent value. Seasonal Naïve utilizes values from the corresponding period in the previous season [26]. It can be expressed as Equation 31.

$$\hat{y}_t = y_{t-m}$$

Equation 31

Where:

m is seasonal frequency.

3.4 Model Validation

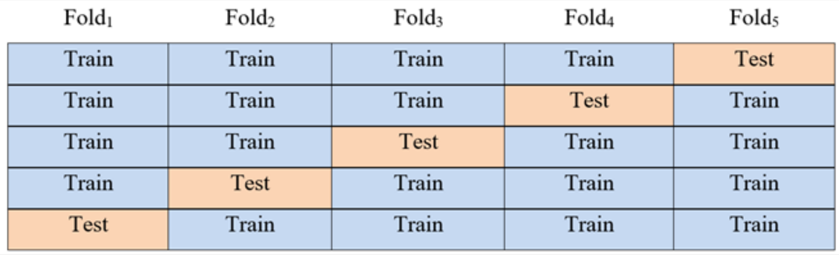
3.4.1 Cross validation

K-Fold Cross-Validation involves utilizing datasets to both test and train AI models [27], with the results indicating their practical accuracy (Figure 4). This method, known for being easy to understand, implement, and providing more reliable estimates than other techniques, is commonly employed for evaluating machine learning models. It maintains time dependence by training, optimizing, and evaluating models across multiple data folds. The crucial parameter in this process is 'k,' which signifies the number of groups into which the data will be split, hence the term "k-fold cross-validation." In this study, 5-fold cross-validation was employed on each parameter at every instance, resulting in the method being specifically referred to as 5-fold cross-validation. During each training iteration, one of the folds was selected as the test data, while the remaining 4 folds served as the training data. This procedure was repeated five times on the entire dataset.

Additionally, a Grid Search was employed to optimize the exploration of each parameter influencing the accuracy of the trained model. Four parameters played a role in determining the accuracy:

- The interpolation method involved two values derived from two interpolation methods: Kriging Ordinary and Universal, resulting in this parameter having two values.
- The variogram model encompassed six models: linear, power, gaussian, spherical, exponential, and hole-effect, giving this parameter six values.
- The number of averaging bins for the semi-variogram consisted of 4, 6, and 8, providing this parameter with three values.
- The weight parameter had two possible values: True or False.

Consequently, the combinations of parameter values resulted in 72 models, each assessed across 360 data folds. The performance evaluation was conducted three times, leading to a total of 1080 training iterations in this study.



$Fold_1$	$Fold_2$	$Fold_3$	$Fold_4$	$Fold_5$
Train	Train	Train	Train	Test
Train	Train	Train	Test	Train
Train	Train	Test	Train	Train
Train	Test	Train	Train	Train
Test	Train	Train	Train	Train

Figure 4: 5-folds cross validation in this study

3.4.2 Performance of the error analysis

For every machine learning model, R-square (R^2), Mean Squared Error (MSE), the Root of the Mean Square Error (RMSE), and the Mean Absolute Error (MAE) were employed as metrics to ascertain the optimal model [28]. These performance indicators are calculated based on Equations 32 to 34.

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(p_k - q_k)^2}$$

Equation 32

$$MAE = \frac{1}{N}\sum_{k=1}^{N}\left|p_k - q_k\right|$$

Equation 33

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Equation 34

Where:

p is experimental value.
q is forecasted value calculated from the model.
N is total number of samples in the database.

In the realm of model validation performance, the RMSE and MAE indicators achieve optimal values when approaching 0, while the R^2 indicator attains its optimal value at 1. This implies that the model exhibits excellent predictive capabilities when these indicators reach the specified values. These three indicators are commonly utilized to assess the predictive prowess of a model. A higher R^2 value signifies a stronger model, indicating a better fit to the dataset, with a maximum value of 1. Conversely, smaller MAE, MSE, and RMSE values indicate a stronger model.

4. Research Results and Discussions

4.1 Results of Data Analysis

The development of accurate air quality forecasts using machine learning models relies heavily on a substantial volume of training datasets. Unfortunately, the ground observation data frequently exhibits uneven distribution and may contain gaps attributed to receiver errors, resulting in records with 0 values or negative values (-9999). Rectifying erroneous data is essential to ensure data integrity, thereby enhancing the training efficacy of machine learning models. Following the implementation of three methods recommended in Section 3.2, datasets were generated, and Figure 5 provides a comprehensive overview of the data status before and after undergoing the data analysis processing.

4.2 Result of PM2.5 Forecast Using Machine Learning Models

Utilizing the machine learning models outlined in Section 3.3, the study can produce results for PM2.5 forecasts using input datasets derived from ground air quality stations (Figure 6).

4.3 Validation of the Results

Criteria for assessing accuracy performance involved values of R^2, MSE, RMSE, and MAE. Initially, the study examined the quality of PM2.5 forecasts based on the dataset and its compatible machine learning models. Despite proposing three interpolation methods in section 3.2 to maintain data integrity, a significant proportion of missing data (16.25%) and occurrences of -9999 (8.13%) still contributed to the lower accuracy of the machine learning models. Figure 7 illustrates the notably high MSE and MAE of the Intermittent Spare time series data. The metric values for MAE and MSE from AutoTheta, STheta, OTM, and DST underscored the specific range (1.03 to 3.33 and 51.13 to 52.07, respectively).

It is evident from Figure 7 that the automatic forecast model (AutoARIMA) and exponential smoothing models such as SES, SSO, VectorETS, Holt, and Holt Winter exhibit similar trends. This suggests that utilizing time series observation data is sufficiently effective for PM2.5 forecasting, with automatic forecast and exponential smoothing models demonstrating greater accuracy compared to other models. Notably, AutoARIMA yields approximately 5% higher results than other models when tested on the same monitoring station.

