Time Series-Based Predictive Modeling of PM2.5 Levels in Chiang Mai, Thailand

Tewa Promnuchanont[1], Theeraphop Saengsri[1*], Rujipan Kosarat[2], Piyaphol Yuenyongsathaworn[2]
[1]Department of Computer Information System, Faculty of Business Administration and Libera Arts,Rajamangala University of Technology Lanna, Chiang Mai, Thailand
[2]Department of Software Engineering, Faculty of Engineering, Rajamangala University of TechnologyLanna, Chiang Mai, Thailand
128 Huay Kaew Road, Muang, Chiang Mai, Thailand, 50300
[*]Corresponding Author: tees@rmutl.ac.th. Phone Number: +66-869202820
Received: 3 December2024, Revised: 7 May 2025, Accepted: 23 May 2025

Abstract
The purpose of this study is to use a variety of models to create prediction models for Chiang Mai's PM2.5 levels. To improve the accuracy of our predictions, we take into account outside variables that might influence PM2.5 levels. Among the variables that we include in the data are PM2.5 concentrations, temperature, wind speed, precipitation, cloud cover, relative humidity, and other external factors. Before using the model, the researcher used basic statistical analysis, seasonal analysis, and stationary analysis to assess the data. The team of researchers carried out both data transformation and data cleansing. We tested the ARIMA, SARIMA, and SARIMAX forecasting models. First, we use ARIMA to forecast and assess results. The SARIMA model more accurately captured the seasonal connection in the data when we included a seasonal component. The model was able to forecast PM2.5 levels more precisely at times when seasonal patterns recurred thanks to this improvement. As the last step, we used the SARIMAX model to improve performance by adding exogenous variables. In the end, we assessed the accuracy and performance of each forecast using the MAE and RMSE numbers. The ARIMA model yielded MAE values of 7.34 and RMSE 7.95. The SARIMA model MAE values of 5.76 and RMSE 6.54. The SARIMAX model, when incorporating humidity, had the lowest MAE values of 4.36 and RMSE 5.25, representing improvements MAE of 40.6% and RMSE 34% compared to ARIMA.
Keywords: PM2.5, ARIMA, SARIMA, SARIMAX, Forecast Model

1. Introduction

Dust, a common environmental issue, has detrimental effects on both human health and the ecosystem. Dust is a term that refers to the minute particles that are floating in the atmosphere. These particles may vary greatly in size and composition, with some even being tiny enough to enter an individual's respiratory system. There are a number of major health hazards associated with small particles, including cardiovascular and pulmonary issues. We often divide small particles into categories of PM10 and PM2.5 categories. Due to its interactions with environmental systems, dust not only impacts health, but also influences climate change; especially when it contains harmful greenhouse gases. Finding solutions that improve the environment and human health requires an understanding of the sources, effects, and necessary mitigation techniques of dust pollution. PM2.5 may originate from a number of sources, such as vehicle emissions, wildfires, and industrial operations. By reducing air quality, PM2.5 worsens environmental problems, including poor sight and climate change, in addition to its negative health impacts. Knowledge of PM2.5 sources, effects, and mitigation strategies is essential for improving human health and protecting the environment. The forecast of PM2.5 levels is essential for maintaining air quality and safeguarding human health. Serious health risks may arise from particulate matter with a diameter of 2.5 micrometers or smaller, or PM2.5, which can enter the bloodstream and penetrate deeply into the lungs. Accurate PM2.5 level estimates may help individuals and communities prevent exposure, take effective action, and anticipate and solve air quality issues.

In today's world of rapid change, data mining has grown in importance due to the exponential growth in data collecting. Data mining uses sophisticated techniques rather than just extracting data to find valuable information hidden in massive datasets. In order to develop sophisticated artificial intelligence models, find patterns, and facilitate informed decision-making, it involves using deep technologies and complex algorithms. Businesses, academic institutions, and other organizations of various kinds may utilize this ability to identify important trends across a range of industries. Many different areas, including forecasting commercial trade, targeted marketing plan development, and medical research, employ data mining.

The use of data mining techniques has made it possible to develop new methods for displaying and analyzing data. Regression analysis accomplishes the separation of dependent factors and independent variables [1], classification models make educated guesses about the categories that individuals belong to, anomaly detection [2] pinpoints problems, association rule mining [3] illustrates the connectivity of data, and text mining [4] extracts meaning from text data. Time series analysis (ARIMA) [5] examines and forecasts data gathered over time. Hierarchical structures are used by decision trees [6] to assist individuals in making decisions. Ensemble approaches use multiple models to make predictions more accurate. And neural networks [7] try to work like the brain. The optimal data mining technique depends on the specific goals and types of data.

This project aims to develop prediction models for PM2.5 levels in the Thai region of Chiang Mai, which has annual PM2.5 dust pollution. We have evaluated the forecast model using three models: SARIMA, ARIMA, and SARIMAX. ARIMA is used for modeling time-dependent data that exhibit trends and autocorrelation. SARIMA is built upon ARIMA by including seasonal components, which are essential for capturing the yearly patterns often observed in environmental data such as PM2.5. Forecasting is further enhanced by SARIMAX through the incorporation of external variables, such as weather conditions, which are known to influence PM2.5 levels. These models are widely applied to support early warning systems and inform environmental policies aimed at reducing public health risks.

In the test, the test component used data from January 2023 to December 2023 for 12 months, while a 60-month data from January 2018 to December 2022 was used to predict the PM2.5 levels for the whole year of 2023. The structure of this work describes the setup as follows: Section 2 provides a review of various earlier studies directly related to this subject, while Section 3 discusses the technique. Section 5 concludes this article with a presentation and discussion of the results from Section 4.

