The University of Tehran Press

Pollution

Print ISSN: 2383-451X
Online ISSN: 2383-4501

https://jpoll.ut.ac.ir/

Multioutput Least Square Support Vector Regression Based on Multivariate Exponentially Weighted Moving Variance Control Chart for Air Quality Monitoring in Makassar City

Rafli Setiawan Nasir | Muh. Afdal | Nur Azizah Amra | Erna Tri Herdiani✉

Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar, Indonesia

Article Info

Article type:
Research Article

Article history:
Received: 6 Dec 2025
Revised: 6 June 2026
Accepted: 11 Sep 2026

Keywords:
ARL,
Autocorrelated,
Performance,
Pollutant,
Process Monitoring

ABSTRACT

Air quality monitoring is important for public health, but the characteristics of environmental data, which are multivariate, non-normally distributed, and highly correlated, automatically render conventional statistical control methods ineffective. This study aims to develop an air quality monitoring system in Makassar City using a statistical approach, integrating the Multioutput Least Square Support Vector Regression (MLS-SVR) method and the Multivariate Exponentially Weighted Moving Variance (MEWMV) control chart. This study uses daily data on Particulate Matter 2.5 ($PM_{2.5}$), Sulfur Dioxide (SO_2), and Carbon Monoxide (CO) concentrations from January 2024 to August 2025. The methodology involves using MLS-SVR as a whitening process to model nonlinear relationships and eliminate autocorrelation structures. The resulting independent residuals are then monitored using MEWMV control charts to detect shifts in process variability. The results show that the MLS-SVR model produces high prediction accuracy, as evidenced by an R^2 value of 0.7362, an MSE of 0.2638, and an RMSE of 0.5136. The model also successfully removes significant autocorrelation from the raw data. The MEWMV control chart shows good performance by remaining within control limits despite a significant spike in pollutants during Independence Day celebrations, demonstrating the control Chart's ability to distinguish between predictable structural variations and actual process anomalies. This study concludes that the proposed hybrid method provides a sensitive and effective approach to environmental monitoring. Therefore, this integration is recommended to policymakers as an effective method for early detection of air quality deterioration.

Cite this article: Setiawan Nasir, R., Afdal, M., Azizah Amra, N., & Tri Herdiani, E. (2026). Multioutput Least Square Support Vector Regression Based on Multivariate Exponentially Weighted Moving Variance Control Chart for Air Quality Monitoring in Makassar City. *Pollution*, 12(3), 799-816.
https://doi.org/10.22059/poll.2026.406521.3226

DOI: https://doi.org/10.22059/poll.2026.406521.3226

INTRODUCTION

Quality control is an important aspect of statistics to ensure process stability and detect changes early on. In its development, this concept is not only applied in the manufacturing industry, but also in the environmental field, particularly in air quality monitoring (Miftahul Ihsan et al., 2025). Monitoring the concentration of air pollutants ($PM_{2.5}$, SO_2, and CO) is crucial because small shifts in air pollutant concentrations can have a major impact on public health, including the risk of small airway dysfunction to the risk of death from ischemic heart disease and indirect death (Lee et al., 2024; Z. Xu et al., 2021). However, unlike industrial data, environmental data is autocorrelated, multivariate, and often not normally distributed due to the influence of weather and anthropogenic activities (Q. Liu et al., 2020; Núñez-Alonso et al., 2019). This condition causes conventional methods such as Shewhart control charts or

*Corresponding Author Email: *herdiani.erna@unhas.ac.id*

classical linear regression to be less effective because the assumption of independence is not met (Montgomery, 2012).

To address these limitations, the Exponentially Weighted Moving Average (EWMA) and its multivariate extension (MEWMA) were developed to detect small process shifts as well as correlations between variables (Reynolds, 1996; Saritha & Varadharajan, 2024). Furthermore, since system instability is reflected in the variance, the Multivariate Exponentially Weighted Moving Variance (MEWMV) was introduced to adaptively monitor process variability (Hawkins & Maboudou-Tchao, 2008). However, these exponential control charts remain limited when dealing with nonlinear characteristics and environmental data uncertainty. Therefore, machine learning has emerged as a key solution for capturing these complex relationships (Kim & Moon, 2025). The advantages of machine learning in handling dynamic environmental variables are well-documented and significantly outperform conventional methods in various applications, such as the use of neural networks for continuous water quality monitoring (Noori et al., 2013) and evolutionary polynomial regression for predicting complex river dispersion (Balf et al., 2018).

One of the most widely used machine learning methods is Support Vector Regression (SVR), which is an extension of Support Vector Machine for regression cases. SVR has the advantage of producing globally optimal solutions by minimizing structural risk, which is the generalization error bound that combines training error and regularization to avoid overfitting (Sato et al., 2008). In its simplified version, SVR is modified into Least Square Support Vector Regression (LS-SVR), which simplifies the optimization process from quadratic to linear, making it more computationally efficient (X. Liu et al., 2023).

However, when more than one output variable is observed, the regression problem develops into a multi-output case, leading to the development of Multioutput Least Square Support Vector Regression (MLS-SVR), which is capable of mapping the relationship between several input and output variables simultaneously by considering the nonlinear interdependence between outputs (S. Xu et al., 2013). Unlike conventional LS-SVR, which trains models separately, MLS-SVR trains all outputs simultaneously so that it can take advantage of the interdependence between variables, making it effective in multivariate modeling (Tran et al., 2024). However, although MLS-SVR excels at capturing nonlinear and complex relationships, this method does not yet consider the variance dynamics and temporal fluctuations that are common in time series data.

Therefore, this study focuses on combining the Multioutput Least Square Support Vector Regression model with Multivariate Exponentially Weighted Moving Variance (MLS-SVR MEWMV). This method is used to produce a model that is not only accurate in multivariate prediction but also adaptive to changes in variance structure and time correlation in the data (Khusna et al., 2018). Based on this, this study aims to analyze and monitor the Air Quality Index in Makassar City, which is the center of transportation growth and the largest industrial activity in Eastern Indonesia, thus making it urgent to develop an early detection system.

MATERIALS & METHODS

Study Area and Data

This study is located in Makassar, the largest metropolitan city in Eastern Indonesia, which has extremely high rates of mobility growth and motor vehicle volume. The increase in exhaust emissions resulting from these urban activities necessitates continuous air quality monitoring as a measure to mitigate public health risks. The observation point is focused on the AQICN network monitoring station at the SDN Sudirman Complex, Jenderal Sudirman Street (90115), whose location is highly representative of the downtown area with its daily traffic density. Using daily average data from January 1, 2024, to August 31, 2025, this study analyzes three critical pollutants: Particulate Matter 2.5 ($PM_{2.5}$) originating from combustion and road dust, Sulfur

Dioxide (SO_2) from heavy vehicles and industrial activities, and Carbon Monoxide (CO), which is predominantly generated by traffic congestion emissions along major thoroughfares.

During the process of recording environmental sensor data, the presence of missing values due to instrument calibration or maintenance is common, but must be addressed to avoid disrupting the autocorrelation structure of the time series data. To address this, the cubic spline interpolation method is used, which can estimate missing values while preserving the original trend and natural fluctuations of pollutants, thereby avoiding the artificial smoothing effect often seen in standard mean imputation methods. Once the data structure is confirmed to be intact, the pre-processing stage concludes with a Z-score standardization process. This step aims to balance the scale ranges between variables, allowing the MLS-SVR algorithm model to run optimally without introducing bias toward pollutant types with higher baseline values.