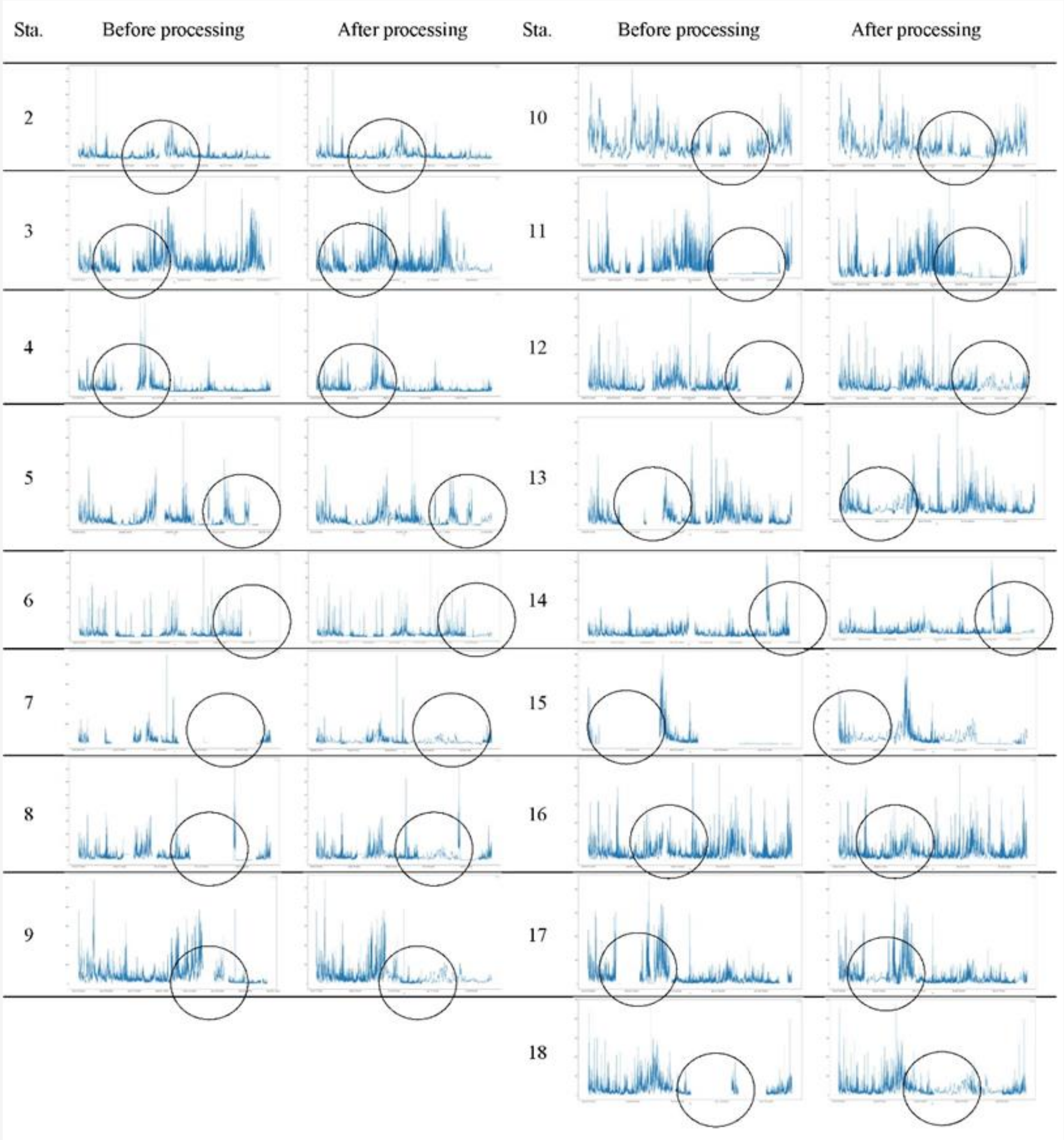


Figure 5: The condition of the data before and after undergoing data processing

In Figure 7, the automatic forecast model, encompassing AutoARIMA and AutoTheta, demonstrates superiority over other AI models. The trend indicates that leveraging multi-temporal data obtained from long-term continuous observation of PM2.5 ground observation data creates a favorable condition for running this model, as evidenced by MAE, MSE, RMSE, and R^2 estimations. AutoARIMA exhibits the lowest MSE at 23.24, MAE at 0.6, and RMSE at 4.7. The smallest RMSE suggests the narrowest difference between forecasted and observed values, signifying the highest accuracy of the PM2.5 forecast model compared to others. Interestingly, AutoARIMA and ETS exhibit a similar trend in MAE (23.2) but with lower RMSE and MSE values (4.7 to 7.06 and 23.24 to 32.56, respectively). Furthermore, the study implemented 5-fold cross-validation (Figure 8) to generate scatter plots illustrating the best fit between two sets of PM2.5 forecasted and actual values, providing a visual representation of their relationship. The scatterplots were generated for forecasted results at intervals of 2 hours, 4 hours, 8 hours, 16 hours, 24 hours, and 48 hours.

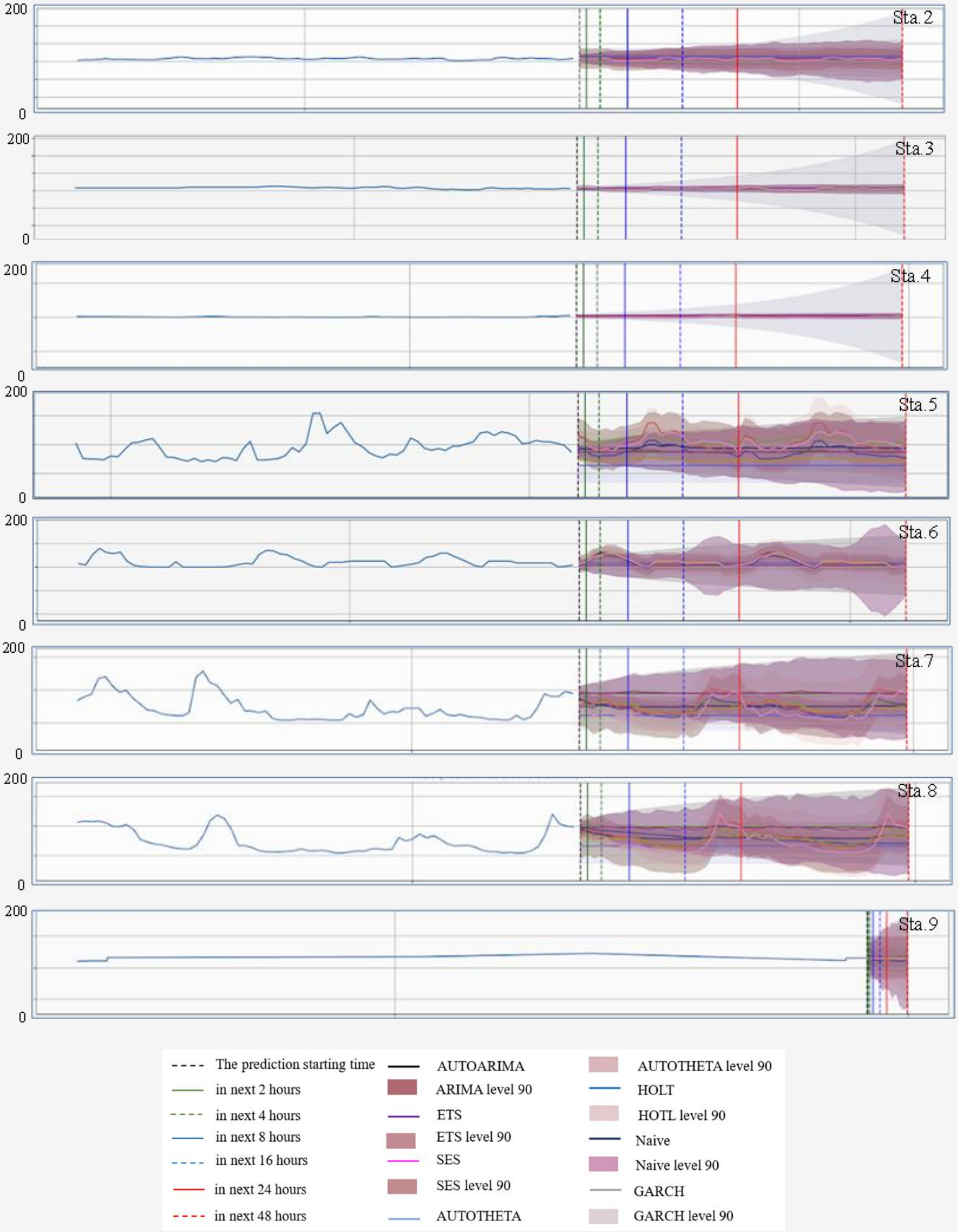


Figure 6: Stimulations of the $PM_{2.5}$ by using machine learning models (Unit of microgram/m^3) (continue next page)