2. Literature review

According to research, PM2.5 and PM10 are two different kinds of dust, each having its own origins and chemical makeup [8]. The particles that make up particle pollution are separated into two categories, according to Chen et al [9], PM10 contains particles as tiny as 10 microns, whereas PM2.5 contains particles as small as 2.5 microns. PM2.5 contains carbonaceous material and metal compounds that may be harmful to both people and the environment The atmosphere disperses these particles in different ways, according to Singh et al [10], and PM10 often reflects coarse-mode PM, such as dust carried by the wind. The study used the Nested Regional Climate and Chemistry Model (NRCM-Chem) to predict PM2.5 concentrations over the northern peninsula of Southeast Asia during 2020–2029 under the RCP8.5 scenario. The model showed good agreement with observed data, with an Index of Agreement (IOA) between 0.63 and 0.80, although it slightly underestimated temperature and precipitation and overestimated PM2.5 levels [11].

The intersection of big data, machine learning, and data mining is a fast-growing subject with enormous potential for a broad variety of possible applications at every stage of its development. Singh highlights how machine learning may be able to help with the difficulties that come with large data analytics [12]. Yang C, Huang Q, and et al [13] both stress the value of data mining as a

technique for drawing insightful conclusions from large and intricate data sets. Furthermore, they look at the different kinds of big data and the difficulties that come with them. Wang S and Cao J's [14] article delves deeper into the use of data mining methods and the applications of big data processing. The research looks more closely at various applications and focuses on the diverse nature of huge data.

ARIMA has been the subject of several research studies in the fields of data mining and projected outcomes. Díaz-Robles LA, Ortega JC and et al [15] used an upgraded ARIMA model for air quality prediction and achieved better results by using a sliding window technique and forecasted data. The integration of the data made this possible. Lee, Dongwon, and et al [16] achieved a low average error rate, while Khashei M and Bijari M [17] likewise discovered that ARIMA was effective in predicting time series data. Shivhare N, Rahul AK, and et al [18] has developed tools for daily weather forecasting and proposed an ARIMA method for weather data mining that is based on Hadoop. These research findings illustrate the utility of ARIMA in data mining and prediction activities. Other research results demonstrate the usefulness of ARIMA in other fields. The results of this research illustrate the adaptability and use of ARIMA in the field of data mining.

According to earlier research, there were a number of obstacles in PM2.5 forecasting, such as significant environmental variability, difficulties in capturing abrupt pollution spikes, and limitations in data completeness and resolution. To address these challenges, models were required that could manage missing or inconsistent data, incorporate external factors, and handle seasonality. The combination of many forecasting models, such as ARIMA, SARIMA, and SARIMAX, plus the addition of other exogenous factors to increase predicted accuracy set this study apart from earlier investigations. This study examined the relative effectiveness of several models and showed how external meteorological elements might greatly improve forecasting accuracy, while previous studies often relied on a single model or just looked at past PM2.5 data.

3. Methodology

A structured time series forecasting flowchart was the primary program of the proposed algorithm, as shown in Figure 1. Data entry comes first, followed by analysis and cleaning. Before establishing the ARIMA, SARIMA, and SARIMAX models, we made changes to the cleaned data. With this model, we made predictions and we corrected for poor model performance. The technique produced excellent analysis and outcomes. This method generated accurate forecasts by fine-tuning the model in response to performance assessments.

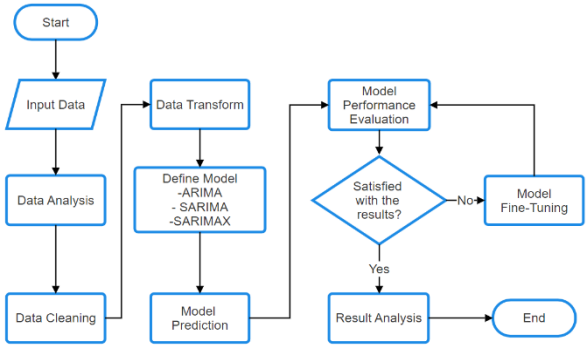

Figure 1 the primary program of the suggested approach.

3.1 Dataset

The comprehensive data from several sources was included in the dataset used to study PM2.5 levels, which improved the analysis's resilience. The PM2.5 information came from Berkeley Earth (https://berkeleyearth.org), which provided detailed records of particulate matter concentrations [19]. Visual Crossing (https://www.visualcrossing.com) provided essential meteorological context by retrieving external data; including temperature, cloud cover, relative humidity, and wind speed. An additional source of environmental information was rainfall data from the Hydrological Information Institute (https://www.hii.or.th) [20]. By combining these datasets, the research was able to more accurately identify the factors influencing PM2.5 levels more accurately, yielding more useful conclusions and recommendations for managing air quality and understanding environmental impacts.

3.2 Data preprocessing

We formatted the obtained data in this stage so that it cloud be used for other forecast. Data preparation is an essential step in the data analysis pipeline that ensured the data was accurate, consistent, and ready for analysis. We converted raw data into a format suitable for analysis, cleaning, and transformation to achieve for more accurate and insightful results. To identify stationary, seasonality, and seasonal decomposition data, we examined the processes involved in the data analysis procedure. After that, the data transformation procedure included converting daily data into monthly data, removing anomalies, and using averages to fill in the missing days.

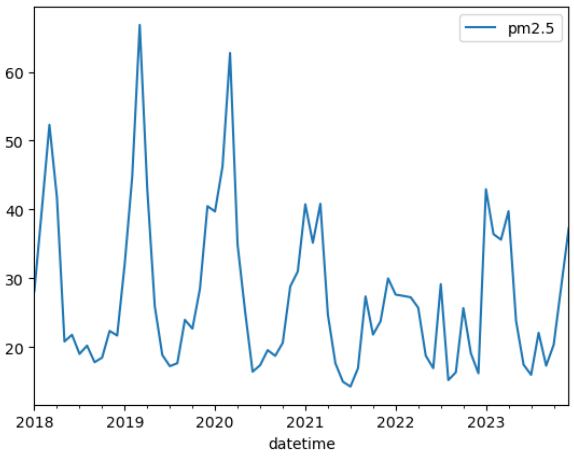

Figure 2 Chiang Mai, Thailand's basic information on PM2.5 levels.

