Short-term Forecasting of PM2.5 via Transformer Encoder with Sparse Attention

7113095001 Kao Che-Kai
Course: Functional Data Analysis

January 6, 2026

1 Abstract

Motivated by the effectiveness of DLinear in decomposing time series components, this study extends the idea toward multivariate dependency modeling by adopting the iTransformer architecture for PM2.5 forecasting. The model explicitly captures cross-variable interactions among multiple environmental factors, and further replaces the conventional Softmax attention with Entmax to enable sparse and adaptive attention distributions that suppress irrelevant information. Experimental results on real-world data demonstrate that the proposed approach achieves an MSE of 42.48, confirming its effectiveness in handling complex multivariate air quality forecasting tasks.

2 Introduction

2.1 Background and Public Health

With rapid industrialization and rising living standards, fine particulate matter ($PM_{2.5}$) has emerged as a critical public health concern. Due to its minute particle size and high permeability, $PM_{2.5}$ can transport toxic substances into the respiratory and circulatory systems, significantly increasing the risk of cardiovascular and respiratory diseases. Accurate air quality forecasting is therefore essential for public health warnings and the formulation of emission reduction strategies.

2.2 Challenges in Taiwan's Context

Predicting $PM_{2.5}$ concentrations in Taiwan is exceptionally challenging due to two primary factors:

- **Complex Topography:** The Central Mountain Range acts as a physical barrier to airflow, causing pollutants to accumulate on the leeward side (e.g., Central and Southern Taiwan), resulting in localized high-concentration episodes.
- **Seasonal Meteorological Variability:** Long-range transport via the Northeast Monsoon in winter and spring elevates baseline concentrations, while low boundary layers

and temperature inversions in autumn/winter hinder vertical dispersion, introducing high non-linearity into temporal forecasting.

2.3 Limitations of Current Models

While traditional chemical transport models (e.g., CMAQ) are physically grounded, they are computationally intensive and sensitive to initial conditions. Although deep learning models like LSTM and Transformers have advanced data-driven forecasting, $PM_{2.5}$ data exhibits strong **inter-variable coupling** (e.g., humidity, wind, and temperature) and **spatial heterogeneity**. Efficiently identifying key predictors while filtering out environmental noise remains a pivotal challenge in data science.

3 Structure Overview

3.1 Problem Definition

In this study, short-term $PM_{2.5}$ concentration forecasting is formulated as a **Multivariate Time Series Forecasting (MTSF)** problem. Given a historical observation sequence $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times N}$, where T denotes the lookback window length and N represents the total number of variables—including $PM_{2.5}$ concentrations, meteorological factors (e.g., wind speed, temperature, humidity), and other relevant air pollutants.

The objective is to learn a mapping function f that accurately predicts the values for the subsequent S time steps, denoted as $\mathbf{Y} = \{\mathbf{x}_{T+1}, \ldots, \mathbf{x}_{T+S}\} \in \mathbb{R}^{S \times N}$:

$$\mathbf{Y} = f(\mathbf{X}; \Theta) \tag{1}$$

where Θ denotes the set of learnable model parameters.

3.2 Core Architecture: Inverted Transformer (iTransformer)

Traditional Transformer models typically perform attention operations across the *temporal dimension* (time steps). However, in multivariate meteorological forecasting, the correlations between different variables (e.g., the influence of wind direction on pollutant dispersion) are often more critical than the temporal evolution of a single variable. Consequently, this study adopts the **iTransformer** architecture, characterized by the following core logic:

3.2.1 Variate Inversion (Tokenization)

Unlike conventional methods that treat each time step as a token, we embed the **entire historical sequence** of a single variable as an independent token. This inversion allows the attention layer to directly compute the interaction between distinct physical features (e.g., $PM_{2.5}$ concentration vs. humidity), capturing the inherent coupling of multivariate meteorological data.

3.2.2 Inter-variate Correlation with Sparse Attention

The model utilizes a multi-head attention mechanism to capture the complex, non-linear dependencies among N variables. To address the noise generated by Taiwan's complex topography, we introduce **Entmax sparse attention**. This mechanism performs active feature selection by assigning strictly zero weights to redundant or irrelevant variables, ensuring that only predictors with strong physical correlations to $PM_{2.5}$ levels are retained.

3.2.3 Temporal Dependency Modeling

Within each variable token, the model employs a shared **Multilayer Perceptron (MLP)** or Feed-Forward Network (FFN) to learn the non-linear temporal evolution. This decoupled design ensures a dual modeling capability: the attention mechanism handles the "horizontal" inter-variate relationships, while the MLP precisely captures the "vertical" longitudinal variations within individual time series.

4 METHOD

Given Taiwan's complex topography, air quality is significantly influenced by microclimates and geographic barriers. To address these challenges, this research implements a **site-specific modeling** strategy, incorporating observations from **spatial neighbors** as exogenous variables. To enable the model to identify the physical characteristics of pollutant transport via seasonal monsoons, we explicitly introduce **Spatial-Geometric Embeddings** into the architecture.