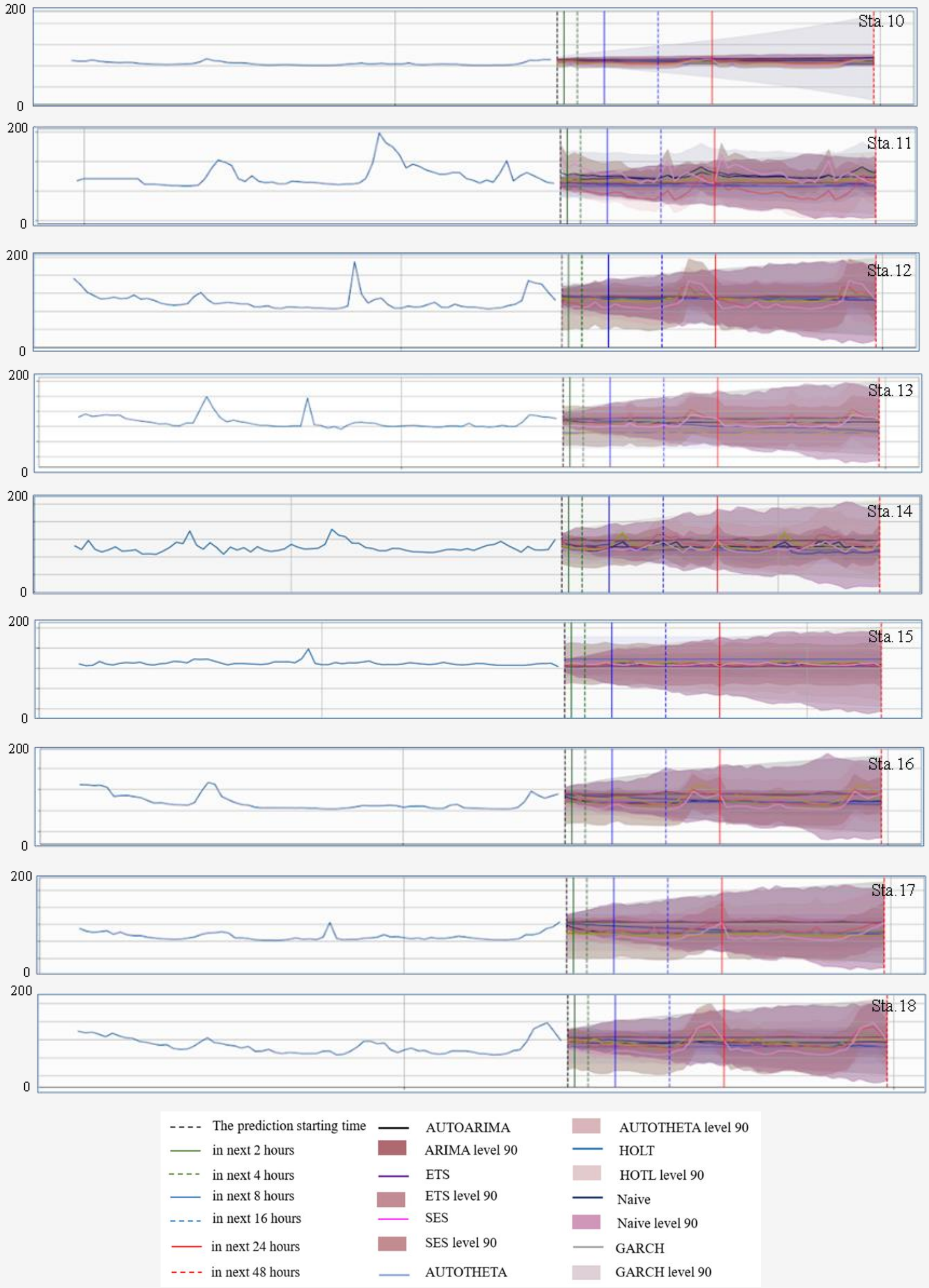


Figure 6: Stimulations of the PM$_{2.5}$ by using machine learning models (Unit of microgram/m^3) (continue from previous page)