The 2218-day collection of PM2.5 data for Thailand's Chiang Mai province provided crucial information on the air quality in that area. Data transformation involved several steps. Outliers were removed, and missing values were filled using the average of the available data. The daily data were then aggregated and converted into monthly data. For model evaluation, the dataset was divided into two parts. The training set consisted of 60 months of monthly data, from January 2018 to December 2022. The testing set included 12 months of data, from January 2023 to December 2023, and was used to forecast the PM2.5 values for the entire year of 2023. The average PM2.5 level was 27.24 µg/m³, which was considered moderate pollution by the Air Quality Index (AQI). Air quality fluctuated significantly, as seen by the standard deviation of 19.41 µg/m³, which demonstrated substantial variability in daily pollution levels. The variability was further shown by the data, which displayed a broad range of PM2.5 values, from a low of 3.67 µg/m³ to a maximum of 126.44 µg/m³. According to the distribution analysis, the 25th percentile (Q1) was 12.51 µg/m³, meaning that the air was comparatively clean for 25% of the days. PM2.5 levels were at or below this threshold on half of those days, as shown by the median (Q2) of 20.84 µg/m³. The 75th percentile (Q3), which indicated that 25% of those days had greater pollution levels, was 37.24 µg/m³. Even though Chiang Mai occasionally experiences low pollution levels, the city frequently experiences moderate to high PM2.5 levels, with notable daily fluctuations. Figure 2 illustrates how important this data is for planning public health initiatives, and for those who are sensitive to changes in air quality.

A time series analysis's ability to determine whether the data is stationary or if its statistical properties remain constant across time was essential for its success. There were a number of ways to do this. For visual assessment, we plotted the data to see whether the mean and variation were seasonal and stable. It was possible to determine the unit root using statistical methods like the Augmented Dickey-Fuller (ADF) test. Without it, the data remains stagnant. When the autocorrelation function (ACF) and the partial autocorrelation function (PACF) were analyzed, it was possible to determine whether or not the relationships in the data had deteriorate with time.

Time series analysis requires determining whether the data was stable and if its statistical features stayed constant. To find a unit root, one can either display the data and determine whether the mean and variance remain constant, or they can use statistical tests such as the Augmented Dickey-Fuller (ADF) test, as outlined in Eq. (1). Both null hypotheses assume unsteady data. This test should have a p-value of 0.05 or less to reject the null hypothesis.

$$\Delta y_t=\alpha+\beta t+\gamma y_{t-1}+\sum_{i=1}^{p}\delta_1\Delta y_{t-i}+\epsilon_t \quad (1)$$

where Δy_t the time series' initial difference,
y_t, γy_{t-1} the time series' delayed level,
α is an optional constant term,
βt is an optional trend term,
ϵt is the error term,
p is the model's number of lag differences,
$\sum_{i=1}^{p}\delta_1\Delta y_{t-i}+\epsilon_t$ is the Sum of time series lagged differences to adjust for higher-order autocorrelation.

The following were the ADF test results: There were 60 observations, the p-value was 0.766, and we employed 11 lags. These findings implied that we were unable to rule out the ADF test's null hypothesis, which showed that the time series was not stationary. The results of the ADF test suggest that we could validate the unit root or non-stationary nature of this data with a p-value of 0.05.

After that, time series analysis used the autocorrelation (ACF) and partial autocorrelation (PACF) functions to understand the correlations between the data [21]. The ACF method evaluated the connection between the present data value and the future data value (lag) without taking into consideration any additional factors in the data. Through the consideration of intermediate lag correlations, PACF was able to discover independent links between the values of the present data and those of the distant data.

The Autocorrelation Function (ACF) for a time series $\{X_t\}$ at lag k is defined as seen by Eq. (2).

$$\rho_k=\frac{Y_k}{Y_o} \quad (2)$$

where Y_k is $Cov(X_t,X_{t-k})$ for any i.
$Cov(X_t,X_{t-k})$ is the covariance between X_t to X_{t-k}.
Y_o is the variance of the stochastic process.

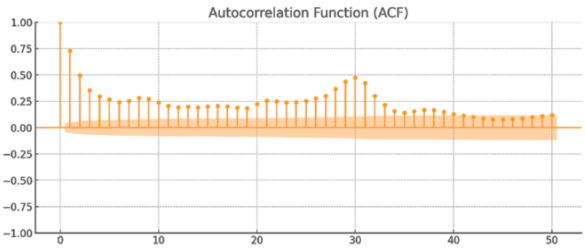

Figure 3 The autocorrelation (ACF).

From Figure 3. The ACF graph shows the relationship between current and lag values. Its periodic decline reveals seasonality. The ACF graph's cyclical rise and fall validates data seasonality by showing a recurrent and consistent correlation across periods.

The Partial Autocorrelation Function (PACF) is a statistical tool that quantifies the degree of correlation between a time series and its delayed values. This function took into consideration the influence of intermediate delays. In the PACF at lag k, the direct link between X_t and X_{t-k}. was calculated after the contributions of intermediate delays had taken into consideration. In the next regression model, Eq. (3) shows how to estimate the PACF at lag k, which is shown by $\emptyset_{kk}$ as the coefficient of X_{t-k}.

$$X_t=\emptyset_{1k}X_{t-1}+\emptyset_{2k}X_{t-2}+\ldots+\emptyset_{(k-1)}X_{t-(k-1)}+\epsilon_t \quad (3)$$

where X_t is the partial autocorrelation coefficient at lag k,
ϵ_t denotes the residual.

