# SARIMAX + LSTM for Tainan PM2.5 Prediction

Machine Learning Final Report

github: https://github.com/HsienLee/SARIMAX-LSTM-for-Tainan-PM2.5-Prediction/tree/main

**Ping-Hsien Lee**

R26124013

January 11, 2025

## 1 Abstract

This study presents a hybrid modeling approach combining Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) and Long Short-Term Memory (LSTM) neural networks for the prediction of PM2.5 air pollution levels. The proposed model aims to leverage the strengths of both statistical and deep learning methods to improve forecasting accuracy.

In the first stage, a SARIMAX model is developed to capture the linear and seasonal patterns inherent in the PM2.5 time series data. Autocorrelation and partial autocorrelation analyses are conducted to determine the optimal parameters for the SARIMAX model. The model is then fitted to the historical data to generate initial forecasts.

The residuals obtained from the SARIMAX model, representing the nonlinear patterns not captured in the first stage, are used as input for the LSTM neural network in the second stage. The LSTM model is trained on these residuals to learn complex nonlinear dependencies and temporal dynamics within the data.

Finally, the forecasts from the SARIMAX model are combined with the predictions of the LSTM model to produce the final PM2.5 concentration forecasts. This integration effectively adjusts the initial SARIMAX forecasts by incorporating the nonlinear corrections learned by the LSTM network. Experimental results demonstrate that the SARIMAX-LSTM hybrid model outperforms traditional single-model approaches in predicting PM2.5 levels, achieving higher accuracy and better capturing of both linear and nonlinear patterns in the data. The enhanced forecasting capability of this hybrid model provides valuable insights for air quality management and can assist policymakers in making informed decisions to mitigate air pollution impacts.

***Keywords***: SARIMAX, LSTM, PM2.5, Air Pollution, Hybrid Model, Residual Analysis, Deep Learning

## 2 Intoduction

Air pollution, particularly PM2.5 (particulate matter with a diameter of 2.5 micrometers or less), has become a critical environmental issue significantly impacting public health and quality of life. In Taiwan, the air pollution situation is especially severe in the Tainan region. According to the 2023 IQAir report, Tainan ranks as the second-worst city in Taiwan for air quality, with approximately 66 presents of PM2.5 emissions originating from domestic power plants and industrial manufacturing. Tainan's unique geographical location, surrounded by towering mountains, combined with the winter northeast monsoon, prevents pollutants from dispersing easily, further deteriorating air quality. These pollutants include not only PM2.5 and other particulate matter but also carbon dioxide, nitrogen oxides, sulfur oxides, and industrial heavy metal emissions, posing serious health threats to local residents, particularly vulnerable groups such as children, the elderly, and pregnant women.

Accurate prediction of PM2.5 concentrations is crucial for implementing effective air quality management strategies and providing timely public warnings. However, PM2.5 concentration patterns exhibit complex characteristics, including linear and non-linear components, seasonal variations,

and multiple influencing factors, making accurate prediction particularly challenging. In Tainan's case, this predictive complexity is further intensified by its unique geographical environment and industrial distribution.

Traditional time series prediction methods, such as Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX), have proven effective in capturing linear relationships and seasonal patterns in PM2.5 data. These statistical models provide interpretable results and handle seasonal variations well. However, they may fall short in capturing the inherent complex non-linear relationships in air pollution dynamics.

On the other hand, deep learning methods, particularly Long Short-Term Memory (LSTM) networks, have demonstrated excellent capability in learning non-linear patterns and long-term dependencies in time series data. LSTM's sophisticated architecture allows it to retain important long-term information while filtering out irrelevant details. Nevertheless, LSTM models may not be as effective as traditional statistical methods in capturing explicit seasonal patterns.

To address these limitations, this study proposes a hybrid modeling approach that combines the advantages of both SARIMAX and LSTM models. The framework first uses SARIMAX to model the linear and seasonal components of PM2.5 concentrations. The residuals from the initial modeling phase, representing non-linear patterns not captured by SARIMAX, are subsequently modeled using LSTM networks. This two-stage approach better captures both linear and non-linear components of PM2.5 time series.

