



Article

Incorporating the Third Law of Geography with Spatial Attention Module–Convolutional Neural Network–Transformer for Fine-Grained Non-Stationary Air Quality Predictive Learning

Shaofu Lin ^{1,†} , Yuying Zhang ^{1,†}, Xiliang Liu ^{1,*,†} , Qiang Mei ², Xiaoying Zhi ¹ and Xingjia Fei ¹

¹ Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; linshaofu@bjut.edu.cn (S.L.); zhangyuying@emails.bjut.edu.cn (Y.Z.); s202375050@emails.bjut.edu.cn (X.Z.); feixingjia@emails.bjut.edu.cn (X.F.)

² Navigation College, Jimei University, Xiamen 361021, China; meiqiang@jmu.edu.cn

* Correspondence: liuxl@bjut.edu.cn

† These authors contributed equally to this work.

Abstract: Accurate air quality prediction is paramount in safeguarding public health and addressing air pollution control. However, previous studies often ignore the geographic similarity among different monitoring stations and face challenges in dynamically capturing different spatial–temporal relationships between stations. To address this, an air quality predictive learning approach incorporating the Third Law of Geography with SAM–CNN–Transformer is proposed. Firstly, the Third Law of Geography is incorporated to fully consider the geographical similarity among stations via a variogram and spatial clustering. Subsequently, a spatial–temporal attention convolutional network that combines the spatial attention module (SAM) with the convolutional neural network (CNN) and Transformer is designed. The SAM is employed to extract spatial–temporal features from the input data. The CNN is utilized to capture local information and relationships among each input feature. The Transformer is applied to capture time dependencies across long-distance time series. Finally, Shapley’s analysis is employed to interpret the model factors. Numerous experiments with two typical air pollutants (PM_{2.5}, PM₁₀) in Haikou City show that the proposed approach has better comprehensive performance than baseline models. The proposed approach offers an effective and practical methodology for fine-grained non-stationary air quality predictive learning.

Keywords: air quality prediction; the third law of geography; spatial–temporal attention convolutional network; Shapley’s analysis

MSC: 37M10; 68T07



Citation: Lin, S.; Zhang, Y.; Liu, X.; Mei, Q.; Zhi, X.; Fei, X. Incorporating the Third Law of Geography with Spatial Attention Module–Convolutional Neural Network–Transformer for Fine-Grained Non-Stationary Air Quality Predictive Learning. *Mathematics* **2024**, *12*, 1457. <https://doi.org/10.3390/math12101457>

Academic Editor: Daniel-Ioan Curiac

Received: 16 April 2024

Revised: 4 May 2024

Accepted: 7 May 2024

Published: 8 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, rapid urbanization and industrialization have led to deteriorating air quality, which poses a serious threat to public health [1,2]. Furthermore, prior studies indicate that prolonged exposure to air pollution can stimulate lung tissue and affect vascular endothelial function, thereby precipitating a range of diseases affecting the respiratory and cardiovascular systems, either directly or indirectly causing harm to the entire body [3,4]. Therefore, the effective control of air pollutants has become an urgent issue in many countries. Providing a comprehensive analysis of air quality and predicting the concentration of various pollutants in the air in advance is essential for controlling air pollution levels and supporting policy formulation. This effort holds paramount significance in safeguarding public health.

According to literature reviews, many factors influence the efficacy of air quality predictive learning, which strongly depends on the following three complex factors: modeling feature engineering [5,6], the application of geography laws [7,8], and the selection of predictive models [9,10].

Firstly, existing studies reveal intricate interactions among multiple air pollutants [5,11]. Consequently, integrating multiple air pollutants into the predictive learning of air quality becomes imperative. For example, Ma et al. considered the interaction between $PM_{2.5}$, PM_{10} , SO_2 , benzene, and other pollutants when predicting the air quality index, resulting in heightened prediction accuracy [12]. However, the studies exclusively considering inter-pollutant interactions often overlook the impacts of geographical location and meteorological factors on air pollution concentrations. To address this, Lin et al. incorporated meteorological and other pollutant exogenous factors in $PM_{2.5}$ concentration prediction to comprehensively capture features influencing the $PM_{2.5}$ concentration sequence [13]. Multivariate models embedding diverse factors have demonstrated significant contributions to enhancing prediction capability [14]. It is worth noting that an efficient method is required to fully capture the interrelationships between multivariate data to obtain effective information while avoiding redundant information that interferes with model training.