For an autoregressive (AR) process, PACF may be computed using recursive connections derived from the Yule-Walker equations. To calculate the PACF at lag k, use Eq. (4).

$$\emptyset_{kk}=Corr(X_t, X_{t-k} \mid X_{t-1},\ldots, X_{t-(k-1)}) \quad (4)$$

The PACF (Partial Autocorrelation Function) graph that is shown in Figure 4 indicates a direct link that exists between the current values and the lagged values. This graph takes into consideration the effect of other lagged data as well. The existence of substantial spots in the PACF graph at lags that correspond to seasonal periods, i.e. around 12 months, is additional evidence that the data exhibited seasonality. This was evidenced by the fact that the PACF graph contains significant spots. The evidence of seasonal patterns was enhanced by these notable spikes, which demonstrate that there was a clear association between the data and values from previous seasonal periods. This gives credence to the notion that seasonal patterns existed.

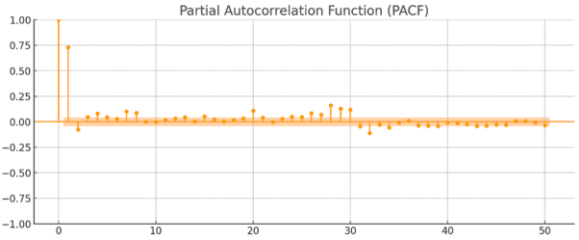

Figure 4 The partial autocorrelation (PACF).

In the last step of data analysis, seasonal decomposition breaks down the data to better understand its patterns and properties. Separating the trend, seasonal, and residual data. This was a typical seasonal breakdown. The trend component showed a long-term data evolution to identify long-term growth or reduction. Seasonality captures trends that occur monthly, quarterly, or annually to indicate short-term changes that follow a cycle. Once we eliminated trends and seasonal effects, the residual component reveals random data fluctuations or deviations that neither a trend nor seasonal factors can explain. Breaking the data into various components helped us understand and analyze the time series.

Figure 5 illustrates the seasonality and non-stationarity of Chiang Mai's PM2.5 data from 2018 to 2023. The ADF test and seasonal decomposition analysis p-values over 0.05 indicated data instability. The breakdown graph shows trends and seasonality components in the data. The monthly PM2.5 data for Chiang Mai shows distinct seasonal patterns and trends.

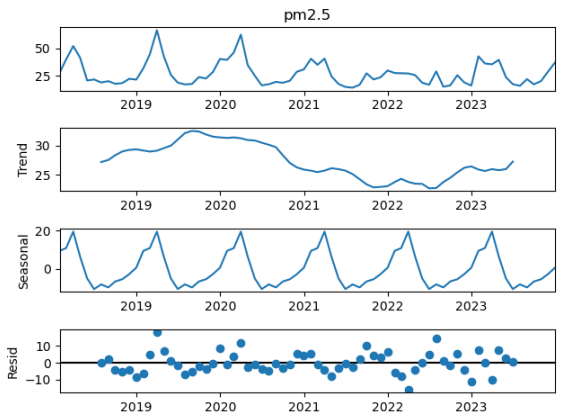

Figure 5 2018–2023 Chiang Mai PM2.5 seasonal component decomposition.

3.3 Model prediction

To assess the study's prediction effectiveness based on the examination of PM 2.5 data in Chiang Mai Province, the researchers used three models: ARIMA, SARIMA, and SARIMAX [22].

A technique known as ARIMA, which is an acronym that stands for autoregressive integrated moving average, was used by us for the aim of assessing and predicting time series data that displays seasonality, patterns, or instability. ARIMA is comprised of three components: MA, I, and AR. We made use of differencing in order to decrease trends and stabilize the data; MA makes use of moving averages of prediction mistakes in order to boost accuracy; and AR makes use of past values in order to anticipate future values. All of these techniques were used in parallel. ARIMA is able to manage time-series data that is both unstable and intricate when both components are mixed inside the model.

It was possible to represent the ARIMA model equation in terms of its constituent parts. Eq. (5) displays the generic form of the ARIMA *(p, d, q)* model for a time series Y_t.

$$y_t = c + \emptyset_1 y_{t-1} + \ldots + \emptyset_p y_{t-p} - \theta_1 \epsilon_{1-1} - \ldots - \theta_q \epsilon_{t-q} + \epsilon_t \quad (5)$$

where y_t is the actual value at time *t*,

c is a constant term,

$\emptyset_i$ are coefficients for autoregressive terms,

θ_i are the coefficients for moving average terms,

ϵ_t is the time-*t* error term.

When $d>0$, replace Y_t with its Δ^d differences $\Delta^d Y_t$ for stationarity, as described in Eq. (6)-(7). ARIMA successfully models and forecasts time series data using AR, differencing, and MA terms.

$$\Delta Y_t = Y_t - Y_{t-1} \quad (6)$$

$$\Delta^d Y_t = \Delta(\Delta^{d-1} Y_t) \quad (7)$$

SARIMA adds seasonal components to the ARIMA model. The SARIMA model is defined as SARIMA $(p, d, q) \times (P, D, Q)_s$, Eq. (8) represents the SARIMA model.

$$y_t=c+\emptyset_1 y_{t-1}+\ldots+\emptyset_p y_{t-p}-\theta_1\epsilon_{1-1}-\ldots-\theta_q\epsilon_{t-q}+\Phi_1 y_{t-s}+\ldots+\Phi_P y_{t-Ps}-\Theta_1\epsilon_{1-s}-\ldots-\Theta_Q\epsilon_{t-Qs}+\epsilon_t \quad (8)$$

where L is the lag operator,

$\emptyset_i$ and Φ_i are Coefficients of the non-seasonal AR terms,

θ_i and Θ_i are Coefficients of the non-seasonal MA terms,

ϵ_t is the white noise,

s is the seasonal period.