This study makes several contributions to the field of air quality prediction: 1. Development of an innovative hybrid method combining statistical and deep learning approaches 2. Improved prediction accuracy through complementary modeling of linear and non-linear patterns 3. Enhanced understanding of PM2.5 concentration dynamics in the Tainan region through comprehensive model analysis 4. Practical application value for air quality management and public health protection in the Tainan area

Subsequent chapters will detail the methodology, experimental setup, results, and implications of this hybrid modeling approach for PM2.5 prediction in the Tainan region. This research is not only significant for air quality management in Tainan but also provides valuable reference for cities with similar geographical environments and industrial structures.

## 3 Material and methods

### 3.1 Data collection and preprocessing

This study analyzes air quality monitoring data from Taiwan Environmental Protection Administration's central monitoring station in Tainan City, covering the observation period from January 1, 2021, to December 31, 2023. Data is updated at a high frequency of once per hour, and this long-term, high-frequency dataset provides a sufficient sample foundation for establishing reliable prediction models. The monitoring data includes seven major air pollutant indicators: PM10 (suspended particulates), CH4 (methane), O3 (ozone), NO (nitrogen monoxide), SO2 (sulfur dioxide), CO (carbon monoxide), and NO2 (nitrogen dioxide). Through Variance Inflation Factor (VIF) analysis, results show that CO and NO2 have notably high VIF values (>10), indicating potential strong multicollinearity between these variables, while PM10, O3, SO2, and NO have lower VIF values, demonstrating better independence. Consequently, CO and NO2 were removed from subsequent analyses.

To ensure data quality, this study employed Forward Filling method during the data preprocessing stage to handle missing values, maintaining temporal continuity of the data and avoiding the introduction of additional data bias. Additionally, scale standardization was performed on all features to eliminate the effects of different measurement scales and improve model stability and comparability.

### 3.2 Model design

This research proposes two innovative hybrid model architectures for predicting PM2.5 concentrations. The first architecture (SARIMAX $\rightarrow$ LSTM) begins by determining the key parameters of the SARIMAX model through ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) analysis, including autoregressive order (p), differencing order (d), and moving average order (q), to generate preliminary predictions. Then, the residuals from the

SARIMAX model are converted into time-step sequence format and input into the LSTM network for training and prediction. Finally, the predictions from both models are integrated to obtain the final predicted values. The second architecture (LSTM → SARIMAX) reverses the processing order, first using LSTM for primary prediction, then applying SARIMAX to model the residuals, and finally integrating both results. The main difference between these two architectures is that Model One, with SARIMAX as the primary predictor, is more suitable for data with strong seasonal and temporal patterns, while Model Two, led by LSTM, is better suited for handling complex nonlinear patterns. This hybrid approach fully utilizes SARIMAX's advantages in capturing linear relationships and seasonality, along with LSTM's strengths in processing complex nonlinear patterns, aiming to provide more accurate PM2.5 concentration predictions by combining the characteristics of deep learning and statistical modeling.

## 4 Results analysis

### 4.1 SARIMAX model

To determine the optimal parameter settings for the SARIMAX model, this study analyzed the characteristics of ACF and PACF plots. As shown in Figure 1, the ACF plot exhibits a slowly decaying pattern, leading to the selection of a moving average order of q=1. The PACF plot shows significant spikes exceeding the confidence intervals at the first two lags (highlighted in yellow boxes), which informed the determination of the autoregressive order p=2. In the prediction results shown on the right, the blue line represents the actual PM2.5 concentration values, while the orange line indicates the SARIMAX model predictions, spanning time points 21000 to 26000, with PM2.5 concentration values ranging approximately from 0 to 0.5. The model parameters were primarily established based on the following criteria: p=2 was derived from the significance of the first two lags in the PACF plot, q=1 was based on the gradual decay pattern in the ACF plot, and a 24-hour periodicity was incorporated as the seasonal parameter.