Data Preprocessing

To ensure data acceptability and rigorous Quality Assurance/Quality Control (QA/QC) practices before modeling, an initial screening was conducted. This step revealed missing values across the PM2.5, SO2, and CO datasets, which if left untreated, could introduce significant bias into the control chart limits. To address this, data cleansing was performed using cubic spline interpolation (Le Thi Thu, 2024). In accordance with international environmental data QA/QC guidelines, arbitrary removal of pollutant data is not permitted in order to preserve time series integrity and ensure traceability of analytical results (ISO/IEC 17025:2017, Clauses 7.7 and 8.4) (Hadeed et al., 2020). Data gaps (<15%) were addressed using cubic spline interpolation (Thu, 2024), which was validated through RMSE comparison and stationarity testing. This approach ensured data completeness exceeding 75% (EPA data capture requirement) without compromising the representation of the volatile patterns of PM2.5, SO2, and CO (Iso, 2017).

Furthermore, because Support Vector Regression (SVR) is highly sensitive to feature scales, and considering the pollutants are measured in different magnitudes (e.g., $\mu g / m^3$ for $PM_{2.5}$ and SO_2 versus mg / m^3 for CO), a standardization process was strictly implemented. The input features were transformed using Z-score standardization, scaling the data to have a mean of zero and a standard deviation of one. This step prevents variables with larger numeric ranges from dominating the model's objective function. Finally, to validate the proposed model, the standardized dataset was partitioned into a Calibration Phase (January 2024 – June 2025) to train the model and establish the baseline control limits, and a Monitoring Phase (July 2025 – August 2025) to evaluate the system's detection capability in a testing scenario.

MEWMV-Multioutput Least Square Support Vector Regression

Multioutput Least Square Support Vector Regression is a machine learning method that studies mapping modeling from multivariate inputs to multivariate outputs. Unlike the LS-SVR algorithm, which processes each output separately, MLS-SVR overcomes this limitation by not ignoring the potential for nonlinear relationships between different output variables. Meanwhile, the MEWMV control chart is designed to detect shifts in process variability in multivariate cases as well as changes in process averages during the control period (Huwang et al., 2007). This approach is supported by Noori et al. (2012),who demonstrated that SVM-based models are highly effective for rapid, online prediction of complex parameters with high accuracy. Furthermore, Najafzadeh et al. (2021) highlighted that SVM provides the lowest level of uncertainty compared to other machine learning models, ensuring robust and stable monitoring performance. To address the autocorrelation problem found in the observation output, the MLS-SVR algorithm is used. The MLS-SVR model is trained using the partial autocorrelation function (PACF) input vector, namely $lag - p_1, p_2, \ldots, p_m$. If $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m$ are observation output vectors containing multivariate autocorrelation, where $\boldsymbol{y}_i = \left(\boldsymbol{y}_{1j}, \boldsymbol{y}_{2j}, \ldots, \boldsymbol{y}_{nj}\right)^T$ with $j = 1, 2, \ldots, m$ is the sum of the MLS-SVR output. This study utilizes the Radial Basis Function (RBF) kernel within the MLS-SVR model. The RBF kernel function was chosen because it is easy and efficient in terms of computation time, and is used to implement models in higher dimensions without having to define a mapping function from the initial dimension to the higher dimension (Zhou et al., 2023). Thus, the input of the MLS-SVR model can be written using the following equation (Khusna et al., 2018).

$$\boldsymbol{X} = \left(\boldsymbol{y}_{1,(i-1)},\ldots,\boldsymbol{y}_{1,(i-p_1)},\ldots,\boldsymbol{y}_{1,(i-1)},\ldots,\boldsymbol{y}_{1,(i-p_j)},\ldots,\boldsymbol{y}_{m,(i-1)},\ldots,\boldsymbol{y}_{m,(i-pm)}\right)$$

The function of the MLS-SVR model with optimal parameter $\hat{f}(\boldsymbol{x})$ allows the residual to be calculated using $e_j = y_j - \hat{f}(\boldsymbol{x}_j)$, which is the difference between the observed value and the predicted value that becomes the independent data from autocorrelation. The residual vector $e_1, e_2, \ldots, e_j, \ldots, e_m$ is defined as the result of the MLS-SVR model with the output vector $y_1, y_2, \ldots, y_j, \ldots, y_m$. Then, the MEWMV control chart formula is obtained as follows (Huwang et al., 2007).

$$\boldsymbol{V}_i = \omega\left(\boldsymbol{e}_i - \mu_i^*\right)\left(\boldsymbol{e}_i - \mu_i^*\right)' + (1-\omega)\boldsymbol{V}_{i-1}$$

Where:
ω : weight value of $0 < \omega < 1$
e_i : MLS-SVR model residual as MEWMV estimate
μ_i^* : natural estimate for the mean process at time i and e_i, and

$$\boldsymbol{V}_0 = \left(\boldsymbol{e}_1 - \mu_i^*\right)\left(\boldsymbol{e}_1 - \mu_i^*\right)'$$

with estimates of μ_i^* for changes in the average process at time i defined as follows.

$$\mu_i^* = \lambda\boldsymbol{e}_i + (1-\lambda)\mu_{i-1}^*, 0 < \lambda < 1, \mu_0^* = 0.$$

Subsequently, to ensure optimal predictive performance, this study employed the Grid Search method (Hsu et al., 2003; Syarif et al., 2016). The Grid Search method was utilized to determine the most effective combination of hyper-parameters by systematically evaluating the error metric across grid points with equal parameter spacing. The hyperparameters subjected to optimization included the RBF kernel parameter set and regularization parameters, specifically $\gamma \in \{2^{-5}, 2^{-3}, \ldots, 2^{15}\}$, $\lambda \in \{2^{-10}, 2^{-8}, \ldots, 2^{10}\}$, and $p \in \{2^{-15}, 2^{-13}, \ldots, 2^{3}\}$ (S. Xu et al., 2013). The Mean Square Error (MSE) was calculated for each output combination, with the parameter set yielding the lowest MSE value being designated as the optimal configuration for the predictive model. Following the optimization of the forecasting model, the final performance of the MEWMV control chart was comprehensively assessed using the Average Run Length (ARL) value. This ARL evaluation step serves to quantify the speed and accuracy with which the control chart can detect abnormal shifts in the controlled process (Montgomery, 2012). The literature provides analytical formulations for calculating the ARL of control charts, including moving average charts. However, (Zhang et al., 2004) noted corrections to the expected run length formula for moving average charts, suggesting the previous formula served only as an upper bound for the actual ARL. Furthermore, (Li et al., 2014) found that classical methods often lack accuracy, positioning Markov chain methods and integral equations as more reliable alternatives for calculating ARL across various control chart types (e.g., CUSUM, EWMA), thereby providing a more realistic and accurate measure of control chart performance.

Evaluation Metrics

To evaluate the predictive performance of the proposed MLS-SVR model, this study used three commonly applied regression metrics, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). These metrics are widely used in machine learning-based prediction studies because they provide complementary information on prediction error and model fit. (Gzar et al., 2022) emphasized the importance of using appropriate error-based measures in regression model assessment, while Almansour and (Almansour & Alsaab, 2025) and (Arunadevi et al., 2025) also employed these metrics to compare the predictive accuracy of machine learning models.

MSE was used to measure the average prediction error by giving greater emphasis to larger deviations between observed and predicted values, whereas RMSE was used to express the prediction error in the same unit as the original data so that it is easier to interpret. Meanwhile, R^2 was used to assess how well the model explains the variability of the observed data. The three metrics are expressed as follows:

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \hat{y}_i)^2}$$

Where:

y_i : observed value

$\hat{y}_i$: predicted value

y : mean of the observed values

n : number of observations

In general, lower MSE and RMSE values indicate better predictive accuracy, while an R^2 value closer to 1 indicates better model fit and stronger explanatory power. Therefore, in this study, the MLS-SVR model was considered to perform well when it produced low error values and a high R^2 value (Chicco et al., 2021).