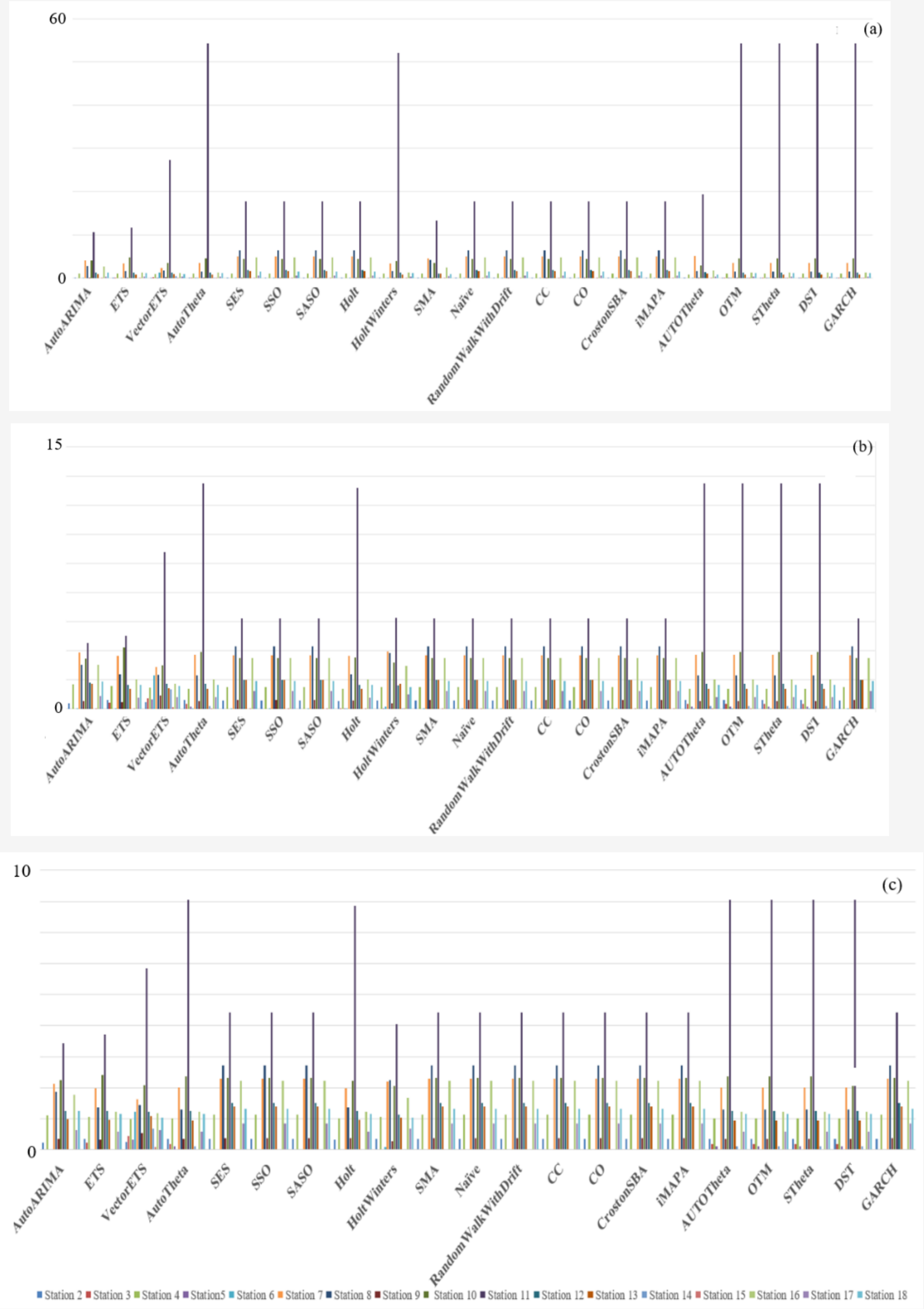


Figure 7: (a) MSE, (b) MAE, (c) RMSE, of forcasted PM2.5

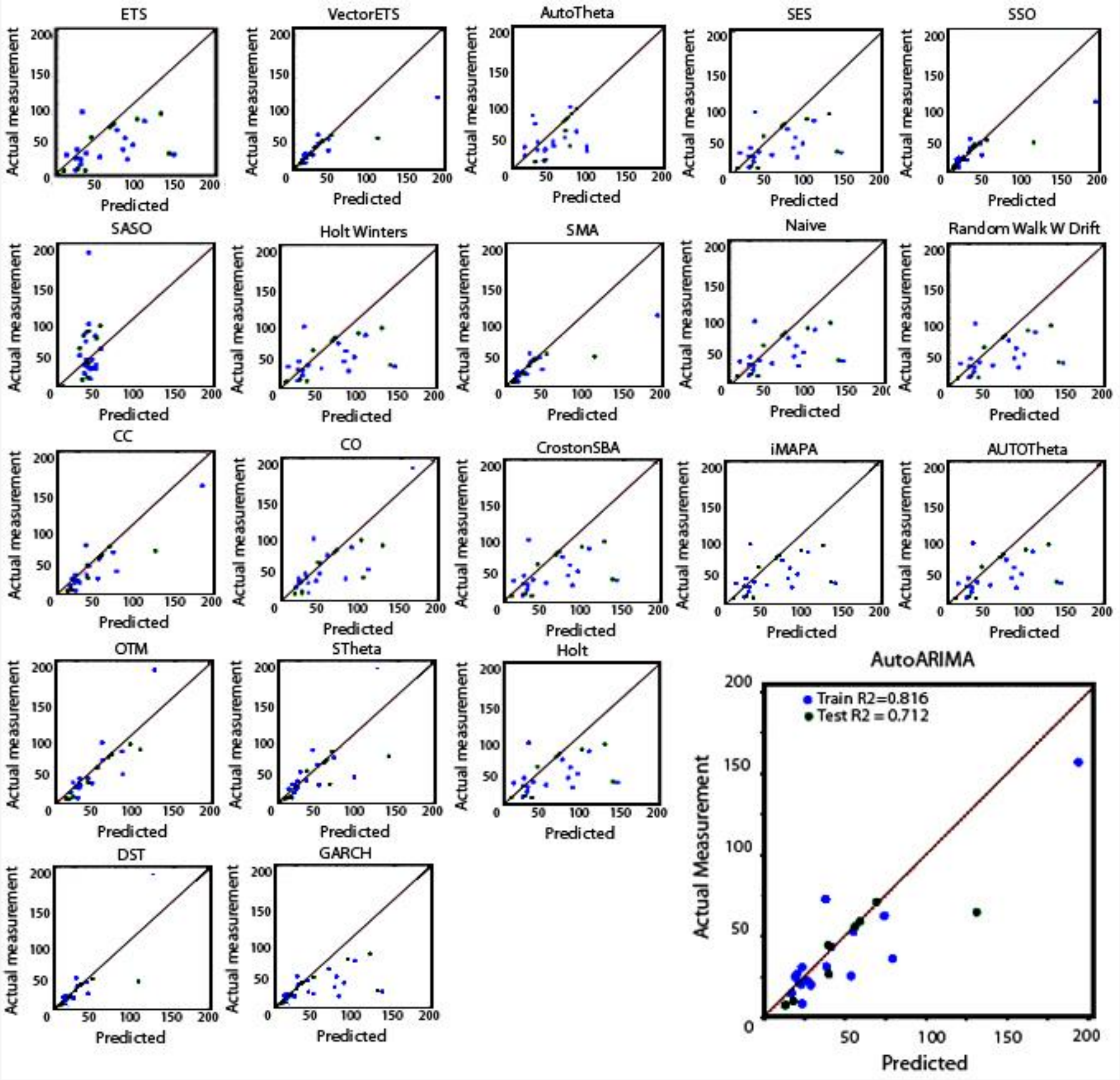


Figure 8: PM2.5 forecasted and actual values established through 5-fold cross-validation

Notably, AutoARIMA demonstrated the highest accuracy among the models, achieving an R^2 value of 0.81. As a result, the performance validation indicates that AutoARIMA stands out as a suitable predictive model when compared to others. Additionally, by incorporating weight updates through the cross-validation method, this model aligns closely with the dataset used in the study. Consequently, AutoARIMA was chosen as the best-fit model for the dataset outlined in Section 3.2.

4.4 Mapping of Predicted PM2.5 Dust Concentration

The mapping of future PM2.5 dust concentration in the study area employs well-established machine learning models to unveil insights into the PM2.5 forecast in BacNinh province, Vietnam. Notably, in this study, AutoARIMA demonstrated its potential for forecasting future dust concentrations after analyzing the previous time series dataset derived from ground air quality stations. The performance analysis allows us to discern the functionalities of the sub-models p, d, and q. Figure 9 illustrates the PM2.5 forecast at intervals of 2 hours, 4 hours, 8 hours, 16 hours, 24 hours, and 48 hours.

4.5 Discussions

Based on a literature review on air pollution in Bac Ninh province, PM2.5 dust emissions are primarily linked to anthropogenic activities such as transportation and industrial emissions (Figure 1). This study has created PM2.5 forecast maps for time intervals of 2 hours, 4 hours, 8 hours, 16 hours, 24 hours, and 48 hours using machine learning models (Figure 9).