As shown in Figure 2, the overall prediction results demonstrate that the SARIMAX model effectively captures

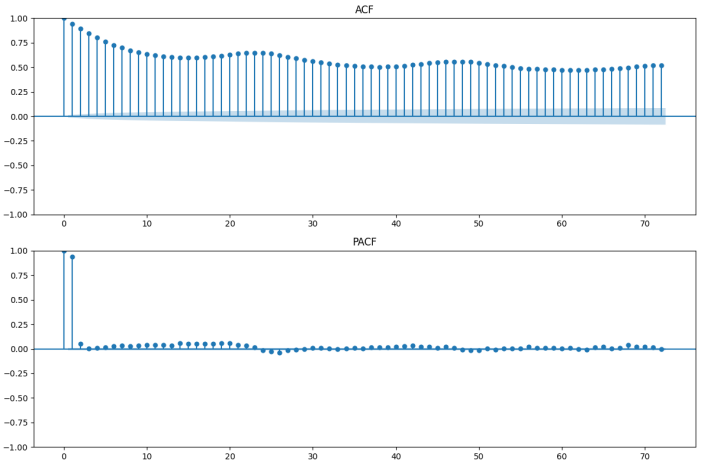


**Figure 1:** ACF and PACF for SARIMAX model of Tainan PM2.5

the trends in PM2.5 concentration variations, although some discrepancies exist in predicting extreme values. This limitation underscores the necessity of incorporating LSTM models in the comprehensive study to further enhance prediction accuracy.

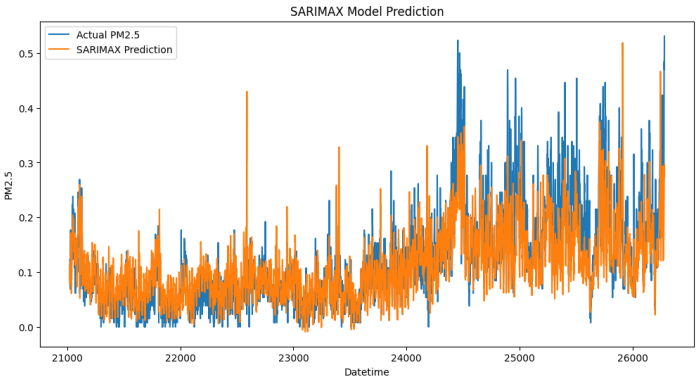


**Figure 2:** SARIMAX model for Tainan PM2.5 prediction

### 4.2 LSTM model

Based on Figure 3, analysis of the LSTM model's prediction results reveals that due to the training dataset (from 2021 to mid-2023) encompassing the COVID-19 pandemic period, during which reduced human activities led to relatively lower PM2.5 emissions, while the test set (late 2023) was no longer affected by the pandemic, the model's predicted values were generally lower than the actual observed values. Specifically, the model demonstrated high prediction accuracy in low concentration ranges (0.0-0.2) and effectively captured medium-scale concentration fluctuations,

but exhibited significant errors in predicting peak values (concentrations around 0.5). From a model characteristics perspective, the prediction curve displays smooth features, reflecting the LSTM model's strong generalization capability, though it adopts a conservative prediction strategy when handling extreme values, tending to output predictions within the medium range. These performance characteristics indicate that while the LSTM model excels in capturing overall PM2.5 concentration trends, there remains room for further optimization in predicting extreme values.

