

Air Quality Prediction Using Support Vector Regression

A Machine Learning–Based Sensor Calibration Study: By Santhosh Kumar K

Project Type

Self-Initiated Machine Learning Project

Objective of the Project

To evaluate the effectiveness and limitations of Support Vector Regression in predicting ground truth air pollutant concentrations using low-cost sensor data and environmental variables.

Domain

Data Science | Machine Learning |
Environmental Analytics

Dataset Source

Open-source Air Quality Monitoring
Dataset (UCI Repository)

Tools & Technologies Used

- Python
- Scikit-learn
- Pandas, NumPy
- Power BI
- Jupyter Notebook

Prepared For: Professional Portfolio

INTRODUCTION

Air quality monitoring plays a critical role in environmental management, public health protection, and urban planning. Accurate measurement of air pollutants such as carbon monoxide, nitrogen oxides, nitrogen dioxide, and benzene is essential for assessing environmental risks and ensuring regulatory compliance.

Traditionally, air quality monitoring relies on reference-grade analytical instruments that provide highly accurate measurements but are expensive to install, operate, and maintain. As a result, their deployment is limited to a small number of monitoring stations.

In recent years, low-cost air quality sensors have emerged as an alternative for large-scale and continuous monitoring. While these sensors are affordable and capable of real-time data collection, they produce indirect and noisy measurements that cannot directly represent true pollutant concentrations. This limitation has created a need for intelligent data-driven approaches that can calibrate sensor outputs and improve their reliability.

Machine learning techniques, particularly non-linear regression models, offer promising solutions for mapping sensor signals to accurate pollutant concentrations.

IDENTIFIED PROBLEM

Despite the availability of low-cost air quality sensors, their practical use is limited due to inaccurate and unstable readings. The key challenges identified in this study are:

- Sensor signals do not directly represent pollutant concentrations
- High levels of noise and skewness in sensor measurements
- Environmental factors such as temperature and humidity influence sensor behavior
- Some pollutants exhibit complex chemical behavior that is difficult to model

These challenges raise the question of whether machine learning models, specifically Support Vector Regression, can reliably estimate ground truth pollutant concentrations using indirect sensor data under real-world conditions.

OBJECTIVES OF THE STUDY

Primary Objective:

To evaluate the effectiveness of Support Vector Regression in predicting ground truth air pollutant concentrations using low-cost sensor signals and environmental variables.

Secondary Objectives:

- To analyze the variability and distribution of air pollutants and sensor signals
- To compare the performance of linear and non-linear SVR models
- To assess differences in predictability across multiple pollutants
- To examine the impact of selective hyperparameter tuning on model performance
- To identify practical limitations of machine learning in sensor-based air quality monitoring

DATA COLLECTION METHODS

The dataset used in this study was obtained from an open-source (Kaggle) air quality monitoring repository. Data was originally collected from a real-world air quality monitoring station equipped with both reference-grade analyzers and low-cost sensor arrays. Measurements were recorded at regular time intervals and include pollutant concentrations, sensor responses, and environmental conditions.

Reference-grade instruments were used to obtain ground truth pollutant values, while metal oxide sensors captured indirect electrical responses to gas presence. Environmental variables such as temperature and humidity were also recorded to account for their influence on sensor behavior.

Data Analysis Tools:

The following tools and technologies were used in this study:

- Python for data processing and machine learning
- Pandas and NumPy for data manipulation and cleaning
- Scikit-learn for implementing Support Vector Regression models
- Matplotlib and Seaborn for exploratory data analysis
- Power BI for interactive data visualization and insight communication

These tools enabled efficient data handling, model training, evaluation, and visualization.

Data Collected:

The dataset consists of multiple air quality parameters collected over an extended period. The key components include:

- Ground truth pollutant concentrations: CO(GT), NO_x(GT), NO₂(GT), and C₆H₆(GT)
- Sensor signals from metal oxide gas sensors (PT08.S1 to PT08.S5)
- Environmental variables: temperature, relative humidity, and absolute humidity

Certain variables, such as NMHC(GT), were excluded due to extensive missing values and unreliable measurements. Only validated and scientifically meaningful variables were retained for analysis.

Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NO _x (GT)	PT08.S3(NO _x)	NO ₂ (GT)	PT08.S4(NO ₂)	PT08.S5(O ₃)	T	RH	AH
3/10/2004	18:00:00	2.6	1360.0	150.0	11.9	1046.0	166.0	1056.0	113.0	1692.0	1268.0	13.6	48.9	0.7578
3/10/2004	19:00:00	2.0	1292.0	112.0	9.4	955.0	103.0	1174.0	92.0	1559.0	972.0	13.3	47.7	0.7255
3/10/2004	20:00:00	2.2	1402.0	88.0	9.0	939.0	131.0	1140.0	114.0	1555.0	1074.0	11.9	54.0	0.7502
3/10/2004	21:00:00	2.2	1376.0	80.0	9.2	948.0	172.0	1092.0	122.0	1584.0	1203.0	11.0	60.0	0.7867
3/10/2004	22:00:00	1.6	1272.0	51.0	6.5	836.0	131.0	1205.0	116.0	1490.0	1110.0	11.2	59.6	0.7888

DATA ANALYSIS AND MACHINE LEARNING METHODOLOGY

Data Preprocessing:

Data preprocessing involved removing irrelevant attributes, handling invalid sensor readings, and treating missing values using statistically justified imputation techniques. Median imputation was applied to skewed sensor signals, while environmental variables were treated using mean or median values based on distribution characteristics. The cleaned dataset was then standardized for machine learning applications.

Machine Learning Model Selection:

Support Vector Regression was selected as the primary modeling technique due to its robustness to noise and ability to capture non-linear relationships. Both linear and RBF-kernel SVR models were trained using the same feature set to ensure fair comparison. Feature scaling was applied prior to model training to support distance-based learning.

Model Training and Evaluation:

Each pollutant was treated as an independent regression task. Models were evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. Baseline comparisons revealed consistent performance improvements when using RBF-kernel SVR, confirming the non-linear nature of sensor–pollutant relationships.

Model Target	MAE		RMSE		R2	
	Linear SVR	RBF SVR	Linear SVR	RBF SVR	Linear SVR	RBF SVR
C6H6(GT)	0.806353	0.050405	1.404246	0.115914	0.964483	0.999758
CO(GT)	0.451716	0.390902	0.751934	0.682580	0.680271	0.736531
NO2(GT)	20.398265	16.297931	27.681287	23.517933	0.609915	0.718431
NOx(GT)	71.425893	51.918043	114.090379	88.942232	0.654247	0.789872

Hyperparameter Tuning Strategy:

Hyperparameter tuning was selectively applied to pollutants exhibiting high prediction error, particularly nitrogen oxides (NO_x). Instead of exhaustive tuning, a focused approach was adopted to demonstrate tuning capability while minimizing overfitting risks. The regularization parameter (C) was adjusted to balance model complexity and noise tolerance.

The tuning process resulted in modest reductions in error metrics but did not eliminate prediction uncertainty, indicating that sensor noise and chemical variability impose inherent limits on achievable accuracy. This outcome reinforces the importance of realistic expectations when applying machine learning to noisy environmental data.

Findings:

- RBF-based SVR consistently outperformed linear SVR models
- Benzene showed near-perfect predictability due to strong sensor alignment
- CO and NO₂ demonstrated moderate prediction accuracy
- NO_x remained the most challenging pollutant due to high variability
- Hyperparameter tuning improved robustness but had limited impact under noisy conditions
- Data quality and sensor sensitivity were more influential than model complexity

DATA ANALYSIS:

Data analysis was conducted in multiple stages. Initially, the dataset was cleaned by removing irrelevant columns and treating invalid sensor readings as missing values. Distribution-based analysis was used to guide appropriate imputation strategies, ensuring minimal distortion of real-world behavior. Exploratory analysis revealed significant differences in pollutant variability and sensor signal distributions.

Machine learning analysis involved training Support Vector Regression models using both linear and radial basis function (RBF) kernels. Feature scaling was applied to ensure effective distance-based learning. Model performance was evaluated using standard regression metrics such as Mean Absolute Error, Root Mean Squared Error, and R^2 score. Baseline models were compared across multiple pollutants to assess differences in predictability. Selective hyperparameter tuning was applied to high-variance targets to evaluate potential performance improvements.





CONCLUSION:

This study confirms that Support Vector Regression is a suitable and effective technique for calibrating low-cost air quality sensors when strong sensor–pollutant relationships exist.

Non-linear SVR models significantly improve prediction accuracy compared to linear approaches; however, their performance is constrained by sensor noise and pollutant complexity.

The findings highlight that machine learning should complement, rather than replace, reliable measurement systems and that careful data preparation and informed model selection are essential for real-world deployment.