4.1 Spatial-Geometric Encoding

To capture the spatial topology relative to the target station, we define a geometric feature vector for each neighboring station $n \in \{1, \ldots, N\}$. Taking the target station as the origin $(0, 0)$, the spatial relationship is encoded as:

$$\mathbf{p}_n = [d_n, \sin(\theta_n), \cos(\theta_n)] \tag{2}$$

where d_n denotes the normalized Euclidean distance and θ_n represents the relative azimuth angle. For the target station itself, this vector is zero-padded. This encoding allows the attention mechanism to remain sensitive to the directionality and decay effects of pollutant dispersion, which are highly correlated with seasonal wind patterns in Taiwan.

4.2 Variate Tokenization and Spatial Augmentation

In contrast to temporal tokenization, each variable's complete historical sequence $\mathbf{x}_{1:T,n}$ is projected into a high-dimensional latent space. This temporal representation is then fused with its

corresponding spatial-geometric information:

$$\mathbf{e}_n = \text{Linear}(\mathbf{x}_{1:T,n}), \quad \mathbf{e}_n \in \mathbb{R}^d \tag{3}$$

$$\mathbf{s}_n = \text{MLP}_{\text{spatial}}(\mathbf{p}_n), \quad \mathbf{s}_n \in \mathbb{R}^d \tag{4}$$

$$\mathbf{h}_n = \mathbf{e}_n + \mathbf{s}_n \tag{5}$$

where $\mathbf{h}_n$ represents the finalized variable token served as input to the Transformer layers. This step ensures that each token possesses *spatial-physical awareness*—integrating its relative geographic position with its temporal evolution—prior to the calculation of inter-variable correlations.

4.3 Inter-variate Interaction with Sparse Attention

A multi-head attention mechanism is employed to capture the complex coupling between the target station, neighboring stations, and various meteorological factors. The interaction is formulated as:

$$\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N] \in \mathbb{R}^{N \times d} \tag{6}$$

$$\mathbf{A} = \text{entmax}_{1.5}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right), \quad \mathbf{Q}, \mathbf{K}, \mathbf{V} = \text{Linear}(\mathbf{H}) \tag{7}$$

By leveraging **entmax-1.5**, the model dynamically identifies "critical upwind stations" and key meteorological trends that contribute most significantly to the forecast. Unlike the traditional softmax, which yields a dense distribution, the entmax mapping facilitates *active variable selection* by assigning strictly zero weights to irrelevant or noisy variables. This ensures that the aggregated features are derived only from the most physically relevant predictors.

4.4 Variable-wise Feed-Forward Network

Following the inter-variate interaction, a shared Feed-Forward Network (FFN), implemented as a Multilayer Perceptron (MLP), is applied to each token to extract deep features along the temporal dimension:

$$\mathbf{z}_n = \text{LayerNorm}(\mathbf{h}_n + \text{MLP}(\mathbf{h}_n)) \tag{8}$$

This operation processes the historical evolution of each variable independently within the latent space, refining the representations before the final projection.

4.5 Generative Projection

The final stage of the architecture involves mapping the refined latent features back to the target forecasting horizon S. To enhance the model's capacity for capturing complex, non-linear dependencies in the output space, we employ a **Multilayer Perceptron (MLP)** as the generative projection head, rather than a simple linear layer. The predicted $PM_{2.5}$ concentrations for the target station are generated as follows:

$$\hat{\mathbf{y}}_{\text{target}} = \text{MLP}_{\text{out}}(\mathbf{z}_{\text{target}}), \quad \hat{\mathbf{y}}_{\text{target}} \in \mathbb{R}^S \tag{9}$$

where $\hat{\mathbf{y}}_{\text{target}}$ represents the predicted sequence for the future S time steps. By utilizing a non-linear projection head, the model more effectively decodes the aggregated information—comprising both the station's historical evolution and the sparsely selected spatial features—into accurate future concentration trajectories.

4.6 Loss Function

The model parameters Θ are optimized by minimizing the Mean Squared Error (MSE) between the predicted and ground-truth $PM_{2.5}$ concentrations. The objective function is defined as:

$$\mathcal{L}(\Theta) = \frac{1}{S}\sum_{s=1}^{S}(y_{T+s} - \hat{y}_{T+s})^2 \tag{10}$$

where y_{T+s} and $\hat{y}_{T+s}$ denote the actual and predicted values at time step $T+s$, respectively. We employ this loss function to prioritize the reduction of large prediction deviations, thereby ensuring stable forecasting performance across the 24-hour horizon.

5 Experimental Results and Analysis

5.1 Evaluation Metrics

To evaluate the forecasting performance for the 24-hour $PM_{2.5}$ concentration horizon, we employ three standard metrics: Mean Squared Error (**MSE**), Mean Absolute Error (**MAE**), and the Coefficient of Determination (R^2).

5.2 Performance Comparison

The experimental results for various model configurations are summarized in Table 1.