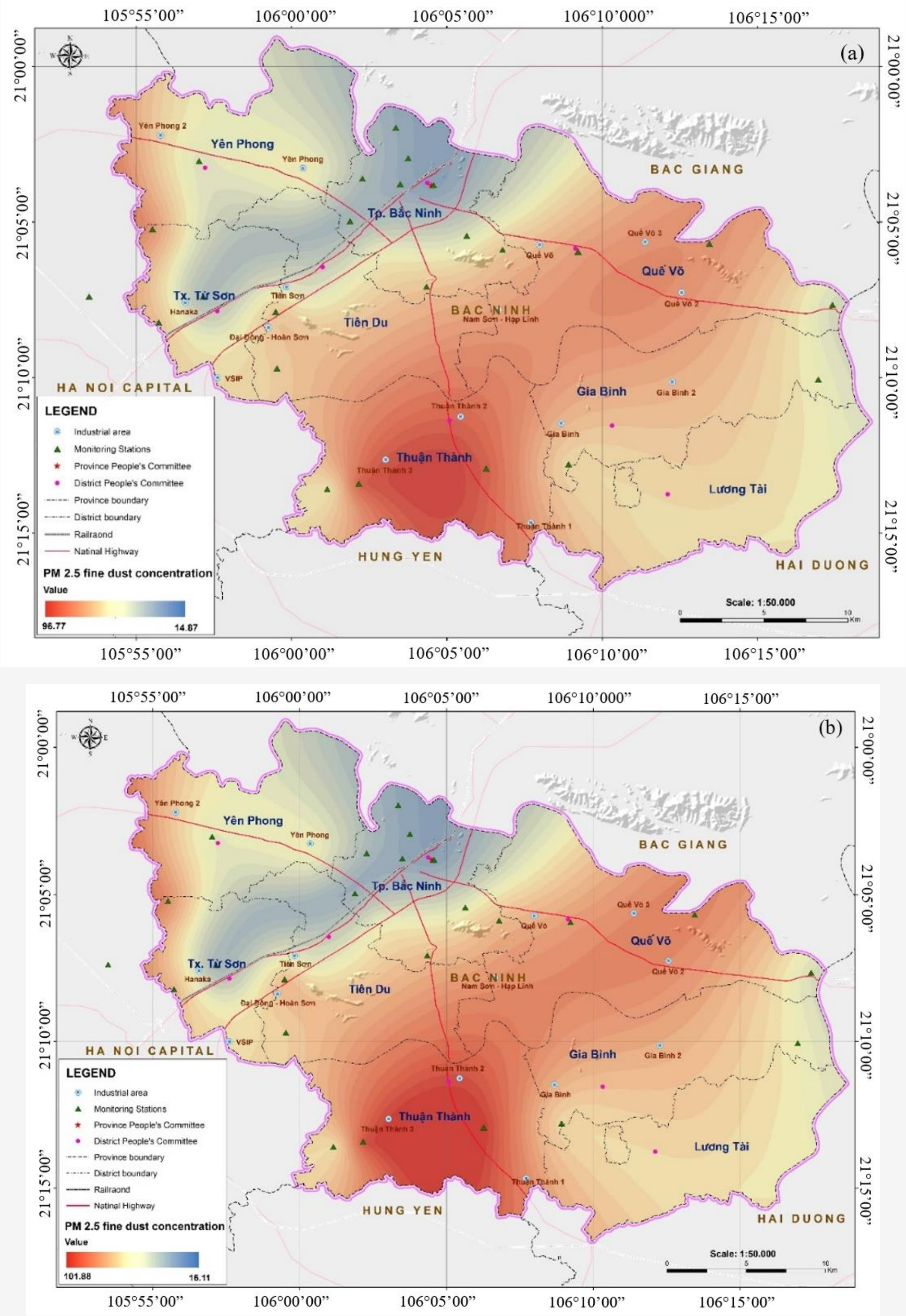


Figure 9: The PM2.5 prediction results generated by the AutoARIMA model at 11:00 AM on September 23, 2022 (a) 0 hour, (b)2 hours, (continue next page)

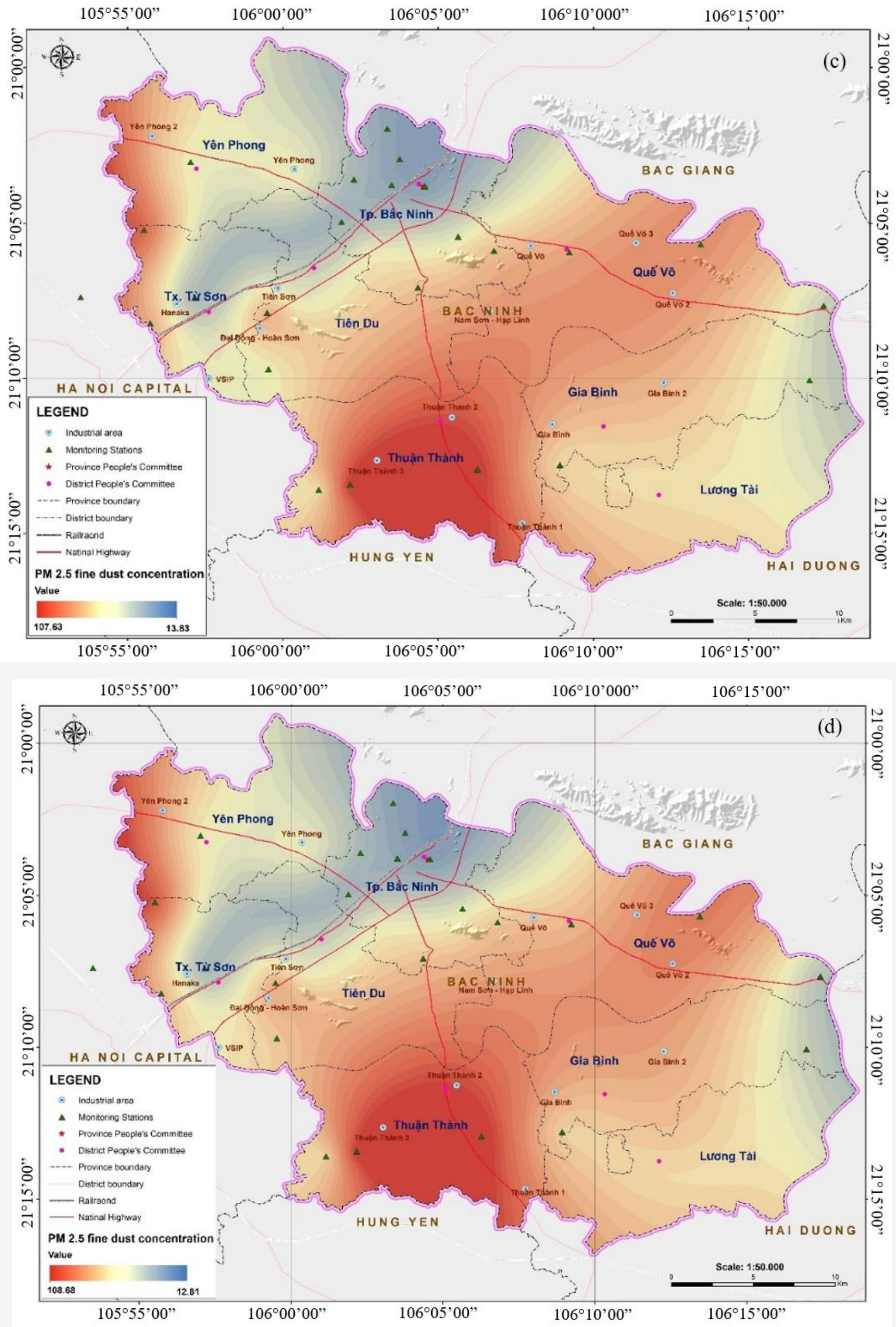


Figure 9: The PM2.5 prediction results generated by the AutoARIMA model at 11:00 AM on September 23, 2022 (c) 4 hours, (d) 8 hours, (continue next page)