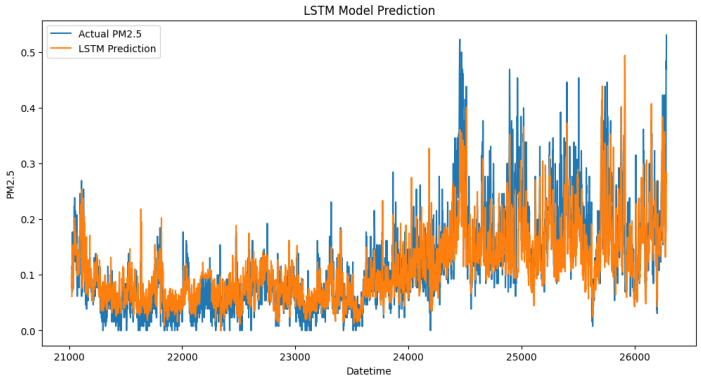


**Figure 3:** SARIMAX model for Tainan PM2.5 prediction

### 4.3 Hybrid Model

Figures 4 and 5 present a comparative analysis of two hybrid modeling approaches, each demonstrating distinct advantages and limitations in PM2.5 concentration prediction. In Model1 (SARIMAX → LSTM configuration), the actual PM2.5 concentrations are represented by the purple line, while the LSTM residual predictions and hybrid model outputs are depicted by green and yellow lines, respectively. Analysis reveals that the LSTM exhibits conservative behavior in residual processing, generating relatively stable predictions—a characteristic potentially advantageous in scenarios demanding reliable forecasting. Notably, the y-axis constraint within the 0-0.5 range indicates superior prediction volatility control, although this conservative approach may compromise sensitivity to extreme values.

Conversely, Model2 (LSTM → SARIMAX configuration) employs an inverse methodology, where the dark blue line represents actual concentrations, the orange line indicates SARIMAX residual predictions, and the light blue line shows the hybrid model outputs. This model demonstrates an expanded prediction range (0-0.8) with SARIMAX residual predictions exhibiting pronounced volatility. The model displays enhanced capability in capturing peak values, potentially more suitable for extreme pollution event early warning applications. Temporal analysis indicates that both models detect a significant PM2.5 concentration elevation post-timestamp 24000, with Model2 showing more pronounced responsiveness.

The selection of an appropriate model should be predicated on specific application requirements: Model1 may be preferable when prediction stability and reliability are paramount, while Model2 might be more suitable for applications requiring high sensitivity to extreme pollution events, despite potential false alarms. This trade-off exemplifies the classical stability-sensitivity compromise in environmental monitoring systems. Furthermore, we propose the potential integration of both models through cross-validation for more comprehensive predictive reference. Future research directions could explore dynamic weight adjustment methodologies for these models to adapt to varying pollution scenarios.

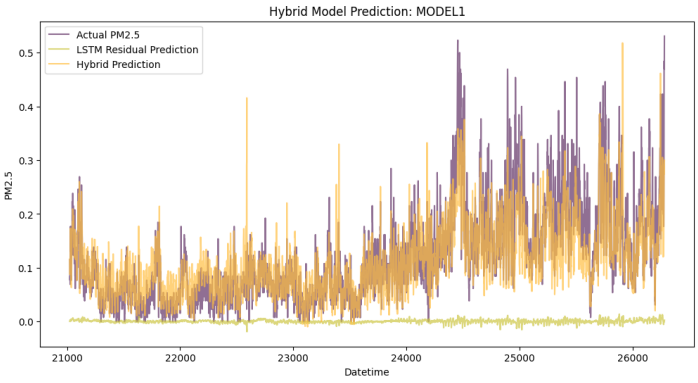


**Figure 4:** SARIMAX model for Tainan PM2.5 prediction

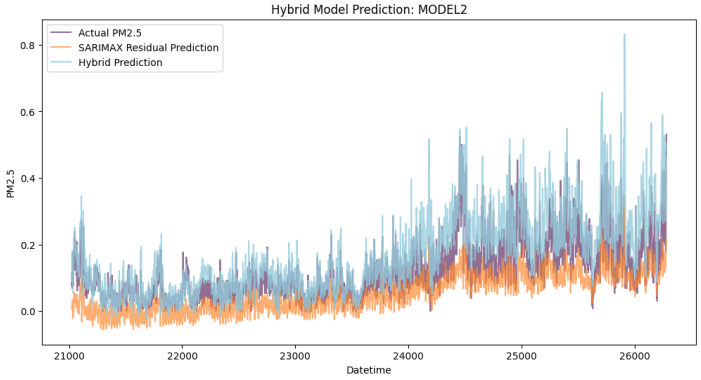


**Figure 5:** SARIMAX model for Tainan PM2.5 prediction

### 4.4 Model evaluation

We using three standard metrics—Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination ($R^2$ Score)—to evaluate model predictive performance. MSE reflects overall prediction accuracy by calculating the average of squared differences between predicted and actual values; RMSE, as the square root of MSE, provides error measurements at the same scale as the original data, enhancing the interpretability of evaluation results; $R^2$ Score indicates the model's capacity to explain data variance, with values approaching 1 representing superior model fit.