Moreover, the majority of current air quality predictive learning methods are based on the principle of the First Law of Geography. This law posits that everything is related to everything else, but near things are more connected than distant things [15]. Based on this theory, the Pearson correlation coefficient and proximity are commonly applied to the selection of correlated stations in most studies [7]. However, these approaches achieve poor results as they solely consider relationships between adjacent stations within the geographic area, overlooking the complex interactions induced by human factors in distinct urban functional areas. The Third Law of Geography can fully play an important role in addressing uncontrollable patterns of spatial variation in air pollutant concentrations. This law points out that the more similar the geographical environment is, the more similar the geographical target characteristics are [16]. The spatial correlation of air pollution concentration at each monitoring station is not only correlated to the distance between stations but also to the environment surrounding the stations [17]. In practice, human production activities and the distribution of Points of Interest (POIs) can lead to differences in the distribution of pollutant concentrations [18]. In addition, considering the influence of wind direction and the effect of airflow and atmospheric circulation, there is a specific correlation and synergy between the concentration of air pollutants at two monitoring stations that may be farther away in a particular direction [19]. Therefore, fully integrating the content of the Third Law of Geography is crucial for predictive learning in air quality.

Last but not least, researchers have proposed many methods based on deep learning architectures to automatically capture dynamic non-linear distribution characteristics. The widely applied foundational models include attention mechanisms [20,21], convolutional neural networks (CNNs) [22,23], and long short-term memory networks (LSTMs) [11,24]. A common strategy for modeling air quality predictive learning is to integrate these individual models, leveraging their respective strengths to enhance prediction accuracy and stability [25]. Although the combination of these models has improved the accuracy of air quality prediction to a certain extent, there are still some limitations. Recently, the Transformer models have been successfully applied in time series forecasting tasks [26]. The model can capture the long-term dependencies in the data, presenting a viable approach for achieving accurate fine-grained air quality predictive learning based on the Transformer model [26].

As far as we know, most of the current studies on air quality predictive learning predominantly center around the identification of correlated stations using the First Law of Geography [7]. However, there is a limited number of studies that take into account regional influences and the anisotropy of spatial air pollutant concentrations. In addition, the existing deep learning models are difficult to fully extract the spatial-temporal features of air pollutants. To address the weaknesses of previous studies, an air quality predictive learning approach incorporating the Third Law of Geography with SAM-CNN-Transformer is proposed. The proposed approach fully considers the interaction between different

air pollutants and meteorological factors, the environmental similarity between stations through the Third Law of Geography, and the spatial–temporal features of air pollutant concentrations through the SAM–CNN–Transformer model. The main contributions of this study can be summarized as follows:

- (1) The Third Law of Geography is incorporated. The spatial clustering results of POI data are used as a characterization parameter to fully consider the correlation and synergism among different geospatial monitoring stations. The spatial anisotropy analysis is also utilized to optimize the impacts of spatial factors to fully consider the spatial variability of the atmospheric physical processes of air pollution.
- (2) This study notes the advantages of the hybrid deep learning model based on fusion mechanisms in dealing with spatial–temporal dependencies. SAM, CNN, and Transformer are integrated with the overall structural design to fully extract the spatial–temporal distribution features of the stations; it overcomes the problems existing in typical deep learning methods, such as gradient vanishing, gradient explosion, etc.
- (3) Shapley’s analysis is employed to assess the importance of air pollutant concentrations, meteorological factors, and correlated stations’ influences on the model predictive learning, providing direction for further modeling.

The rest of this paper is organized as follows: Section 2 briefly summarizes the related work. Section 3 reviews the theoretical principles of the involved methods. Section 4 presents several experiments to analyze the obtained results. Section 5 outlines the discussion. Finally, the conclusion and future work are presented in Section 6.