Air Quality Index Control

The air quality index is a tool for measuring air pollution levels, which includes measuring the concentration of various pollutants, such as ozone, particulates, carbon monoxide, sulfur dioxide, and nitrogen dioxide. Monitoring the air quality index is vital because it has a direct impact on health. In particular, monitoring the concentration of air pollutants ($PM_{2.5}$, SO_2, and CO) is crucial because small shifts in air pollutant concentrations can have a major impact on public health, including the risk of small airway dysfunction to the risk of death from ischemic heart disease and indirect death (Lee et al., 2024; Xu et al., 2021). Previous studies, Xie & Qiu (2023), proposed a control chart for dynamic processes with applications in air pollution monitoring. Insiyah et al. (2023), proposed air monitoring with 2T -Hotelling using a Vector Autoregressive model. Y. Liu & Zi (2021) used MEWMA for pollutant data. Alfasanah et al. (2024) used individual charts and EWMA with XGBoost and SVR residual models.

Research Workflow

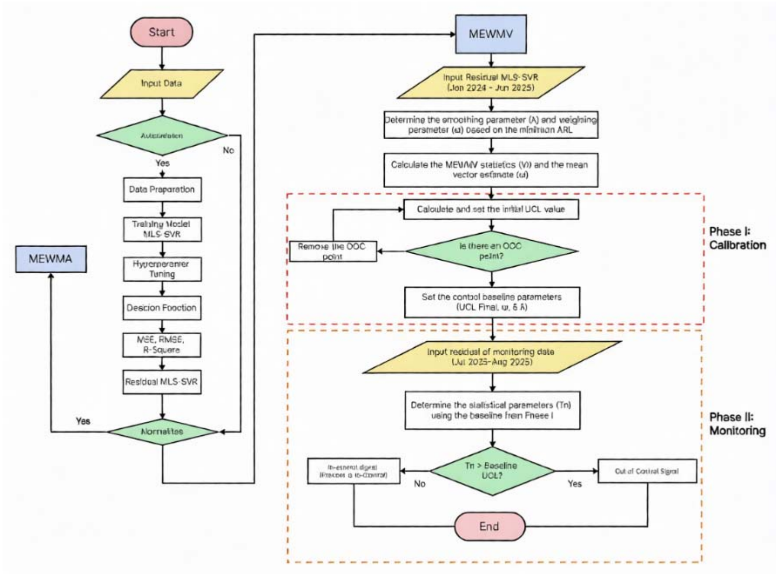


Fig. 1. Workflow MEWMV based residual MLS-SVR

The proposed methodological framework is depicted in Figure 1, illustrating the integration of MLS-SVR modeling with MEWMV control chart procedures across the calibration and monitoring phases.

Based on Figure 1, the study consists of two main phases: calibration and monitoring. The process begins with data preprocessing and MLS-SVR model development, including hyperparameter tuning and performance evaluation. The resulting residuals are then used in the MEWMV framework. In the calibration phase, optimal parameters and control limits are determined, while out-of-control points are iteratively removed. In the monitoring phase, new data are evaluated against the established control limits to detect whether the process is in-control or out-of-control.

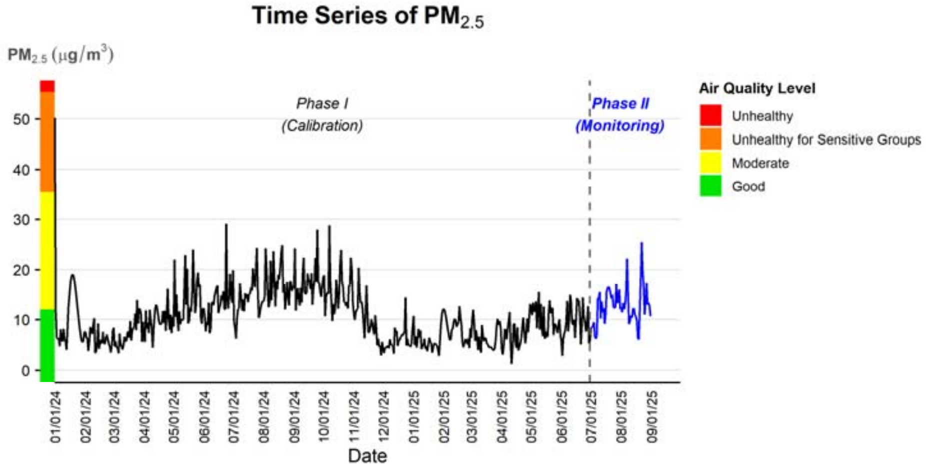
RESULTS AND DISCUSSION

Data Characteristic

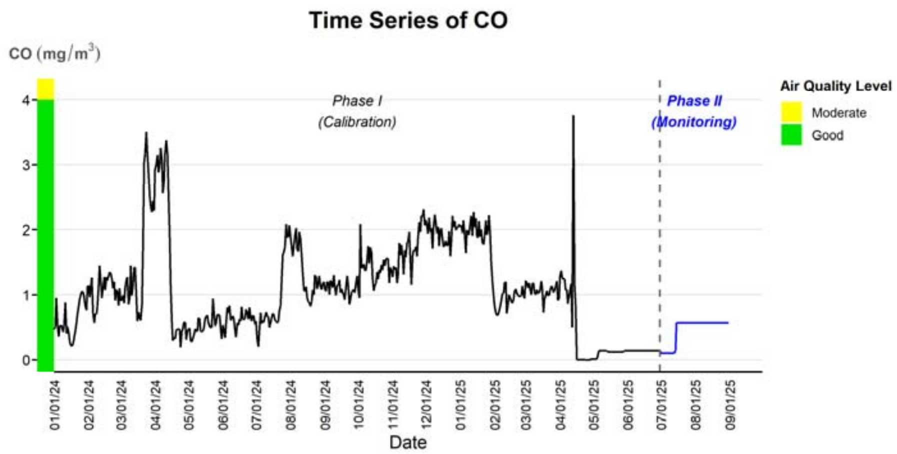
Figure 2 illustrates the concentration levels of $PM_{2.5}$, SO_2, and CO in Makassar City during the Calibration (Phase I) and Monitoring (Phase II) periods. Prior to this visualization, the dataset was preprocessed using cubic spline interpolation to effectively handle missing values while preserving the natural volatility of the pollutants. These measurements are analyzed alongside the air quality categories outlined in the reference table, which utilizes a color-coded system to define safety thresholds from 'Good' (Green) to 'Hazardous' (Black).

During Phase I, the data reveals frequent spikes into the 'Unhealthy' category, particularly for $PM_{2.5}$ and SO_2. From an environmental perspective, these persistent elevations are likely driven by intensive anthropogenic factors such as high traffic volumes during peak commuting hours and local industrial emissions within the densely populated urban center. Conversely, CO concentrations remained largely within the 'Good' to 'Moderate' ranges. This suggests that carbon monoxide emissions, which are subject to strict regulations in modern vehicle exhausts, were more effectively managed and dispersed during this period.

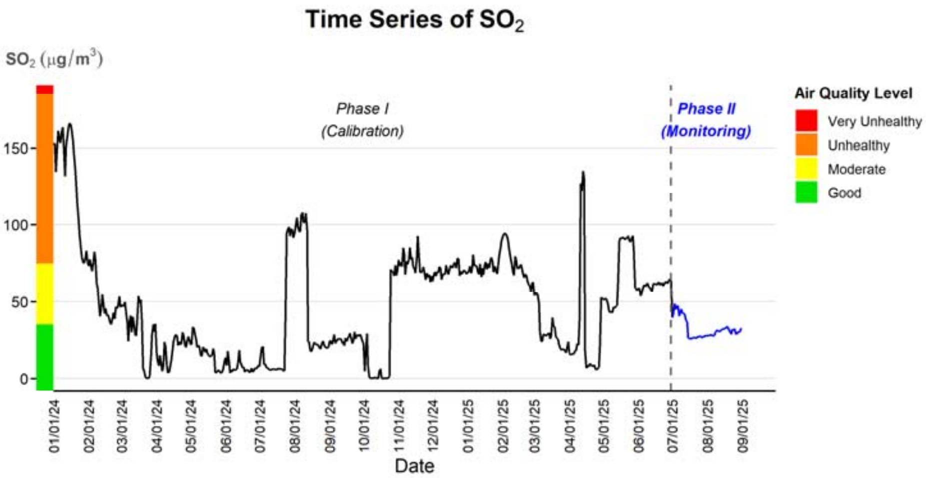
In Phase II, the data indicates a general stabilization across all parameters within safer limits. However, the intermittent 'Moderate' levels of $PM_{2.5}$ observed in this phase confirm that baseline pollution remains a continuous threat. This condition underscores the necessity for a sensitive surveillance system capable of detecting subtle environmental deteriorations before



(a)



(b)



(c)

Fig. 2. Time series of air pollutants in Makassar City (Jan 2024 – Aug 2025): (a) $PM_{2.5}$, (b) CO, and (c) SO_2.