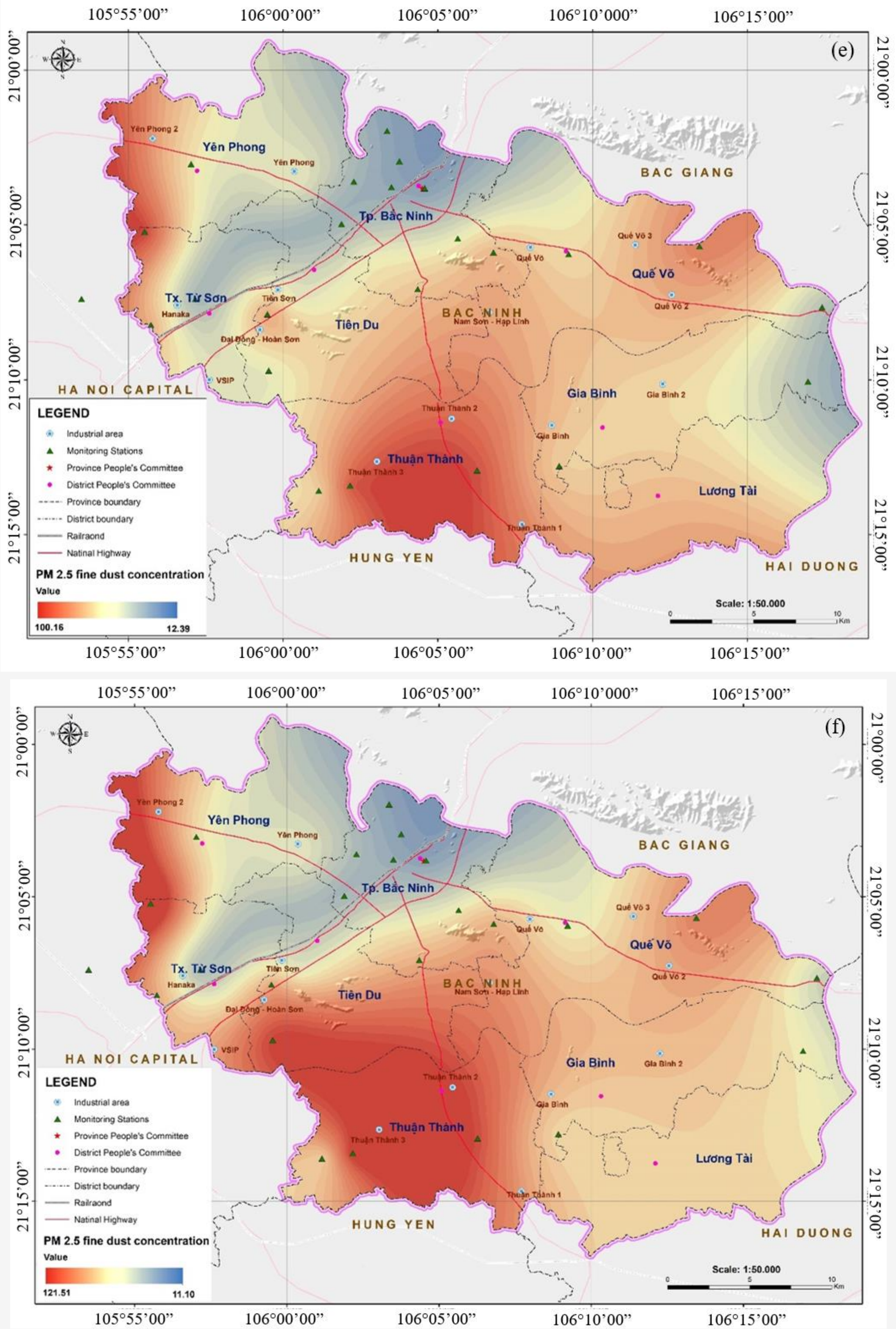


Figure 9: The PM2.5 prediction results generated by the AutoARIMA model at 11:00 AM on September 23, 2022 (e) 16 hours, (f) 24 hours, (continue next page)

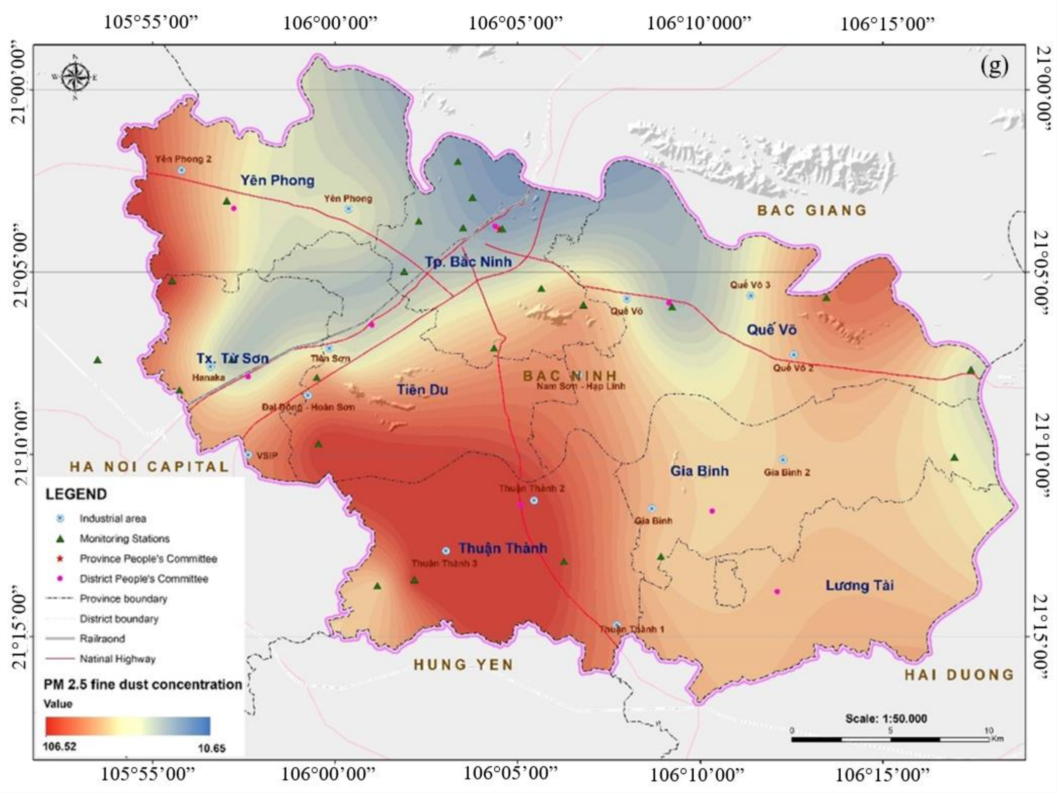


Figure 9: The PM2.5 prediction results generated by the AutoARIMA model at 11:00 AM on September 23, 2022 (g) after the 48-hours, (continue from previous page)

Despite the average level of PM2.5 dust emissions during this period, the distribution is uneven, leading to localized air pollution in neighboring areas. Notably, the annual PM2.5 dust emissions from industrial areas significantly exceed the province's average PM2.5 dust emissions (90 micrograms/m³ compared to 50 micrograms/m³, respectively).