2. Related Works

2.1. The Laws of Geography in Spatial–Temporal Forecasting

The First Law of Geography [15], describing a spatial similarity and autocorrelation of geographic phenomena, has gained attention in spatial–temporal air quality predictive learning [7,8,27]. For example, Seng et al. utilized spatial information from five nearby stations with the highest Pearson correlation coefficients to improve prediction accuracy [7]. Mao et al. treated pollutants at different stations as a spatial adjacency matrix, employing graph convolutional networks for spatial dependency modeling [8]. However, these studies often rely on Pearson correlation coefficients or identify strongly correlated stations based on proximity, introducing subjectivity. There is limited emphasis on extracting spatial–temporal correlation features, particularly concerning uncontrollable spatial change patterns in geographic phenomena and their variability. For instance, geographic phenomena may decrease with distance due to wind direction, air movement, and atmospheric circulation, resulting in spatial correlations between monitoring stations more pronounced in specific directions [28]. The First Law of Geography primarily focuses on the spatial distance between two stations, overlooking the interaction among different geographical elements in natural phenomena. We are motivated by the Third Law of Geography [16], which asserts that “similar geographical environments lead to analogous features in geographical targets”. The theory of environmental similarity is broadly applied across multiple domains, including the groundwater level [29] and soil organic carbon predictive learning [22]. However, in the field of air quality, few researchers have thoroughly examined the similarity between the geographical environments of monitoring stations when analyzing the spatial and temporal characteristics of air pollution. Figure 1 illustrates this concept, with circles representing air pollution monitoring stations categorized into different regions based on POI information around their locations. According to the First Law of Geography, monitoring stations A and B, being geographically adjacent, are expected to show similarities in air quality. Furthermore, both station A and station C are in proximity to power plants, suggesting, in accordance with the Third Law of Geography, that stations A and C may exhibit similar patterns despite their considerable geographical separation. Therefore, it becomes evident that fully considering the geographical similarity of all stations is crucial for enhancing the accuracy of air quality predictive learning.

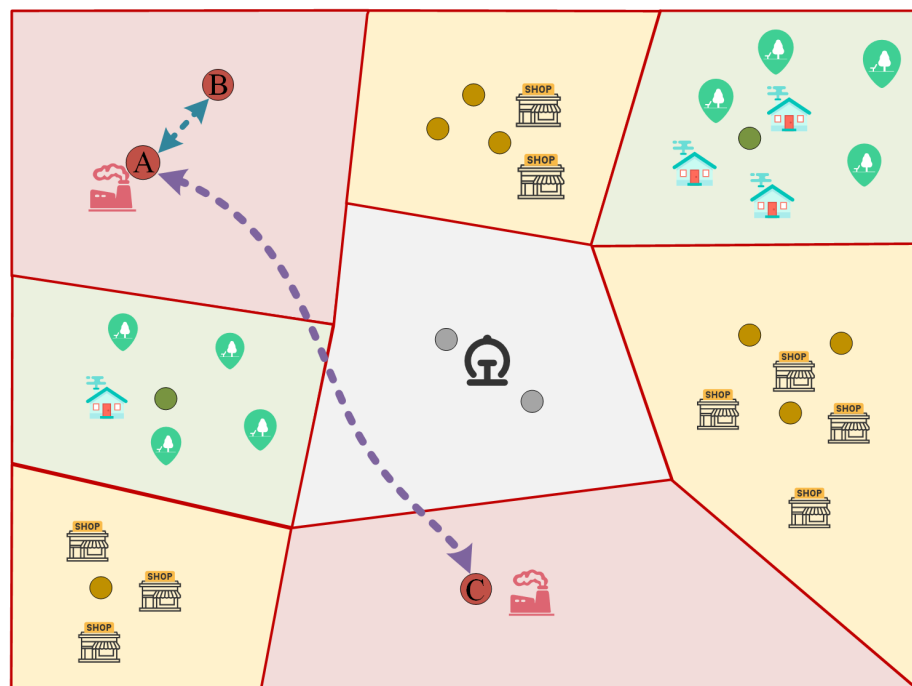


Figure 1. A demonstration of the effectiveness of the First and Third Geographic Laws. A–C represent air quality monitoring stations.

2.2. Spatial Feature Extraction

CNN demonstrates great potential in extracting spatial features, and its utilization for mining spatial dependencies between different stations is widely employed in air quality prediction. For example, when predicting temperature, Hou et al. utilized the CNN network to extract features for obtaining local spatial temperature characteristics [30]. Mao et al. employed deep convolutional neural networks to extract the sequential features of time series data, and the experimental results show that CNN performs well for air quality prediction [31]. However, the uneven distribution of monitoring stations, dynamically changing spatial relationships, and CNNs' limitation in capturing localized areas pose challenges in fully mining irregular topological information in the pollutant monitoring network, thereby affecting predictive learning accuracy.