Table 1: Performance Comparison for 24-hour $PM_{2.5}$ Forecasting

Configuration	**MSE** ↓	**MAE** ↓	R^2 ↑
Baseline (iTransformer)	44.5304	4.7415	0.4447
+ Spatial Neighbors & Embeddings	42.4625	4.6362	0.4676
+ Entmax Sparse Attention (Proposed)	**42.2954**	**4.6079**	**0.4711**

5.3 Result Analysis

The empirical results demonstrate a progressive improvement in forecasting accuracy with the integration of spatial information and sparse attention mechanisms:

1. **Impact of Spatial Information:** Incorporating neighboring stations alongside spatial-geometric embeddings led to a significant reduction in MSE (from 44.53 to 42.46). This indicates that explicit spatial awareness helps the model capture the geographical dispersion patterns of $PM_{2.5}$, which are often overlooked by single-station models.

2. **Efficacy of Sparse Attention:** The proposed *iTransformer-Entmax* achieved the best performance across all metrics ($R^2 = 0.4711$). By utilizing the Entmax mechanism, the model effectively filters out environmental noise from irrelevant stations, focusing its attention budget on the most influential upwind predictors.

3. **Non-linear Projection:** The transition from a linear layer to an MLP in the generative projection phase further stabilized the 24-hour predictions, allowing for a better fit of the complex, non-linear concentration trajectories common in Taiwan's air quality data.

6 Conclusion

In this study, we successfully developed the **iTransformer-Entmax** framework, specifically tailored for short-term $PM_{2.5}$ forecasting within the context of Taiwan's complex topographical and meteorological conditions. Our experimental results demonstrate that the integration of **Spatial-Geometric Embeddings** with a **differentiable sparse attention mechanism** significantly outperforms traditional multivariate time series models.

The primary contributions of this work are twofold:

- **Methodological Advancement:** By shifting from a temporal-step to a variate-based embedding approach and incorporating geographic orientation ($\sin\theta, \cos\theta$), the model achieves a deeper physical understanding of pollutant dispersion.

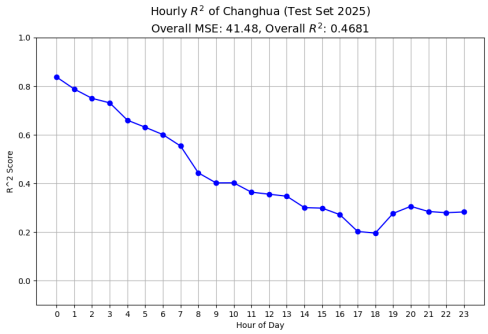
- **Interpretability and Robustness:** The adoption of the Entmax mechanism not only enhances predictive accuracy but also provides a more interpretable machine learning solution by autonomously filtering out environmental noise and identifying key "upwind" spatial predictors.

Ultimately, this research provides a robust data-driven tool for air quality management, offering higher precision for public health warnings and a scalable architecture for atmospheric science applications.

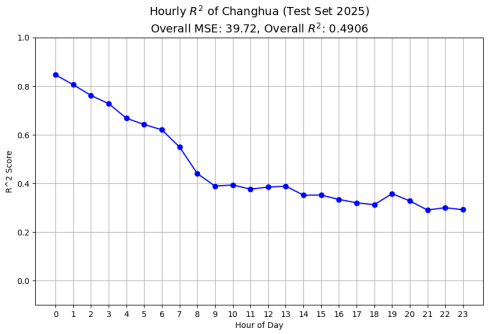
A Appendix: Performance Visualization (Changhua & Xianxi)

This appendix compares the 24-hour $PM_{2.5}$ forecasting performance between the baseline Softmax and the proposed Entmax mechanism. As shown in the figures below, replacing the dense Softmax with Entmax allows the model to ignore environmental noise and better track concentration fluctuations.

A.1 Changhua Station



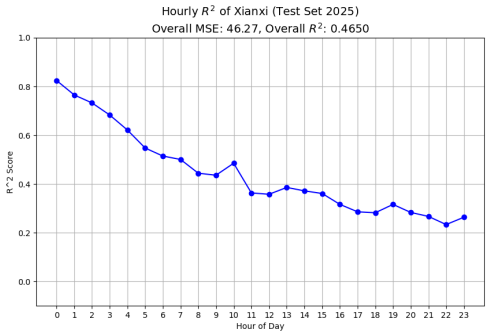
(a) iTransformer (Softmax)



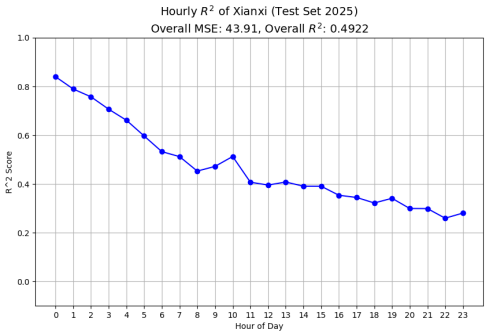
(b) iTransformer (Entmax)

Figure 1: Forecasting comparison at Changhua Station.

A.2 Xianxi Station



(a) iTransformer (Softmax)



(b) iTransformer (Entmax)

Figure 2: Forecasting comparison at Xianxi Station.