SARIMAX adds exogenous regressors to the SARIMA model. SARIMAX incorporates wind speed, rainfall, relative humidity, temperature, and cloud cover. We added one external variable to each test. The SARIMAX model is defined as $(p, d, q) \times (P, D, Q)_s$ with exogenous variables X_{it}. Eq. (9) represents the SARIMAX model.

$$y_t=c+\emptyset_1 y_{t-1}+\ldots+\emptyset_p y_{t-p}-\theta_1\epsilon_{1-1}-\ldots-\theta_q\epsilon_{t-q}+\Phi_1 y_{t-s}+\ldots+\Phi_P y_{t-Ps}-\Theta_1\epsilon_{1-s}-\ldots-\Theta_Q\epsilon_{t-Qs}+\beta_1 X_{1,t}+\ldots+\beta_k X_{k,t}+\epsilon_t \quad (9)$$

where L is the operator of lag,

$\emptyset_i$ and Φ_i are the parameters of AR, θ_i ,

Θ_i are the parameters of MA,

β_k are coefficients of the exogenous variables.

ϵ_t is the forecast error,

X_{it} are the exogenous variables.

3.4 Model performance evaluation

For the purpose of the test, the data was collected over a period of twelve months, commencing in January 2023 and finishing in December 2023. On the other hand, the data for the train component was collected over a period of sixty months, beginning in January 2018 and ending in December 2022. A total of three models, namely ARIA, SARIMA, and SARIMAX, was used by us in order to carry out the procedure. Through the use of two distinct optimization methodologies, we were able to enhance the correctness of the product. First, the SARIMA model's parameters were adjusted using grid search. Grid Search lets you change the model's parameters by giving each parameter a range, and carefully examining every possible combination. This process finds the set of parameters that yields the highest accuracy. Grid search systematically looks at every potential combination of parameters to improve prediction accuracy and optimize model performance.

In an effort to achieve more precision, we included external components into the SARIMAX model. We made observations on the amount of precipitation, relative humidity, temperature, cloud cover, and wind speed. The model incorporates these variables to increase forecast accuracy by adding crucial components that may affect time series data. We used MAE and RMSE to test the SARIMAX model with these external regressors. These metrics, which compute the average magnitude of errors and the square root of the average of squared errors, demonstrate the accuracy of the model. Our prediction model accuracy statistic is Mean Absolute Error (MAE). Mean absolute error (MAE) shows prediction errors' average magnitude regardless of direction. Since it provides the precise difference between predicted and actual values, this statistic helps evaluate a model. The formula calculates the average absolute discrepancies between the predicted and actual values. The MAE formula is in Eq. (10).

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|Y_i-\hat{Y}_i\right| \quad (10)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2} \quad (11)$$

where n is the quantity of occurrences.,

Y_i is the actual value at occurrences i,

$\hat{Y}_i$ is the predicted value at occurrences i

$\left|Y_i-\hat{Y}_i\right|$ reflects each occurrence's absolute error.

The Root Mean Squared Error (RMSE) statistic is often used by predictive models in order to carry out an evaluation with a degree of accuracy that they possess. We were able to calculate the difference between the numbers that were predicted and those that were actually observed by using the square root of the average of the squared deviations. This allows us to determine the difference between the two sets of numbers. RMSE is one of the measures that was used to evaluate how well the predictions made by the model fit the actual data. Due to the fact that this

measure squares differences, it provides more weight to larger errors that are the result of the squared differences approach. As a consequence of this, major deviations and outliers have the ability to have a large impact on the root mean square error (RMSE). Because it presents a degree of mistakes in the same units as the data that was initially gathered, Eq. (11) was a useful indicator for assessing the overall performance of a model. This was because it displays the errors in the same units.

4. Result and discussion

The dataset gathers PM2.5 data for the Thai province of Chiang Mai over a period of 2,218 days, spanning 60 months, from January 2018 to December 2022. The test data was obtained between January 2023 and December 2023. In addition to exterior factors like temperature, cloud cover, relative humidity, wind speed, and rainfall, it includes 52,584 records with PM2.5 values. This large dataset made it easier to accurately measure and predict PM2.5 levels while taking a variety of meteorological factors into consideration.

Table 1 2023 monthly actual data vs. numerous predicting systems.

Year 2023	Actual	ARIMA	Best SARIMA	SARIMAX (temp)
January	43.446	33.618	36.128	36.562
February	36.796	31.159	28.069	28.186
March	34.467	29.069	31.440	31.596
April	37.329	28.022	26.984	25.839
May	24.240	26.419	18.962	17.965
June	17.774	26.636	20.838	20.311
July	17.697	27.769	28.543	28.317
August	20.482	28.506	18.045	17.806
September	17.800	28.844	24.221	23.661
October	20.949	28.918	27.487	27.339
November	29.446	28.579	24.572	24.244
December	37.117	28.187	36.822	35.973

Table 2 2023 monthly actual data vs. numerous predicting systems (con.).

Year 2023	SARIMAX(Humidity)	SARIMAX (Cloud Cover)	SARIMAX (Wind Speed)	SARIMAX(Rainfall)
January	36.040	37.094	35.842	35.995
February	29.131	30.272	27.094	28.516
March	33.864	33.683	30.662	32.459
April	33.207	30.380	27.190	27.465
May	21.957	21.737	19.795	20.996
June	20.701	20.622	22.288	20.620
July	28.667	28.642	30.571	28.790
August	18.818	18.182	17.942	20.428
September	23.140	24.241	25.184	23.613
October	24.270	25.824	28.374	25.515
November	24.876	25.296	24.706	25.331
December	35.649	36.808	36.538	36.938

On the other hand, the PM2.5 levels for each month of 2023 are shown in Tables 1 and 2, and the forecasts for each model reveal the PM2.5 values in Chiang Mai Province. Please refer to both tables displayed down below. Examples of models that are provided in this package include ARIMA, the best SARIMA, and a number of other SARIMAX models. These models consider a variety of parameters, including temperature, humidity, cloud cover, wind speed, and rainfall. One example of a forecast with an actual value of 43.446, was recorded for the month of January, as shown by the data in Table 1. It is worth noting that the year 2023 saw the highest rating for PM2.5. The ARIMA model anticipated 33.618, whereas the best SARIMA model projected 36.128. Additionally, for this month, a number of models generated forecasts that were different from one another. The SARIMAX models, while factoring in a wide range of meteorological factors, generated the following predictions: 36.562 for temperature, 36.040 for humidity, 37.094 for cloud cover, 35.842 for wind speed, and 35.995 for rainfall.