According to the analysis results in Table 1, Model1 (SARIMAX + LSTM) demonstrates excellent performance across all three evaluation metrics: achieving an MSE of 0.002265, RMSE of 0.047593, and $R^2$ Score of 0.697447. These metrics represent the optimal results among all tested models, particularly noteworthy is the $R^2$ Score approaching 0.7, indicating the model's ability to explain nearly 70 presents of data variance—a significant achievement in the complex domain of air pollution prediction. However, it is worth noting that while Model1 outperforms the standalone LSTM model, the advantage is not substantial, possibly attributable to relatively stable data trends and low volatility during the study period. In contrast, the LSTM→SARIMAX combination sequence (Model2) results in significantly degraded predictive performance, strongly indicating the critical impact of model combination ordering on prediction effectiveness.

Based on these findings, it is recommended to implement Model1's configuration in practical applications, specifically utilizing SARIMAX to capture linear and seasonal characteristics followed by LSTM processing of residual non-linear patterns, thereby fully leveraging the advantages of both models to achieve optimal predictive performance.

| **Model** | **MSE** | **RMSE** | **$R^2$ Score** |
|---|---|---|---|
| SARIMAX | 0.002296 | 0.047915 | 0.693340 |
| LSTM | 0.002267 | 0.047608 | 0.697256 |
| **Hybrid Models** | | | |
| MODEL1: SARIMAX + LSTM | **0.002265** | **0.047593** | **0.697447** |
| MODEL2: LSTM + SARIMAX | 0.005128 | 0.071610 | 0.315057 |

**Table 1:** Performance Comparison of Different Models

## 5 Conclusion

This study developed a hybrid model combining SARIMAX and LSTM for predicting PM2.5 concentrations in the Tainan region. The results demonstrate that the SARIMAX→LSTM sequential hybrid model (Model1) achieved superior performance among all tested models, not only obtaining optimal results across all evaluation metrics (MSE: 0.002265, RMSE: 0.047593, $R^2$ Score: 0.697447) but also effectively capturing both linear and non-linear characteristics of PM2.5 concentration variations.

The research reveals that the sequence of model combination has a decisive impact on prediction effectiveness. The approach using SARIMAX as the primary predictor followed by LSTM for residual processing (Model1) significantly outperforms the reverse configuration (Model2), confirming the superior efficacy of first capturing linear and seasonal features before addressing non-linear patterns. Notably, due to the inclusion of COVID-19 pandemic period data in the training dataset, the model exhibited underestimation when predicting PM2.5 concentrations in late 2023 (post-pandemic period), highlighting the significant influence of environmental factors on prediction accuracy.

Hybrid model not only provides a reliable prediction tool for air quality management in the Tainan region but also offers valuable references for policy makers and public health protection measures. Looking forward, the model's appli-

cation could be extended to more volatile domains, such as cryptocurrency and stock market predictions, to further validate the advantages of hybrid models in handling complex time series data. Additionally, there is potential to explore optimization of the model's capability in predicting extreme events and its adaptive applications under various environmental conditions. The methodology of this study is equally applicable to other cities with similar geographical environments and industrial structures.

## References


[1] Wu, J., Zhang, X., Huang, F., Zhou, H., & Chandra, R. (2024). Review of deep learning models for crypto price prediction: implementation and evaluation.

[2] Taiwan Air Quality Monitoring Network

[3] Taiwan Air Quality Monitoring Network Historical monitoring data.