The Graph Convolutional Neural Network (GCN) establishes inter-nodal connections through an adjacency matrix, rendering it particularly apt for handling irregular spatial data, especially when contrasted with CNN [32]. For instance, Qi et al. employed an undirected graph to capture the topological relationships among air quality monitoring sites [24]. Wu et al. adaptively integrated monitoring network topology, ancillary pollutants, and meteorological factors into the GCN model to discern spatial dependence for air quality $PM_{2.5}$ [10]. The existing GCN-based air quality predictive learning research aims to capture the spatial dependence between multiple air quality monitoring stations. However, since the modeling process is a hypothetical static topology structure, it may ignore the spatial–temporal dynamics of the connection between actual monitoring stations.

The attention mechanism has demonstrated significant advantages in capturing crucial information from the past state among monitoring stations [21]. For instance, Yang et al. applied a spatial–temporal attention mechanism in air quality predictive learning, extracting spatial–temporal features from embedded data. This mechanism incorporates self-attention calculations in spatial and temporal dimensions, resulting in more accurate predictions [20]. Wang et al. proposed a Spatial and Channel Calibration Network (SCCNet), integrating spatial and channel attention to effectively extract spatial–temporal dependencies of air pollutants [33]. Therefore, the attention mechanism can adaptively learn the importance of different stations and features, enabling a more flexible and accurate capture of critical information in spatial feature extraction.

2.3. Temporal Feature Extraction

Typical time series prediction models based on deep learning include recurrent neural network (RNN) [34–36], Gated Recurrent Unit (GRU) [37,38], long short-term memory (LSTM) [6,11], Transformer [26,39], and so on. The prediction models based on RNN usually cannot capture the long-term dependence in the input sequences, and it is accompanied by a gradient explosion and gradient disappearance during the training process [40]. RNNs are challenged by long-term dependencies and gradient issues, shortcomings mitigated by GRU and LSTM interventions [6,41]. However, GRU and LSTM confront obstacles in parallel processing and managing long-distance dependencies. Recent methodologies incorporating multi-head attention mechanisms present promising resolutions [26,39,42]. Diverging from RNN and LSTM, the Transformer model exclusively relies on the self-attention mechanism, adept at assimilating global information and modeling extensive dependencies. Exemplarily, Reza, et al. proposed a multi-attention-based Transformer model for traffic flow prediction, demonstrating efficacy in forecasting prolonged patterns [43]. Pundir et al. utilized the Transformer model for air quality index prediction and showed that the Transformer model outperforms the widely used RNN-LSTM model and regression model compared to these two models [44]. These inquiries underscore the Transformer's prowess in handling protracted time series for air quality predictive learning. Nonetheless, for heightened precision in air quality predictive learning, a diligent consideration of temporal correlation characteristics and comprehensive analysis of spatial and temporal features remain imperative.

2.4. Spatial–Temporal Feature Extraction

While deep learning methods have significantly enhanced air quality prediction accuracy, the performance of individual models may be limited for complex non-linear problems [25]. Typically, a single model can only capture either the spatial or temporal characteristics of air pollutants, making it challenging to simultaneously extract deep spatial and temporal dependencies [45]. Hence, coupling-based spatial–temporal joint prediction models have become pivotal in air quality predictive learning. Specifically, the hybrid deep learning model that combines CNN and LSTM finds widespread application in air quality predictive learning. For instance, Zhang et al. integrated CNN and LSTM for air quality predictive learning, demonstrating superior prediction performance compared to the standalone CNN and LSTM models [5]. Wen et al. employed a combination of 3D-CNN and LSTM to extract advanced spatial–temporal features for PM_{2.5} concentration prediction, outperforming other models [46]. Additionally, the fusion of GCN and LSTM has gained popularity for air quality spatial–temporal prediction modeling [10,47]. In these studies, CNN and GCN are employed for extracting spatial features between air quality monitoring stations, while LSTM models are typically used to uncover temporal dependencies in historical time series. With the recent introduction of attentional mechanisms, incorporating these mechanisms to capture future spatial–temporal relationships in potential spaces has become a prominent research focus. For example, Huang et al. introduced SpAttRNN, a novel spatial attention-embedded recurrent neural network for AQI prediction [48]. Wang et al. combined spatial and channel attention to enhance the extraction of global information in air quality predictive learning [33]. Despite the recent success of joint prediction models in air quality predictive learning, limited attention has been given to the application of spatial attention and Transformer technology in air quality predictive learning. Furthermore, CNN, a typical deep learning model, effectively capturing local features, integrating convolutional networks, spatial attention, and Transformer techniques to enhance air quality prediction accuracy, poses a challenging task, requiring a comprehensive exploration of their synergistic effects.