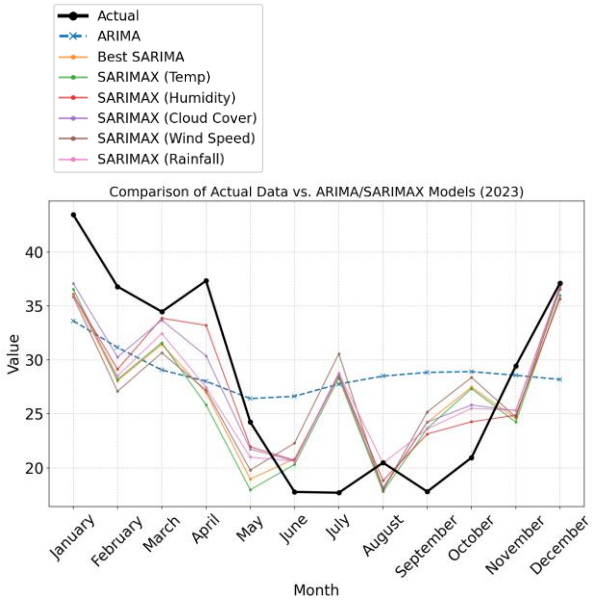

Figure 6 Many predicting methods compared to 2023 monthly data.

The graph in Figure 6 presents a comparison between the actual levels of PM2.5 in Chiang Mai for the year 2023 and the predictions generated by a number of models. These models include

ARIMA, SARIMA, and numerous SARIMAX models, all of which consider a variety of meteorological factors, such as temperature, humidity, cloud cover, wind speed, and rainfall. During the month of January, the actual PM2.5 value reached its maximum point, which was far greater than any model's projection. The SARIMAX model, which took into consideration cloud cover, produced the most accurate estimate of the data. These models have a tendency to reflect the trend of the actual values during the course of the year; nevertheless, they typically overestimate or underestimate the highs and lows that occur throughout the year. The existence of this gap draws attention to the challenges associated with accurately calculating PM2.5 levels and the influence that a wide range of environmental factors has on air quality. Table 3 and Figure 7 both show a comparison of the prediction error to the MAE and RMSE values. We can use metrics such as the RMSE and MAE to compare the forecast error. These metrics provide a numerical depiction of the prediction accuracy.

Table 3 Compare predicted error to MAE and RMSE.

Model	MAE	RMSE
1. ARIMA	7.34	7.95
2. Best SARIMA	5.76	6.54
3. SARIMAX (Cloud Cover)	4.58	5.43
4. SARIMAX (Humidity)	4.36	5.25
5. SARIMAX (Rain)	4.96	6.05
6. SARIMAX (Temperature)	5.85	6.60
7. SARIMAX (Wind Speed)	6.31	7.15

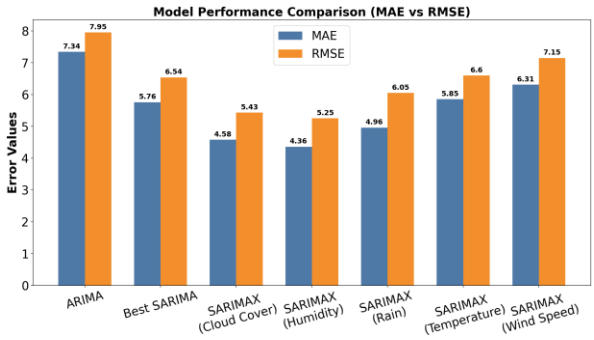

Figure 7 RMSE and MAE should be compared to prediction error.

The mean absolute error (MAE) and the root mean square error (RMSE) are two measurements that may be used in order to assess the prediction mistakes that are associated with the different models. The ARIMA model has a more significant degree of prediction error, as seen by its 7.95 RMSE and 7.34 MAE values. An RMSE of 6.54 and an MAE of 5.75 were the results of the best SARIMA model, which performed better. The SARIMAX models, which took into account a wide variety of climate-related parameters, provided a variety of results. The humidity SARIMAX model performed better than the temperature model, which has an RMSE of 6.60 and an MAE of 5.85. The humidity model has an RMSE of 5.25 and an MAE of 4.36 with a mean absolute error of 4.36. The SARIMAX cloud cover model, which has an RMSE of 5.43 and an MAE of 4.58, is shown to have better accuracy (see Figure 6), which displays the enhanced accuracy. The SARIMAX (wind speed) model, on the other hand, results in a greater number of mistakes, with an RMSE of 7.15 and an MAE of 6.31. The SARIMAX rainfall model is in the middle of the pack as it has a root mean square error of 6.05, and a mean absolute error of 4.96. When it comes to reliably estimating PM2.5 levels, the SARIMAX models which took into consideration cloud cover and humidity performed the best overall. This highlights the significance of these meteorological elements. We can direct future improvements in prediction accuracy by comparing these error measurements to acquire a better understanding of the relative advantages and downsides of each model. This has allowed us to guide future forecasting improvements.

Table 4 MAE's performance metrics.