Table 1. Air Quality Level Thresholds

Air Quality Category	Color	$PM_{2.5}$ ($\mu g/m^3$)	SO_2 ($\mu g/m^3$)	CO (mg/m^3)
Good	Green	0 – 15.5	0 – 35	0 – 4
Moderate	Yellow	15.6 – 55.4	36 – 75	4.1 – 9
Unhealthy	Orange	55.5 – 150.4	76 – 180	9.1 – 15
Very Unhealthy	Red	150.5 – 250.4	181 – 300	15.1 – 30
Hazardous	Black	> 250.5	> 300	> 30

Reference: Indonesian Ministry of Environment and Forestry Regulation No. 14 of 2020

Table 2. Univariate Normality Test Result

Parameter	W	*p-value*	Conclusion
$PM_{2.5}$	0.9542	<0.000	
CO	0.9444	<0.000	Not normal
SO_2	0.9558	<0.000	

they reach hazardous levels.

Pre-modeling Analysis

Before proceeding with the development of the control chart, it is essential to verify the stochastic properties of the pollutant time series to ensure the validity of our model. First, we assessed the distributional assumptions using the multivariate Henze-Zirkler test, which was selected for its proven consistency and superior statistical power compared to other multivariate alternatives (Mecklin & Mundfrom, 2005). The test yielded a *p-value* < 0.001, which was further supported by univariate Shapiro-Wilk tests for PM2.5, SO2, and CO, all of which also returned *p-value* below 0.001 as presented in Table 2. The Shapiro-Wilk test was specifically chosen as it is widely recognized as the most reliable normality test for univariate data across various sample sizes (Razali & Wah, 2011). Collectively, these findings confirm that the dataset does not follow a normal distribution.

From an environmental perspective, this non-normality reflects the dynamic nature of urban air quality in Makassar City. Instead of clustering neatly around a static average, pollutant concentrations are subject to sudden, extreme fluctuations. These spikes are primarily driven by human activities, including heavy traffic during rush hours, industrial emissions and meteorological constraints (e.g., wind stagnation preventing pollutant dispersion).

Furthermore, a critical challenge in environmental monitoring is the presence of autocorrelation. The Multivariate Cross-Correlation Function (MCCF) test revealed significant systematic correlations in the raw data up to lag 14. This is visually demonstrated in Figure 3(a), where the “Raw Data” panel is dominated by colored tiles (red and blue), indicating strong interdependencies between variables across time lags. Practically, this signifies a “long memory effect” in the atmosphere of Makassar which leads to condition where the pollutant particles accumulated today do not dissipate immediately but continue to influence air quality for up to 14 lags. This persistence creates a systematic dependency that violates the assumption of independence required by standard control charts.

Further analysis of these cross-correlations confirms that environmental variables are not isolated from one another. A univariate approach monitoring each gas separately would likely overlook this crucial synergistic information. Moreover, if conventional quality control methods like the Hotelling T^2 or standard MEWMA were applied directly to this raw data will make the high autocorrelation would be misinterpreted as a process shift. This would inevitably trigger excessive false alarms, causing the system to signal a danger state when the process is simply exhibiting its natural persistence. Consequently, the MLS-SVR modeling step is not just an

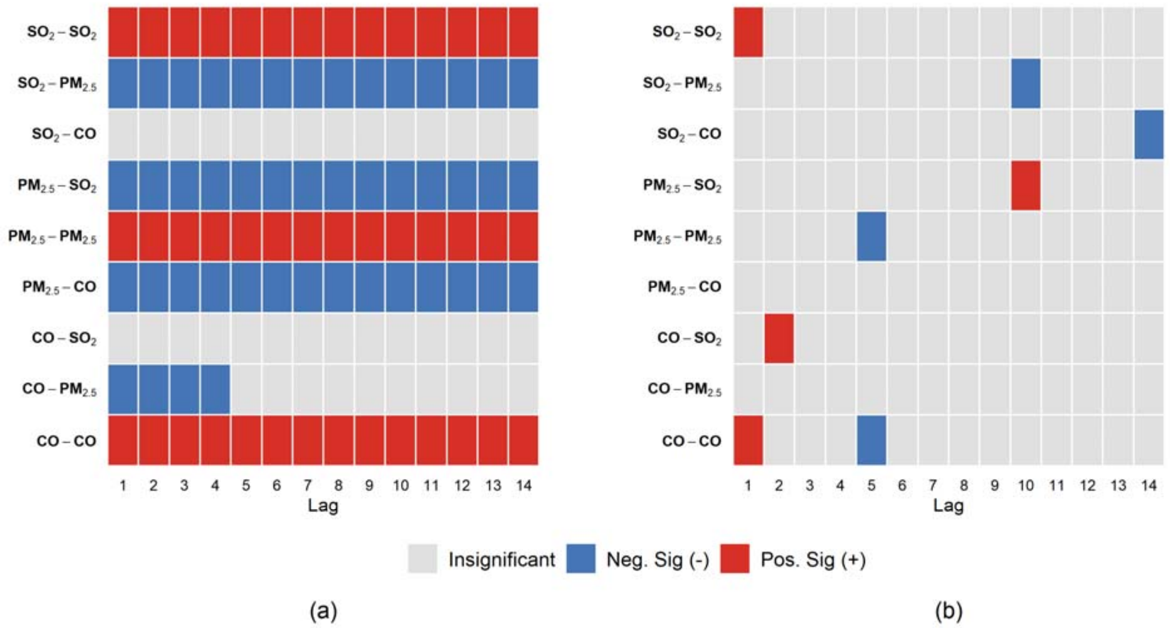


Fig. 3. Schematic visualization of the Multivariate Cross-Correlation Function (MCCF) matrices. (a) The significant autocorrelation structure in the raw AQI data (Phase I), indicating strong temporal dependence. (b) The uncorrelated residuals after MLS-SVR modeling, indicating successful whitening.

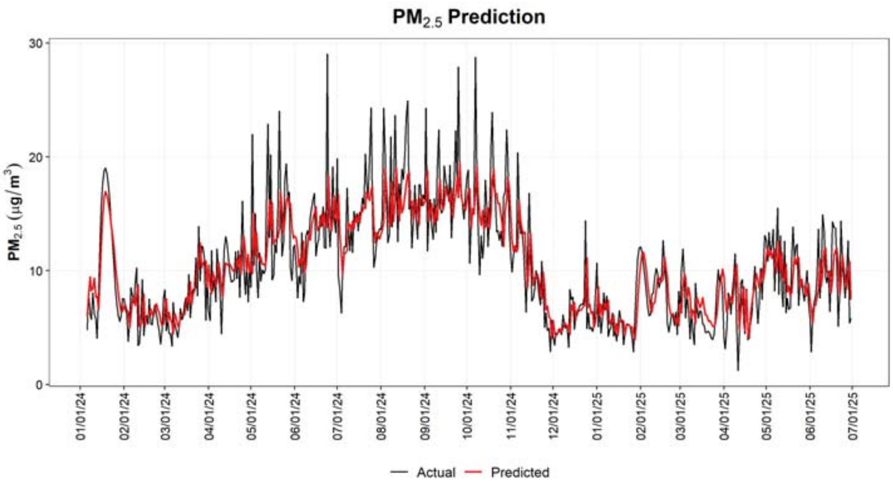
enhancement but a statistical necessity to whiten the data and filter out this autocorrelation structure before any variance monitoring can be validly performed.