3. Methodology

3.1. The Framework of The Proposed Approach

The overall framework is depicted in Figure 2. The proposed approach followed multi-source data collection, feature selection, and spatial–temporal dependency extraction to conduct air quality predictive learning. The multi-source data collection segment details the data utilized in this study. The feature selection part outlines the process of completing the correlation stations following the Third Law of Geography. The spatial–temporal feature extraction elucidates the modeling process of this study. The detailed process of the three main stages is shown below.

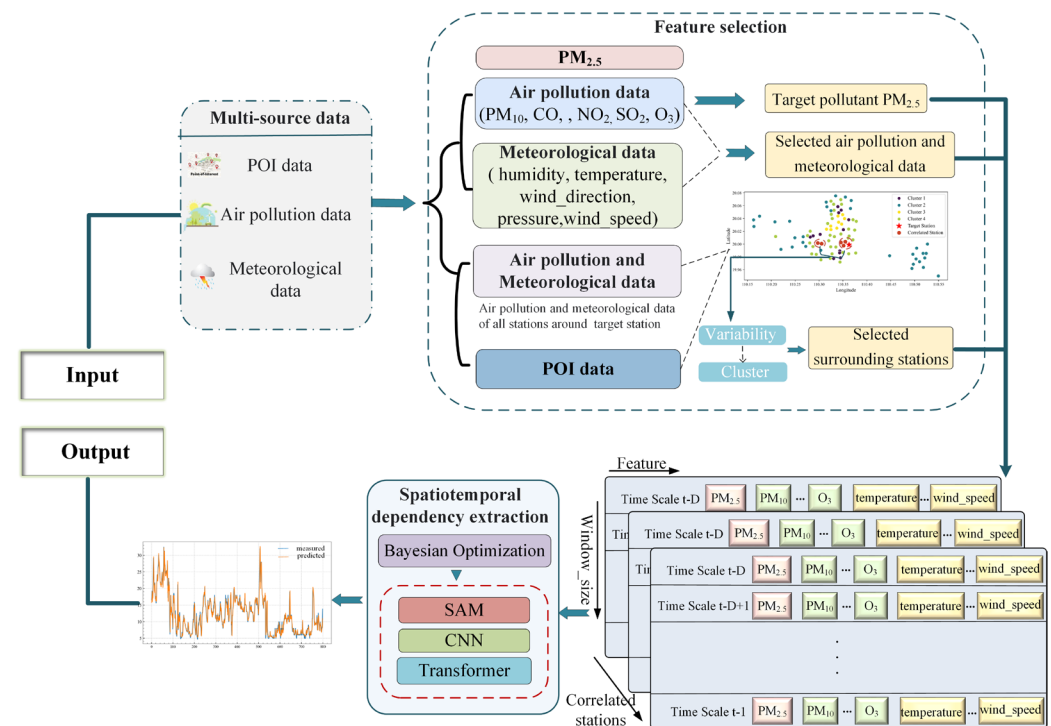


Figure 2. The framework of the proposed approach.