Additionally, the average daily PM2.5 dust emissions recorded at all ground air quality stations surpass the allowable standards. Notably, a substantial area exhibits a PM2.5 concentration exceeding 80 micrograms/m³, extending into residential areas along major traffic routes within the study area. Consequently, the PM2.5 concentration along these traffic routes significantly contributes to elevated PM2.5 levels, including dust carried by vehicles, impacting the air quality of the surrounding region. Another significant source of PM2.5 dust emissions in Bac Ninh Province is attributed to industrial zones. Currently, Bac Ninh hosts 15 industrial zones distributed on both the west and east sides of the province. Figure 9 reveals that the average PM2.5 dust value in these areas exceeds 90 micrograms/m³, reaching a level deemed harmful to human health according to Decision No. 1459, which promulgates Technical Instructions for calculating and announcing the Vietnam Air Quality Index (VN_AQI) as released by the Ministry of Natural Resources and Environment (MONRE) in 2019. Figure 1 provides a visual representation of the impact of industrial zones on air pollution in the province. Thuan Thanh, situated in the southern district of Bac Ninh Province, hosts three concentrated industrial zones namely Thuan Thanh 1, Thuan Thanh 2, and Thuan Thanh 3which displaying the highest pollution levels, reaching approximately 110 micrograms/m³ (Figure 9).

In its strategies and socio-economic development plans, the Bac Ninh government has been actively implementing various environmental protection measures, exemplified by policies aimed at reducing air pollution from dust emissions. An illustration of this commitment is found in Decision No. 222/QĐ-UBND, outlining the Project for Environmental Protection of Bac Ninh province. The government has set increasingly stringent standards for vehicle inspections, advocated solutions for transitioning to renewable energy, promoted the reduction of coal-fired thermal electric power, and mandated the installation of air treatment systems in industrial areas.

Additionally, the government is proactively addressing PM2.5 dust emissions from anthropogenic activities, emphasizing the use of renewable energy and upgrading the transportation system with a focus on public convenience. Bac Ninh, now one of the highest GDP provinces in Vietnam, has recently invested significantly in science and technology, particularly evident in the installation of additional air environment monitoring stations throughout the province. This initiative plays a crucial role in monitoring and mitigating the impacts of PM2.5 dust pollution on the community's health. As living standards improve and environmental awareness grows among the local population, people recognize the significant harm caused by air pollution, which is also identified as a contributing factor to waste generation, either directly or indirectly. Therefore, enhancing public awareness about air pollution is deemed a prerequisite to ensure a reduction in PM2.5 dust emissions in Bac Ninh in the near future.

5. Conclusions

This study developed an advanced machine learning (ML) approach to forecast PM2.5 air pollution within the Bac Ninh Province of Vietnam. Emphasizing the importance of cleaning input data before employing ML models, the study proposed three interpolation methods to preserve ground monitoring data. The selection of suitable machine learning models was based on dataset characteristics, including: 1) Long-term time series; 2) Scattered data due to uneven distribution of ground observation stations; 3) Seasonal variation; 4) Intermittent volatility; and 5) Univariate time series. ML models recommended for time series data included Level time series, Intermittent time series, and Seasonal time series models. Among these, the AutoARIMA model demonstrated the best performance, exhibiting superior accuracy compared to other models (R^2 = 0.81, MSE = 3.5, MAE = 23.24, RMSE = 0.17). AutoARIMA achieved its highest accuracy of R^2 by optimizing its three sub-models *p*, *d*, and *q*. The results revealed a PM2.5 dust concentration in Bac Ninh Province ranging from 40 micrograms/m³ to nearly 100 micrograms/m³, consistent with values obtained at the corresponding ground stations during the accuracy assessment. Thuan Thanh, the southern district of Bac Ninh Province, exhibited the highest pollution level with a dust concentration value of approximately 110 micrograms/m³. Consequently, PM2.5 forecast maps were established for intervals of 2 hours, 4 hours, 8 hours, 16 hours, 24 hours, and 48 hours. These maps vividly depict the impact of industrial zones on PM2.5 air pollution in the southern district, particularly in the concentrated industrial zones of Thuan Thanh 1, Thuan Thanh 2, and Thuan Thanh 3.

This study employed advanced machine learning (ML) models to forecast PM2.5 air pollution in the Bac Ninh Province of Vietnam. Emphasizing the significance of cleaning input data before employing machine learning models, the study proposed three interpolation methods to preserve ground monitoring data. The selection of suitable machine learning models was based on dataset characteristics, encompassing long-term time series, scattered data due to the uneven distribution of ground observation stations, seasonal variation, intermittent volatility, and univariate time series. The ML models recommended for time series data included Level time series, Intermittent time series, and Seasonal time series models. Among these, the AutoARIMA model exhibited the best performance, achieving superior accuracy compared to other models (R^2 = 0.81, MSE = 3.5, MAE = 23.24, RMSE = 0.17). AutoARIMA attained its highest accuracy of R^2 by optimizing its three sub-models p, d, and q (Section 3.3.1). The results indicated that PM2.5 dust concentration in Bac Ninh Province ranged from 40 micrograms/m³ to nearly 100 micrograms/m³. This consistency was evident in the accuracy assessment process with values obtained concurrently at the corresponding ground stations. Thuan Thanh, the southern district of Bac Ninh Province, exhibited the highest pollution level, with a dust concentration value of approximately 110 micrograms/m³. Consequently, PM2.5 forecast maps were established for intervals of 2 hours, 4 hours, 8 hours, 16 hours, 24 hours, and 48 hours, highlighting the impact of industrial zones on PM2.5 air pollution in the southern district, particularly in the concentrated industrial zones of Thuan Thanh 1, Thuan Thanh 2, and Thuan Thanh 3.

In conclusion, the widespread applications of machine learning have generated significant interest, particularly in the realm of forecasting air pollution. This paper focuses on exploring innovative machine learning models for predicting future air pollution in Bac Ninh province, Vietnam. The study underscores the critical importance of pairing a suitable machine learning model with the dataset, as it directly influences result accuracy and reduces computational time. Notably, the limitations of the ground air quality stations in the study area, where only 17 out of 20 stations are operational, impacted the accuracy of research results, particularly in interpolating PM2.5 prediction maps. Future research will consider the implementation of Internet of Things (IoT) technology to enhance PM2.5 monitoring capabilities.