MLS-SVR Modelling Results

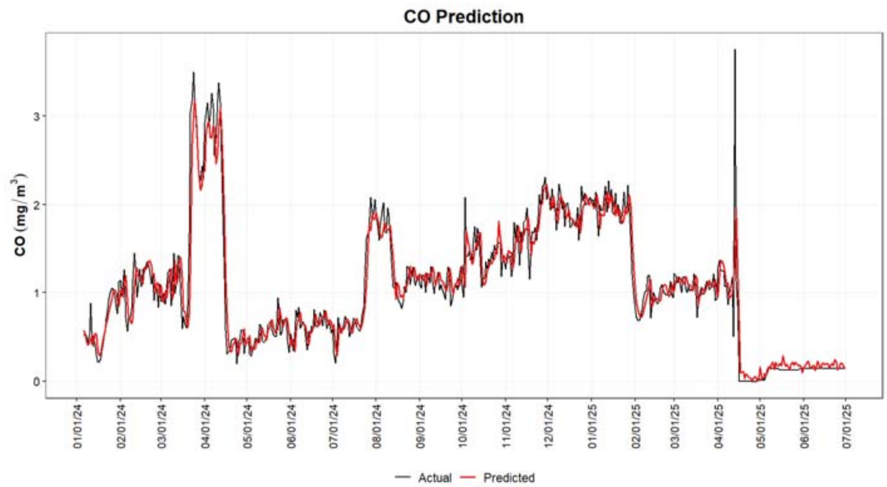
The MLS-SVR algorithm was employed to capture the nonlinear relationships and autocorrelation within the multivariate time series. To achieve optimal generalization performance, we utilized a grid search algorithm with 5-fold cross-validation was utilized to tune the hyperparameters of the RBF kernel. Proper tuning of these parameters is crucial, because the performance of the RBF kernel in SVR is highly sensitive to the selection of regularization and kernel width, which dictates the trade-off between model complexity and error tolerance as emphasized both by Cervantes et al. (2020) and Noori et al. (2015).

Based on the minimum Mean Square Error (MSE) criterion, we identified the optimal configuration from 1210 possible combinations as $\gamma = 2^{-5}, \lambda = 2^{-6}$, and $p = 2^{3}$. This specific configuration yielded an MSE of 0.2638, an RMSE of 0.5136, and an R^2 value of 0.7362. These results indicate that the model can explain approximately 73.62% of the variance in the pollutant data, reflecting high predictive accuracy for environmental monitoring. Figures 4, 5, and 6 sequentially visualize the model's performance, demonstrating that the predicted values closely track the actual fluctuations of $PM_{2.5}$, SO_2, and CO while capturing both the underlying trends and local volatility.

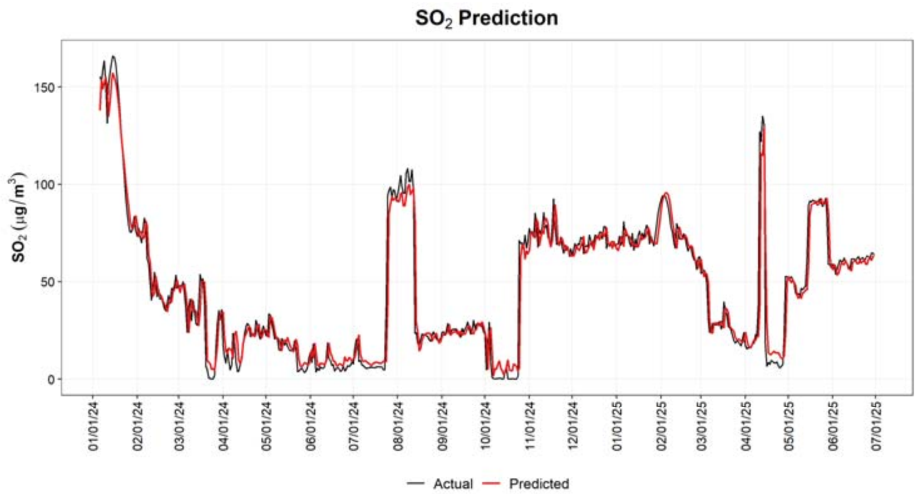
Most importantly, the model's ability to whiten the data is confirmed in Figure 3(b). In stark contrast to the raw data shown in Figure 3(a), the residual heatmap is dominated by white tiles, which represent insignificant correlations. This visual evidence confirms that the MLS-SVR successfully removed the systematic autocorrelation structure. The residuals now satisfy the white noise assumption. Consequently, these residuals represent the pure stochastic variability of the process, making them suitable inputs for the MEWMV control chart. This approach aligns with the nonlinear SVR-based control chart methodology proposed by (Khediri et al., 2010), which effectively isolates process faults by monitoring the uncorrelated residuals derived from the SVR prediction model.



(a)



(b)



(c)

Fig. 4. Time series comparison of Actual vs. Predicted SO_2 parameter using the optimized MLS-SVR model. (a) $PM_{2.5}$, (b) CO, and (c) SO_2.

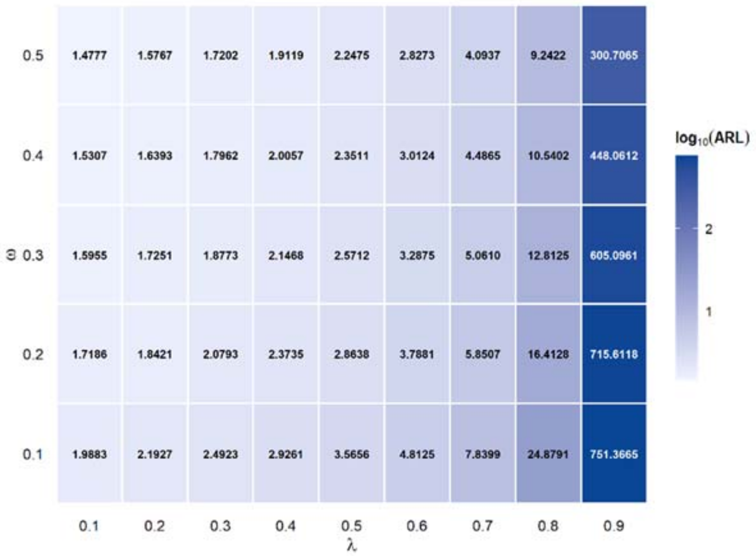


Fig. 5. ARL Performance Heatmap.
Darker colour indicates higher ARL (slower detection), while lighter colour indicates lower ARL (faster detection).

Table 3. Cleaning Iteration History

Iteration	Number of Observations	UCL	Number of OOC	Status
1	542	24,4610	10	Unstable
2	532	14,1214	7	Unstable
3	525	11,5262	7	Unstable
4	518	10,3730	1	Unstable
5	517	10,2340	0	Stable

MEWMV Control Chart Phase I: Calibration

The calibration phase aims to establish a robust baseline for "normal" air quality variability in Makassar. In the MEWMV control chart, the selection of the smoothing parameter (λ) and the weighting parameter (ω) is critical for balancing sensitivity and stability. These parameters were determined based on the Average Run Length (ARL) performance to ensure the fastest detection of shifts (lowest ARL_1). A performance heatmap (Figure 5) was generated for candidate parameters $\lambda \in [0.1, 0.9]$ and $\omega \in [0.1, 0.5]$. The analysis identified the optimal combination as $\omega = 0.5$ and $\lambda = 0.1$, which produced an ARL_1 of 1.4777. This indicates the chart's high sensitivity in detecting process shifts as early as the first or second observation after a disturbance.