Model	Q1 MAE	Q2 MAE	Q3 MAE	Q4 MAE
1. Best SARIMA	6.36	6.23	6.57	3.9
2. ARIMA	6.95	6.78	9.71	5.92
3. SARIMAX (Cloud Cover)	4.55	4.1	6.56	3.11
4. SARIMAX (Humidity)	5.22	3.11	5.99	3.12
5. SARIMAX (Rain)	5.91	5.32	5.65	2.95
6. SARIMAX (Temperature)	6.12	6.77	6.39	4.25
7. SARIMAX (Wind Speed)	7.03	6.37	7.6	4.24

The mean absolute error (MAE) and the root mean square error (RMSE) are two-me. The mean absolute error (MAE) and the root mean square

error (RMSE) have two measurements. RMSE and MAE performance indicators for every model in 2023 are shown in Table 4-5 and Figure 8-9. These metrics have been computed for each of the different time periods (First Quarter, Second Quarter, Third Quarter, Fourth Quarter, and Fourth Quarter). There were several metrics that were considered, and they are included in this table.

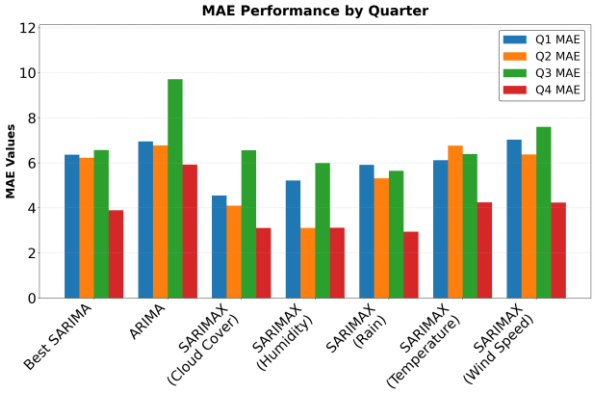

Figure 8 Time-series model performance (MAE).

Table 5 RMSE's performance metrics.

Model	Q1 RMAE	Q2 RMAE	Q3 RMAE	Q4 RMAE
1. Best SARIMA	6.8	6.93	7.41	4.71
2. ARIMA	7.25	7.53	9.79	6.93
3. SARIMAX (Cloud Cover)	5.28	4.57	7.45	3.7
4. SARIMAX (Humidity)	6.16	3.2	7.11	3.37
5. SARIMAX (Rain)	6.53	6.22	7.23	3.55
6. SARIMAX (Temperature)	6.58	7.7	7.17	4.8
7. SARIMAX (Wind Speed)	7.44	6.9	8.69	5.1

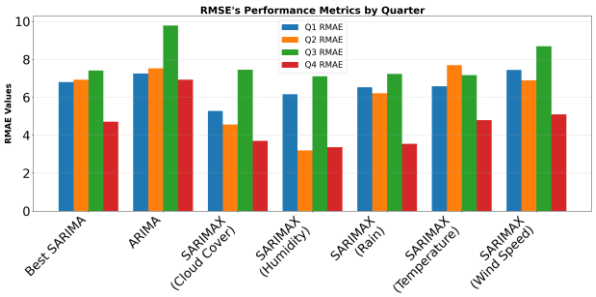

Figure 9 Time-series model performance (RMSE).

The mean absolute error (MAE) and the root mean square error (RMSE) are two-me. According to the information shown in Tables 4 and 5, overall, the SARIMAX model, which uses humidity as an exogenous variable, performs best; in most quarters, it regularly achieves the smallest MAE and RMSE values. Generally speaking, the SARIMAX models perform better than the best SARIMA models and the basic ARIMA models, which suggests that the incorporation of exogenous elements offers a significant improvement in prediction accuracy. In a rather interesting turn of events, the ARIMA model displays the largest errors, notably in Q3. The third quarter was a difficult time for all models, but the fourth quarter was a much better time, especially for the SARIMAX models which took into account the rain and humidity. When every aspect was taken into consideration, the SARIMAX (humidity) model emerges as the most trustworthy model for producing accurate predictions.

The ARIMA model was a key technique in time series forecasting those accounts for temporal correlations and variability throughout time. Its inability to take seasonality into account, particularly when forecasting PM2.5 levels that exhibit clear seasonal variations, was a significant disadvantage. Because ARIMA produced a mean absolute error (MAE) of 7.34 and a root mean squared error (RMSE) of 7.95, the model's performance at measuring clearly showed this issue. Understanding the importance of seasonality in accurately predicting PM2.5 levels, we created the SARIMA model and included a seasonal component to better represent the data's periodic oscillations. This enhancement significantly improved the accuracy of the model and led to a notable drop in both MAE and RMSE, which dropped by 17.6% (down to 6.55) and 21.5% (down to 5.76), respectively.

In order to build the SARIMAX model, we incorporated additional exogenous elements with the SARIMA model. These exogenous factors were temperature, relative humidity, cloud cover, wind speed, and rainfall. The model alone was able to account for variables that impacted PM2.5 variability beyond ARIMA and SARIMA by including these external variables. One important factor that greatly improved the model's predictive power was relative humidity. By applying these enhancements, the SARIMAX model dramatically lowered error metrics: the RMSE declined by 34% to 5.25, and the MAE decreased from the original ARIMA model by 40.6% to 4.36. These findings highlight how crucial it is to take seasonality and other exogenous variables into account when

predicting time series, especially for complex environmental data like PM2.5 levels where variations are driven by a number of interrelated causes.

5. Conclusion

Following the examination of PM2.5 data from Chiang Mai Province, the researchers used three models in order to assess the accuracy of the forecasting process. These models were ARIMA, SARIMA, and SARIMAX. We compared the MAE and RMSE values of each model and found that the SARIMAX model demonstrated the highest level of accuracy. We determined that SARIMAX, using humidity as the external variable, had the lowest MAE and RMSE values. The SARIMAX model, which used cloud cover as an external variable, demonstrated decreased MAE values in both the first quarter and the fourth quarter, as seen by a comparison of similar findings from quarter to quarter. SARIMAX was the most successful model overall because it used humidity as an external variable.