Acknowledgement
This study is supported by Vietnam Ministry of Natural Resources and Environment under the grant number TNMT.2022.04.06.

References

[1] World Health Organization, (2018). Air pollution. Available: https://www.who.int/en/newsroom/factsheets/detail/ambient-(outdoor)-air-quality-and-health. [Accessed September 7, 2023].

[2] Pratyush, M., Dawn, C., Chisato, F. and Mohammad, P., (2022). PM2.5 Air Pollution Forecast through Deep Learning Using Multisource Meteorological, Wildfire, and Heat Data. *Atmosphere*, Vol. 13(5), 1-20, https://doi.org/10.3390/atmos13050822.

[3] Jamali, A., (2020). Sentinel-1 Image Classification Using Machine Learning Algorithms Based on the Support Vector Machine and Random Forest. *International Journal of Geoinformatics*, Vol. 16(2), 15–22. https://journals.sfu.ca/ijg/index.php/journal/article/view/1809.

[4] Li. A. and Xu, X., (2018). A New PM2.5 (2009) Air Pollution Forecasting Model Based on Data Mining and BP Neural Network Model. *Proceedings of the 2018 3rd International Conference on Communications, Information Management and Network Security (CIMNS 2018)*, 110-113. https://doi.org/10.2991/cimns-18.2018.25.

[5] Nicolas, V., (2019). *Forecasting Intermittent Demand with the Croston Model*. Towards Data Science, https://towardsdatascience.com/croston-forecast-model-for-intermittent-demand-360287a17f5f.

[6] Shengjun, W. Q., Feng, Q. and Dong, L., (2011). Artificial Neural Network Models for Daily PM 10 Air Pollution Index Forecast in the Urban Area of Wuhan, China. *Environmental Engineering Science*, Vol 28(5), 357-363. https://doi.org/10.1089/ees.2010.0219.

[7] Leong, W. C., Kelani, R. O. and Ahmad, Z., (2020). Prediction of Air Pollution Index (API) Using Support Vector Machine (SVM). *Journal of Environmental Chemical Engineering,* Vol 8(3). https://doi.org/10.1016/j.jece.2019.103208.

[8] Jumaah, H., Mansor, S. and Pradhan, B., (2018). UAV-based PM2.5 Monitoring for Small-Scale Urban Areas. *International Journal of Geoinformatics*, Vol. 14(4). 61-69. https://journals.sfu.ca/ijg/index.php/journal/article/view/1234.

[9] Joharestani, M. Z., Cao, C., Ni X. and Talebiesfandarani, S., (2019). PM2.5 Forecast Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere*, Vol. 10(7), https://doi.org/10.3390/atmos10070373.

[10] Sachit, M., Ling, C. and Tsai, T. C., (2018). Short-Term PM2.5 Forecasting Using Exponential Smoothing Method: A Comparative Analysis. *Sensors*, Vol. 18(10), https://doi.org/10.3390/s18103223.

[11] Iqra, S., Muhammad, A. A. and Víctor, L., (2023). Machine Learning and Automatic ARIMA/Prophet Models-Based Forecasting of COVID-19: Methodology, Evaluation, and Case Study in SAARC Countries. *Stoch Environ Res Risk Assess*, Vol 37(1), 345–359. https://doi.org/10.1007/s00477-022-02307-x.

[12] Bermudez, J. D., Segura, J. V. and Velcher, E., (2006). Improving Demand Forecasting Accuracy Using Nonlinear Programming Software. *Journal of the Operational Research Society*, Vol 57, 94-100. https://doi.org/10.1057/palgrave.jors.2601941.

[13] Altexsoft, (2021). Preparing Your Dataset for Machine Learning: 10 Basic Techniques That Make Your Data Better. Available: https://www.altexsoft.com/blog/datascience/preparing-your-dataset-for-machine-learning-8-basic-techniques-that-make-your-data-better/. [Accessed Sept. 1, 2023].

[14] Bac Ninh Portal, (2023). Website: https://bacninh.gov.vn. [Accessed Sep. 10th 2023].

[15] Rob, J. H. and George, A., (2023). *Forecasting: Principles and Practice*. (2nd ed.) OTexts. 1-384. https://otexts.org/fpp2/.

[16] Breiman, L., (2001). Random Forests. *Machine Learning*. Vol. 45, 5-32. http://dx.doi.org/10.1023/A:1010950718922.

[17] Dimitris, E., (2020). Time Series Analysis with Theory, Plots, and Code Part 1. Available: https://towardsdatascience.com/time-series-analysis-with-theory-plots-and-code-part-1-dd3ea417d8c4. [Accessed Sept. 1, 2023].

[18] Bollerslev, T., (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, Vol. 31, 307-327.

[19] José, A. F., Tiago, R. P. and Francisco, L., (2015). Models for Optimising the Theta Method and their Relationship to State Space Models. *International Journal of Forecasting*, Vol. 32(4), 1151-1161. https://doi.org/10.1016/j.ijforecast.2016.02.005.

[20] Pablo, R., (2019). ML Approaches for Time Series. https://towardsdatascience.com/ml-approaches-for-time-series-4d44722e48fe

[21] Aishwarya, S., (2023). Multivariate Time Series Analysis for Forecasting & Modeling. Available: https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/. [Accessed Sept. 1, 2023].

[22] Charoenpanyanet, A. and Hemwan, P., (2019). Suitable Model for Estimation of PM2.5 Concentration Using Aerosol Optical Thickness (AOT) and Ground based Station: Under the Dome in Upper Northern, Thailand. *International Journal of Geoinformatics*, Vol. 15(3), 33–43.

[23] John, E. B. and Aris, A. S., (2021). Intermittent Demand Forecasting Companion Site. Wiley. 1-400.

[24] Assimakopoulos, V., and Nikolopoulos, K., (2000). The Theta Model: A Decomposition Approach to Forecasting. *International Journal of Forecasting*, Vol 16 (4), 521-530. https://doi.org/10.1016/S0169-2070(00)00066-2.

[25] Svetunkov, I. and Petropoulos, F. (2018). Old Dog, New Tricks: A Modelling View of Simple Moving Averages. *International Journal of Production Research*, Vol 56, 6034–6047. https://doi.org/10.1080/00207543.2017.1380326.

[26] Ivan, S. and Kourentzes, N., (2015). *Complex Exponential Smoothing*. Research Article, Wiley. http://dx.doi.org/10.13140/RG.2.1.3757.2562.

[27] Rob, J. H. and George, A., (2023). Forecasting: Principles and Practice.

[28] George, S., (2019). Understanding the 3 most Common Loss Functions for Machine Learning Regression. Towards Data Science. https://towardsdatascience.com/understanding-the-3-most-common-loss-functions-for-machine-learning-regression-23e0ef3e14d3.