The establishment of robust control limits in Phase I involved a rigorous iterative cleaning procedure to ensure that the baseline truly represents the natural process variability, free from external anomalies. As detailed in Table 3, the process required five iterations to achieve statistical stability. Throughout these iterations, the number of observations was refined from 542 to 517, resulting in a progressive tightening of the control limit from 24.4610 to a final value of 10.2340. The result of this cleaning process is visualized in Figure 7, which demonstrates a statistically stable condition (zero out of control points). This drastic reduction is not merely a numerical adjustment but the result of systematically eliminating historical assignable causes.

Furthermore, the Out-of-Control (OOC) points removed during the initial iterations, particularly the cluster observed in April 2024, reveals a strong correlation with a major social event, such as *Eid al-Fitr.* During this period, Makassar typically experiences extreme changes

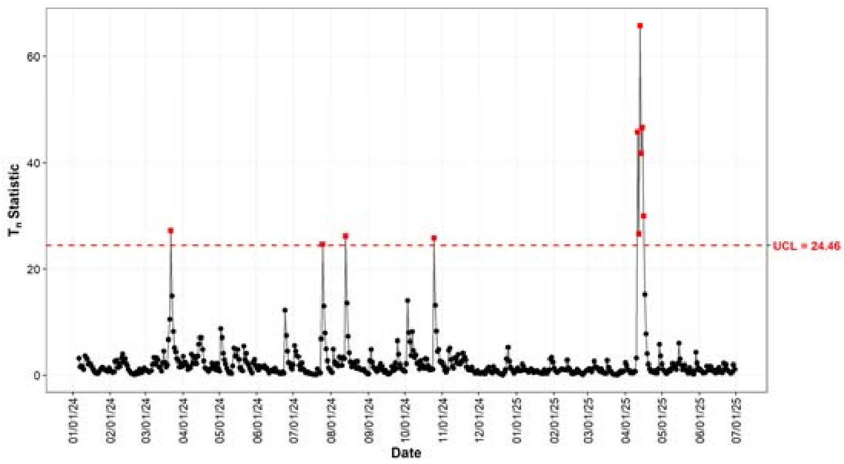


Fig. 6. Uncleaned Phase I MEWMV Chart

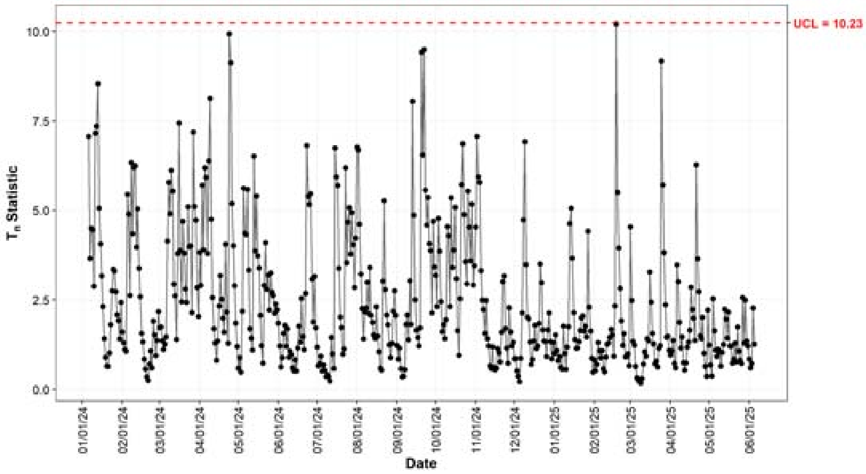


Fig. 7. Cleaned Phase I MEWMV Chart

in mobility patterns, ranging from the massive exodus of 'mudik' (homecoming) traffic to surges in domestic activities. By excluding this holiday period from the calibration phase, the MEWMV model is effectively trained to recognize the pollution patterns of standard working days as the proper baseline. This exclusion ensures that the resulting control chart is more sensitive to genuine anomalies in the future, rather than being desensitized by predictable annual outliers.

MEWMV Control Chart Phase II: Monitoring

The established baseline parameters ($\lambda = 0.1, \omega = 0.5$, and $UCL = 10.2340$) were applied to the monitoring dataset spanning from July to August 2025. As shown in Figure 8, the multivariate monitoring statistic (T_n) fluctuated throughout this period, but remained consistently below the UCL or entirely within the control limits. This in-control status implies that the variability of air pollutants in Makassar during this period followed the historical stochastic patterns identified in Phase I. Furthermore, it indicates an absence of extreme shock events that might have disrupted the regular pollution dynamics of the city.

The ultimate test of the model's robustness occurred around the 45th observation in Phase II,

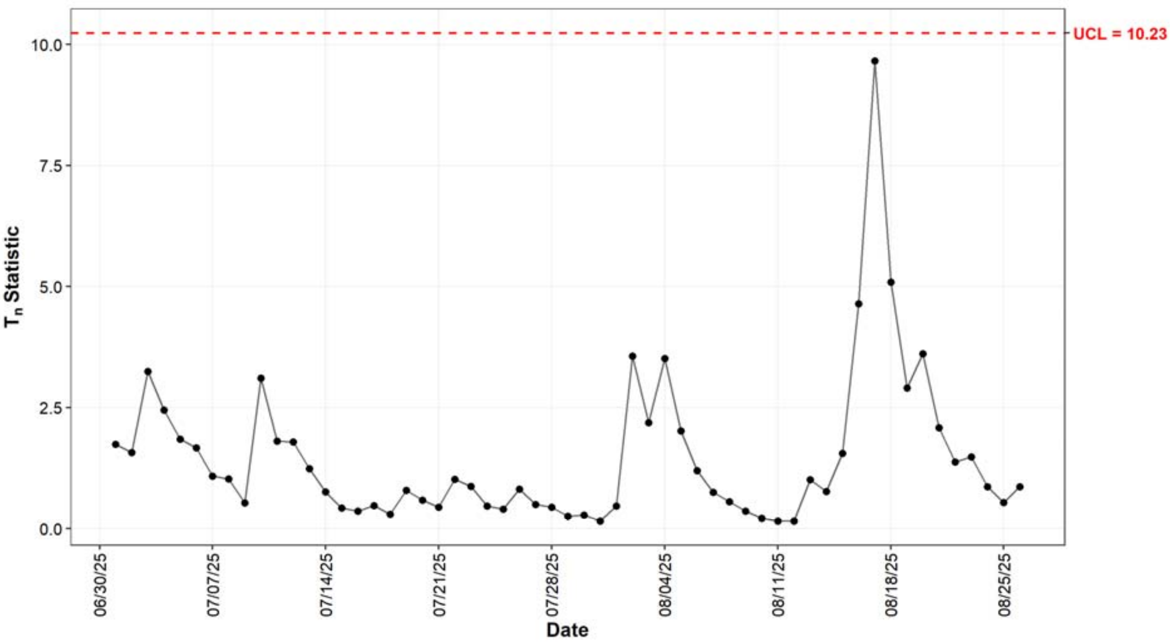


Fig. 8. Phase II MEWMV Control Chart

which corresponds to August 17, 2025 (Indonesia's Independence Day). During this event, the monitoring data recorded a significant surge in pollutant activity. As detailed in Table 4, the T_n statistic reached its maximum value of 9.6616. Contextually, this spike can be attributed to the increased mass mobilization and vehicular parades occurring throughout Makassar City. This observation aligns with the documented "Holiday Effect" in urban environments, where such public celebrations are known to trigger temporary surges in anthropogenic emissions (K. Xu et al., 2020).

However, the most critical finding from this event is the response of the MLS-SVR MEWMV control system. Despite the drastic surge that approached the upper limit, the T_n statistic remained strictly in-control (below the UCL of 10.2340). This result demonstrates the superior robustness of the proposed method, as the system successfully distinguished between "high-variance fluctuations that are structurally consistent" and "chaotic process anomalies." The model recognized that although the magnitude of pollution increased on Independence Day, the covariance structure between variables ($PM_{2.5}$, SO_2, and CO) remained consistent with the historical patterns learned during the training phase.