Stage 1: Multi-source data collection

In this stage, multi-source data in the study area are collected. Firstly, the study area's POI data are obtained through Baidu's (<http://api.map.baidu.com>, accessed on 25 January 2024) open API. The rectangular box search method given by official Baidu Maps is chosen for POI retrieval. Based on the POI industry classification provided by Baidu Maps, we used the rectangular box search method to obtain various types of POI information for each administrative district in the study area. For densely populated POIs such as schools and supermarkets, we further refined the latitude and longitude intervals of the delineation grid to enhance the completeness of POI acquisition. This ensures the reliability and completeness of the acquired POI data. POI data are categorized into 14 types such as restaurant and food, business and residential, tourist attractions, etc. Each POI point contains six attributes: the name, first-level classification, second-level classification, longitude, latitude, and region. Each POI point includes six attributes. This study selects the first-level classification, longitude, and latitude for application. The POI data are used for the spatial clustering analysis of all monitoring stations. The latitude and longitude are used to calculate the distance between each POI and each monitoring station in the study area. In addition, the number of each type of POI around each station is counted to be used for the spatial clustering analysis. Secondly, six air pollutant factors and five meteorological factors are collected from the monitoring stations in the study area.

Stage 2: Feature selection

The target pollutant (e.g., PM_{2.5}) at the target stations, serving as a pivotal indicator for the direct assessment of air quality, is utilized as an input feature. Additional air pollutants and meteorological factors at the target station are considered for predictive learning, potentially influencing PM_{2.5} concentration. This study assesses the impact of all stations on the target stations from two perspectives as spatial anisotropy and spatial clustering. Initially, the spatial variability of PM_{2.5} concentrations at various time points across all monitoring stations in the study area is analyzed. Spatial clustering at all monitoring stations, based on POI data using hierarchical clustering, is then conducted. The clustering containing the target station is selected and further analyzed, considering the results of the spatial variability analysis, to identify stations with a strong correlation and synergy.

Stage 3: Spatial–temporal dependency extraction

Combine spatial attention mechanism (SAM), convolutional neural network (CNN), and Transformer techniques through a fusion mechanism for comprehensive spatial–temporal dependency extraction. The entire process comprises two primary components. In terms of modeling, spatial distribution features between strongly correlated stations are extracted using SAM. Specifically, self-attention computations are performed in the spatial dimension and spatial–temporal features are extracted from the strong-correlation 3D matrix. SAM is able to capture the spatial–temporal correlation information in air quality data well. The SAM is initially employed to perform self-attention calculations in the spatial dimension, extracting spatial–temporal features from the embedded data. Subsequently, CNN is utilized to capture local information and relationships among each input feature. Finally, the Transformer is applied to capture time dependencies across long-distance time series. Concerning model optimization, the Bayesian optimization is employed to fine-tune hyperparameters, ensuring optimal prediction performance.

3.2. Analysis of Spatial Anisotropy

Various geographic phenomena significantly impact the spatial variation of air pollutant concentrations. An anisotropy analysis can help to understand the differences in the distribution of air pollutants in different geographic directions, thus revealing the mechanisms by which multiple factors in geographic space affect air quality. This analytical method primarily examines the variation patterns of spatial data in different directions, assessing whether the data exhibit differences or variations along various directions. The core of an anisotropy analysis is the variogram, also known as the semi-variogram, which describes the degree of variation in the data within a specific distance range. The formula for calculating semi-variance is as follows:

$$Y(h) = \frac{1}{2N(h)} \sum_{i=1}^N (Z(x_i) - Z(x_i + h))^2 \quad (1)$$

where $Y(h)$ is the semi-variance at lag distance h ; $N(h)$ is the number of pairs of sample points separated by lag distance h ; $Z(x_i)$ and $Z(x_i + h)$ are the values of the variable at locations x_i and $x_i + h$, respectively.

The semi-variance function describes the average semi-variance between pairs of spatial points at a distance h . By calculating the semi-variant at different distances h , semi-variograms can be constructed for visualizing the variability of spatial data.

3.3. Spatial Clustering

It is clear from the Third Law of Geography that a synergistic relationship often exists between one element of a geographic phenomenon and other geographic elements. To explore the similarity and spatial association patterns among monitoring stations, this study acquires POI data of the study area through Baidu's open API. The POI data encompasses various geographical entities such as commercial districts, cultural facilities, transportation hubs, etc. Subsequently, hierarchical clustering is applied to spatially

group all monitoring stations. Hierarchical clustering methods construct a dendrogram by grouping spatially adjacent and similar stations into clusters. This approach circumvents the need to pre-determine the number of clusters, facilitating the exploration of potential geographical patterns within the research area without prior knowledge. Analyzing the clustering results contributes to a comprehensive understanding of the geographic relationships among monitoring stations, revealing combinations of stations that exhibit spatially close correlations with synergistic changes. The pseudo-code of the module spatial clustering is shown as Algorithm 1, and the formulas involved in the algorithm are given in Equations (2)–(4).