The shift from the ARIMA model to the SARIMA and SARIMAX models reveals a considerable increase in the ability to anticipate the level of PM2.5. Because it took into consideration external factors, and was able to accurately capture the complex external influences on PM2.5 levels, the SARIMAX model offers the greatest degree of projected accuracy. This research emphasizes how crucial it was to have complete models that take into account both external factors and seasonal components in order to attain high predictive accuracy in air quality forecasting.

The results of this study will be used in air quality forecasting systems, enabling the public and government agencies to prepare for PM2.5 pollution events effectively. For example, there will be notifications through mobile applications or systems for planning outdoor activities in high-risk areas. Additionally, public health authorities will be able to use the forecast data to plan the management of medical resources during PM2.5 pollution crises.

6. Acknowledgment

The authors wish to express their sincere appreciation to everyone who provided essential support and assistance for this research. We are especially grateful to Rajamangala University of Technology Lanna for their contribution. We also thank our coworkers and advisors for their invaluable guidance and feedback throughout the process, which was crucial in achieving our goals.

7. References

[1] Adiat KAN, Akeredolu BE, Akinlalu AA, Olayanju GM. Application of logistic regression analysis in prediction of groundwater vulnerability in gold mining environment: a case of Ilesa gold mining area, southwestern, Nigeria. Environ Monit Assess. 2020 Aug; 192(9).

[2] Robles-Velasco A, Cortés P, Muñuzuri J, Baets BD. Prediction of pipe failures in water supply networks for longer time periods through multi-label classification. Expert Syst Appl. 2023 Mar; 213:119050.

[3] Telikani A, Gandomi AH, Shahbahrami A. A survey of evolutionary computation for association rule mining. Inf Sci. 2020 Mar ; 524:318–52.

[4] Hickman L, Thapa S, Tay L, Cao M, Srinivasan P. Text preprocessing for text mining in organizational research: review and recommendations. Organ Res Methods. 2020 Nov ; 25(1):114–46.

[5] Cuenca E, Sallaberry A, Wang FY, Poncelet P. MultiStream: a multiresolution streamgraph approach to explore hierarchical time series. IEEE Trans Vis Comput Graph. 2018 Dec; 24(12):3160–73.

[6] Felzenszwalb PF, Huttenlocher DP. Efficient graph-based image segmentation. Int J Comput Vis. 2004 Apr; 59(2):167–81.

[7] Wang S, Cao J, Yu PS. Deep learning for spatio-temporal data mining: a survey. IEEE Trans Knowl Data Eng. 2020 Sep; 34(8):3681–700.

[8] Kong S, Ji Y, Lü B, Zhao X, Han B, Bai Z. Similarities and differences in PM2.5, PM10 and TSP chemical profiles of fugitive dust sources in a coastal oilfield city in China. Aerosol Air Qual Res. 2014 Jan; 14(7):2017–28.

[9] Chen Y, Shah N, Huggins FE, Huffman GP. Investigation of the microcharacteristics of PM2.5 in residual oil fly ash by analytical transmission electron microscopy. Environ Sci Technol. 2004 Nov; 38(24):6553–60.

[10] Singh K, Singh S, Jha AK, Aggarwal SG, Bisht DS, Murty BP, et al. Mass-size distribution of PM10 and its characterization of ionic species in fine (PM2.5) and coarse (PM10−2.5) mode, New Delhi, India. Nat Hazards. 2013 Mar; 68(2):775–89.

[11] Amnuaylojaroen T, Surapipith V, Macatangay RC. Projection of the near-future PM2.5 in Northern Peninsular Southeast Asia under RCP8.5. Atmosphere. 2022; 13(2):305.

[12] Lakshmi P, Yarlagadda S, Akkineni H, Reddy AM. Big data analytics applying the fusion approach of multicriteria decision making with deep learning algorithms. Int J Eng Trends Technol. 2021 Jan; 69(1):24–8.

[13] Yang C, Huang Q, Li Z, Liu K, Hu F. Big data and cloud computing: innovation opportunities and challenges. Int J Digit Earth. 2016 Nov; 10(1):13–53.

[14] Wang S, Cao J, Yu PS. Deep Learning for Spatio-Temporal Data Mining: A Survey. IEEE Transactions on Knowledge and Data Engineering. 2020 Sep; 34(8):3681–700.

[15] Díaz-Robles LA, Ortega JC, Fu JS, Reed GD, Chow JC, Watson JG, et al. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. Atmos Environ. 2008 Jul; 42(35):8331–40.

[16] Lee D, Lee D, Choi M, Lee J. Prediction of network throughput using ARIMA. In: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC); 2020; Fukuoka, Japan. IEEE; 2020. p. 1–5.

[17] Khashei M, Bijari M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Appl Soft Comput. 2010 Nov; 11(2):2664–75.

[18] Shivhare N, Rahul AK, Dwivedi SB, Dikshit PKS. ARIMA based daily weather forecasting tool: a case study for Varanasi. Mausam. 2021 Nov;70(1):133–40.

[19] Berkeleyearth. Org [Internet]. Independent Data for a Resilient Future, Available from: https://berkeleyearth.org/.

[20] An additional source of environmental information [Internet]. the Hydrological Information Institute, Available from: https://www.hii.or.th/.

[21] Yakubu UA, Saputra MPA. Time series model analysis using autocorrelation function (ACF) and partial autocorrelation function (PACF) for e-wallet transactions during a pandemic. Int J Glob Oper Res. 2022 Aug ;3(3):80–5.

[22] Manigandan P, Alam MS, Alharthi M, Khan U, Alagirisamy K, Pachiyappan D, et al. Forecasting natural gas production and consumption in United States—evidence from SARIMA and SARIMAX models. Energies. 2021 Sep;14(19):6021.