Unlike mean-based control charts which would likely have flagged this celebration as an out-of-control, the MEWMV chart correctly categorized it as an allowable variation within a known event, thereby preventing false alarms. This capability supports the arguments made by Abbas et al. (2014), who emphasize that independent dispersion monitoring is critical to uncover process instabilities often missed by mean-based controls. Ultimately, this performance proves that the integration of MLS-SVR and MEWMV provides a highly adaptive framework for real-time environmental monitoring policies.

CONCLUSIONS

This study demonstrates that the integration of the MLS-SVR and MEWMV algorithms is effective in monitoring air quality dynamics in the city of Makassar. The MLS-SVR model has proven reliable as a whitening process capable of eliminating complex autocorrelation

Table 4. Plotting Statistics of Phase II

Date	Tn	Status
07/01/25	1,7371	In-Control
07/02/25	1,5707	In-Control
07/03/25	3,2475	In-Control
07/04/25	2,4506	In-Control
07/05/25	1,8468	In-Control
07/06/25	1,6633	In-Control
07/07/25	1,0791	In-Control
07/08/25	1,0284	In-Control
07/09/25	0,5282	In-Control
07/10/25	3,1050	In-Control
07/11/25	1,8050	In-Control
07/12/25	1,7861	In-Control
07/13/25	1,2367	In-Control
07/14/25	0,7555	In-Control
07/15/25	0,4242	In-Control
07/16/25	0,3570	In-Control
07/17/25	0,4692	In-Control
07/18/25	0,2973	In-Control
07/19/25	0,7875	In-Control
07/20/25	0,5885	In-Control
07/21/25	0,4382	In-Control
07/22/25	1,0139	In-Control
07/23/25	0,8737	In-Control
07/24/25	0,4644	In-Control
07/25/25	0,3987	In-Control
07/26/25	0,8126	In-Control
07/27/25	0,5001	In-Control
07/28/25	0,4386	In-Control
07/29/25	0,2557	In-Control
07/30/25	0,2783	In-Control
07/31/25	0,1586	In-Control
08/01/25	0,4674	In-Control
08/02/25	3,5620	In-Control
08/03/25	2,1899	In-Control
08/04/25	3,5152	In-Control
08/05/25	2,0202	In-Control
08/06/25	1,1964	In-Control
08/07/25	0,7494	In-Control
08/08/25	0,5519	In-Control
08/09/25	0,3596	In-Control
08/10/25	0,2159	In-Control
08/11/25	0,1581	In-Control
08/12/25	0,1572	In-Control
08/13/25	1,0097	In-Control
08/14/25	0,7615	In-Control
08/15/25	1,5523	In-Control
08/16/25	4,6423	In-Control
08/17/25	9,6616	In-Control
08/18/25	5,0918	In-Control
08/19/25	2,9005	In-Control
08/20/25	3,6072	In-Control
08/21/25	2,0813	In-Control
08/22/25	1,3779	In-Control
08/23/25	1,4800	In-Control
08/24/25	0,8656	In-Control
08/25/25	0,5383	In-Control
08/26/25	0,8587	In-Control

structures in raw pollutant data such as $PM_{2.5}$, SO_2, and CO, thereby producing residuals that satisfy statistical assumptions without autocorrelation. The use of the MEWMV control band provides a high degree of adaptability, allowing the system to maintain an in-control status even during emission spikes at specific times, such as Independence Day celebrations, which typically trigger false alarms in conventional methods. This capability demonstrates the system's added value in distinguishing between normal structural fluctuations and actual environmental anomalies.

From a policy perspective, this MLS-SVR MEWMV method offers tangible benefits to environmental authorities in the form of a more precise early-warning system. Its utility lies in its ability to monitor variability, which often serves as an early indicator of deteriorating air quality before average pollutant levels exceed hazardous thresholds. With more accurate information and fewer false alarms, policymakers can allocate resources more efficiently. Such policies include the implementation of situational traffic engineering or timely monitoring of industrial emissions at specific locations. This is crucial for preventing long-term exposure to pollutants without resorting to drastic and unnecessary public interventions.

Although it offers promising results for addressing pollution issues in Makassar City, this study has several limitations. The current model focuses solely on historical pollutant concentration data without incorporating exogenous meteorological variables such as wind speed, wind direction, temperature, and humidity, which physically influence air dispersion patterns. Furthermore, the study's geographical scope is currently limited to monitoring points in Makassar City, so the generalizability of the model to regions with different geographical characteristics requires further testing.

As a recommendation for future research, integrating climate and weather variables into the MLS-SVR model is highly recommended to improve the prediction accuracy and sensitivity of the control chart method. Furthermore, future studies could compare the efficiency of this model with more advanced deep learning architectures, such as Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), or Generative Adversarial Network (GAN), to explore optimization potential in broader real-time air quality monitoring systems.

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare that there is not any conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy has been completely observed by the authors.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

Abbas, N., Riaz, M., & Does, R. J. M. M. (2014). Memory-type Control Charts for Monitoring the Process Dispersion. *Quality and Reliability Engineering International*, *30*(5), 623–632. https://doi.org/10.1002/qre.1514

Alfasanah, Z., Niam, M. Z. H., Ahsan, M., Wardiani, S., & Lee, M. H. (2024). *Monitoring Air Quality*

Index with EWMA and Individual Charts Using Xgboost and SVR Residuals. https://doi.org/10.2139/ssrn.4948975

Almansour, K., & Alsaab, H. O. (2025). Machine learning analysis and simulation of pharmaceutical drying process based on prediction of concentration distribution and mass transfer. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-22276-9

Arunadevi, M., Karthikeyan, B., Shrihari, A., Saravanan, S., Sundararaju, K., Palanisamy, R., Awad, M., Mahmoud, M. M., Mbadjoun Wapet, D. E., Ayidh, A. Al, Hussein, H. S., Hussein, M. M., & Omar, A. I. (2025). Prediction of Optimum Operating Parameters to Enhance the Performance of PEMFC Using Machine Learning Algorithms. *Energy Exploration and Exploitation*, *43*(2), 676–698. https://doi.org/10.1177/01445987241290535

Balf, M. R., Noori, R., Berndtsson, R., Ghaemi, A., & Ghiasi, B. (2018). Evolutionary polynomial regression approach to predict longitudinal dispersion coefficient in rivers. *Journal of Water Supply: Research and Technology - AQUA*, *67*(5), 447–457. https://doi.org/10.2166/aqua.2018.021

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trend. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623

Gzar, D. A., Mahmood, A. M., & Abbas, M. K. (2022). A Comparative Study of Regression Machine Learning Algorithms: Tradeoff Between Accuracy and Computational Complexity. *Mathematical Modelling of Engineering Problems*, *9*(5), 1217–1224. https://doi.org/10.18280/mmep.090508

Hadeed, S. J., O'Rourke, M. K., Burgess, J. L., Harris, R. B., & Canales, R. A. (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of the Total Environment*, *730*. https://doi.org/10.1016/j.scitotenv.2020.139140

Hawkins, D. M., & Maboudou-Tchao, E. M. (2008). Multivariate Exponentially Weighted Moving Covariance Matrix. *Technometrics*, *50*(2), 155–166. https://doi.org/10.1198/004017008000000163

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). *A Practical Guide to Support Vector Classification*. http://www.csie.ntu.edu.tw/~cjlin

Huwang, L., Yeh, A. B., & Wu, C. W. (2007). Monitoring Multivariate Process Variability for Individual Observation. *Journal of Quality Technology*, *39*(3), 258–278. https://doi.org/10.1080/00224065.2007.11917692

Insiyah, J., Astutik, S., & Soehono, L. A. (2023). Air Quality Control Using 2 Thotelling Double Bootstrap Control Chart with Vector Autoregressive Model Approach. *Journal of Theoretical and Applied Information Technology*, *2*(101).

Iso. (2017). *General requirements for the competence of testing and calibration laboratories*. www.iso.org

Khediri, I. Ben, Weihs, C., & Limam, M. (2010). Support Vector Regression control charts for multivariate nonlinear autocorrelated processes. *Chemometrics and Intelligent Laboratory Systems*, *103*(1), 76–81. https://doi.org/10.1016/j.chemolab.2010.05.021

Khusna, H., Mashuri, M., Suhartono, S., Prastyo, D. D., & Ahsan, M. (2018). Multioutput least square SVR-based multivariate EWMA control chart: The performance evaluation and application. *Cogent Engineering*, *5*(1), 1531456. https://doi.org/10.1080/23311916.2018.1531456

Kim, Y.-H., & Moon, S.-H. (2025). Machine Learning-Based Quality Control for Low-Cost Air Quality Monitoring: A Comprehensive Review of the Past Decade. *Atmosphere*, *16*(10), 1136. https://doi.org/10.3390/atmos16101136

Le Thi Thu. (2024). Using the Cubic Spline Interpolation Method to Approximate Some Real Data. *Journal of Thu Dau Mot University*, *6*, 93–103. https://doi.org/10.37550/tdmu.ejs/2024.01.521

Lee, H. W., Lee, H. J., Oh, S., Lee, J., Heo, E. Y., & Kim, D. K. (2024). Combined Effect of Changes in NO2, O3, PM2.5, SO2 and CO Concentrations on Small Airway Dysfunction. *Respirology*, *29*(5), 379–386. https://doi.org/10.1111/resp.14687

Li, Z., Zou, C., Gong, Z., & Wang, Z. (2014). The Computation of Average Run Length and Average Time to Signal: An Overview. *Journal of Statistical Computation and Simulation*, *84*(8), 1779–1802. https://doi.org/10.1080/00949655.2013.766737

Liu, Q., Li, X., Liu, T., & Zhao, X. (2020). Spatio-Temporal Correlation Analysis of Air Quality in China: Evidence from Provincial Capitals Data. *Sustainability*, *12*(6), 2486. https://doi.org/10.3390/

su12062486

Liu, X., Dong, X., Zhang, L., Chen, J., & Wang, C. (2023). Least Squares Support Vector Regression for Complex Censored Data. *Artificial Intelligence in Medicine*, *136*, 102497. https://doi.org/10.1016/j.artmed.2023.102497

Liu, Y., & Zi, X. (2021). Air Quality Index Monitoring Based on Profile-MEWMA Control Chart. *Journal of Physics: Conference Series*, *1955*(1), 012051. https://doi.org/10.1088/1742-6596/1955/1/012051

Mecklin, C. J., & Mundfrom, D. J. (2005). A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation*, *75*(2), 93–107. https://doi.org/10.1080/0094965042000193233

Miftahul Ihsan, I., Ma'rufatin, A., Zahroh, N. F., Nurul Ikhsan, I., Suwedi, N., Pratama, R. A., Adhi, R. P., Handika, R., Lusia, A., Nishihashi, M., Terao, Y., Hashimoto, S., Nara, H., & Mukai, H. (2025). Air Quality Assessment Based on Real-Time Continuous Monitoring: Particulate and Nitrogen Dioxide Concentrations in South Tangerang. *Jurnal Teknologi Lingkungan*, *26*(1), 97–104.

Montgomery, D. C. (2012). *Introduction to Statistical Quality Control Seventh Edition*. John Wiley & Sons, Inc.

Najafzadeh, M., Noori, R., Afroozi, D., Ghiasi, B., Hosseini-Moghari, S. M., Mirchi, A., Torabi Haghighi, A., & Kløve, B. (2021). A comprehensive uncertainty analysis of model-estimated longitudinal and lateral dispersion coefficients in open channels. *Journal of Hydrology*, *603*. https://doi.org/10.1016/j.jhydrol.2021.126850

Noori, R., Karbassi, A., Ashrafi, K., Ardestani, M., Mehrdadi, N., & Bidhendi, G. R. N. (2012). Active and online prediction of BOD 5 in river systems using reduced-order support vector machine. *Environmental Earth Sciences*, *67*(1), 141–149. https://doi.org/10.1007/s12665-011-1487-9

Noori, R., Karbassi, A. R., Ashrafi, K., Ardestani, M., & Mehrdadi, N. (2013). Development and application of reduced-order neural network model based on proper orthogonal decomposition for BOD5 monitoring: Active and online prediction. *Environmental Progress and Sustainable Energy*, *32*(1), 120–127. https://doi.org/10.1002/ep.10611

Noori, R., Yeh, H. Der, Abbasi, M., Kachoosangi, F. T., & Moazami, S. (2015). Uncertainty analysis of support vector machine for online prediction of five-day biochemical oxygen demand. *Journal of Hydrology*, *527*, 833–843. https://doi.org/10.1016/j.jhydrol.2015.05.046

Núñez-Alonso, D., Pérez-Arribas, L. V., Manzoor, S., & Cáceres, J. O. (2019). Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies in the Madrid Region. *Journal of Analytical Methods in Chemistry*, *2019*, 1–9. https://doi.org/10.1155/2019/9753927

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. In *Journal of Statistical Modeling and Analytics* (Vol. 2).

Reynolds, M. R. (1996). Shewhart and EWMA Variable Sampling Interval Control Charts with Sampling at Fixed Times. *Journal of Quality Technology*, *28*(2), 199–212. https://doi.org/10.1080/00224065.1996.11979660

Sato, J. R., Costafreda, S., Morettin, P. A., & Brammer, M. J. (2008). Measuring Time Series Predictability Using Support Vector Regression. *Communications in Statistics - Simulation and Computation*, *37*(6), 1183–1197. https://doi.org/10.1080/03610910801942422

Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *14*(4), 1502. https://doi.org/10.12928/telkomnika.v14i4.3956

Tran, N. K., Kühle, L. C., & Klau, G. W. (2024). A Critical Review of Multi-output Support Vector Regression. *Pattern Recognition Letters*, *178*, 69–75. https://doi.org/10.1016/j.patrec.2023.12.007

Xie, X., & Qiu, P. (2023). Control Charts for Dynamic Process Monitoring with an Application to Air Pollution Surveillance. *The Annals of Applied Statistics*, *17*(1). https://doi.org/10.1214/22-AOAS1615

Xu, K., Cui, K., Young, L. H., Hsieh, Y. K., Wang, Y. F., Zhang, J., & Wan, S. (2020). Impact of the COVID-19 Event on Air Quality in Central China. *Aerosol and Air Quality Research*, *20*(5), 915–929. https://doi.org/10.4209/aaqr.2020.04.0150

Xu, S., An, X., Qiao, X., Zhu, L., & Li, L. (2013). Multi-output Least-squares Support Vector Regression Machines. *Pattern Recognition Letters*, *34*(9), 1078–1084. https://doi.org/10.1016/j.patrec.2013.01.015

Xu, Z., Xiong, L., Jin, D., & Tan, J. (2021). Association Between Short-term Exposure to Sulfur Dioxide and Carbon Monoxide and Ischemic Heart Disease and Non-accidental Death in Changsha City, China. *PLOS One*, *16*(5), e0251108. https://doi.org/10.1371/journal.pone.0251108

Zhang, L., Lai, C. D., Govindaraju, K., & Bebbington, M. (2004). A Note on Average Run Lengths of Moving Average Control Charts. *Economic Quality Control*, *19*(1). https://doi.org/10.1515/EQC.2004.23

Zhou, J., Wang, Z., Li, C., Qiu, Y., & Tao, M. (2023). *Metaheuristic-hybridized MLS-SVR: a Multi-output Model for Predicting Continuous Shear Displacements of Rock Fractures*. https://doi.org/10.21203/rs.3.rs-3450476/v1