$$d(S_i, P_j) = 2\arcsin \sqrt{\sin^2 \frac{(plat - slat)}{2} + \cos(plat) \times \cos(slat) \times \sin^2 \frac{(p \ln g - s \ln g)}{2}} \times 6378.137 \quad (2)$$

$$dist(p_a, q_b) = \sqrt{\sum_{k=1}^n \left(\frac{x_k^{(p_a)} - x_k^{(q_b)}}{s_k} \right)^2} \quad (3)$$

$$D(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{p_a \in C_i, q_b \in C_j} dist(p_a, q_b) \quad (4)$$

where $p \ln g, plat$ denote the latitude and longitude of the POI. $s \ln g, slat$ denote the latitude and longitude of the monitoring stations. The radius of the Earth's equator in kilometers is 6378.137. $dist(p_a, q_b)$ represents the Normalized Euclidean distance between data points p_a and q_b ; n represents the number of dimensions for the data point; $x_k^{(p_a)}$ and $x_k^{(q_b)}$, respectively, represent the values of data points p_a and q_b in the k dimension; s_k is the standard deviation on the k dimension; $D(C_i, C_j)$ represents the similarity between clusters C_i and C_j ; and $|C_i|$ and $|C_j|$, respectively, represent the number of samples in the cluster.

Algorithm 1 Proposed “Spatial clustering” Approach

Input: $S = \{S_1, S_2, \dots, S_n\} (n \in 1, \dots, N); P = \{P_1, P_2, \dots, P_m\} (m \in 1, \dots, M)$. // S_n, P_m represents station location information and POI Information, respectively;

Output: C ;

```

1:  $S^* = \{S_1^*, S_2^*, \dots, S_n^*\}$  // initialize  $S^*$  to a matrix of  $n \times k$  dimensions,  $S_n^*$  to a matrix of  $1 \times k$  dimensions;
2: for  $S_i$  in  $\{S_1, S_2, \dots, S_n\}$  do
3:   for  $P_j$  in  $\{P_1, P_2, \dots, P_m\}$  do
4:     compute  $d(S_i, P_j)$ . according to Equation (2);
5:   if  $d(S_i, P_j) < 1\text{km}$  do
6:     update  $S_i^*$ ;
7:    $C = \{C_1, \dots, C_n\}$  // Each  $S_n^*$  is regarded as a separate cluster;
8:   while  $n > 1$  do
9:     for  $C_i$  in  $\{C_1, \dots, C_n\}$  do
10:      for  $C_j$  in  $\{C_1, \dots, C_n\}$  do
11:         $M(i, j) = D(C_i, C_j)$  according to Equations (3) and (4);
12:         $M(j, i) = M(i, j)$ ;
13:      find the most similar clusters:  $C_{i*}$  and  $C_{j*}$ ;
14:      merge  $C_{i*}$  and  $C_{j*}$ :  $C_{i*} = C_{i*} \cup C_{j*}$ ;
15:      for  $k = j* + 1, j* + 2, \dots, n$  do
16:         $C_k = C_{k+1}$ ;
17:       $n = n - 1$ ;
18:   return  $C$ ;
```

3.4. SAM–CNN–Transformer Network

This study utilizes multivariate series data from the target stations as well as the historical time of strongly correlated stations to predict the target pollution concentrations at future time points. This study employs a spatial variogram analysis and spatial clustering to identify the correlated stations that exhibit mutual interaction with the target station, denoted by $X_* = \{X_1, X_2, \dots, X_k\} (k \in 1, \dots, K)$, where $X_k \in R^{D \times L}$ denotes the feature matrix of the target station as well as the k th station associated with it, D denotes the size of the historical time step, and L denotes the feature variable for each time step. $X_* \in R^{K \times D \times L}$ denotes the 3D feature matrix consisting of the target station and the correlated stations, where $K \leq N$ and N denotes the number of all stations.

The network structure is shown in Figure 3. The specific steps are as follows:

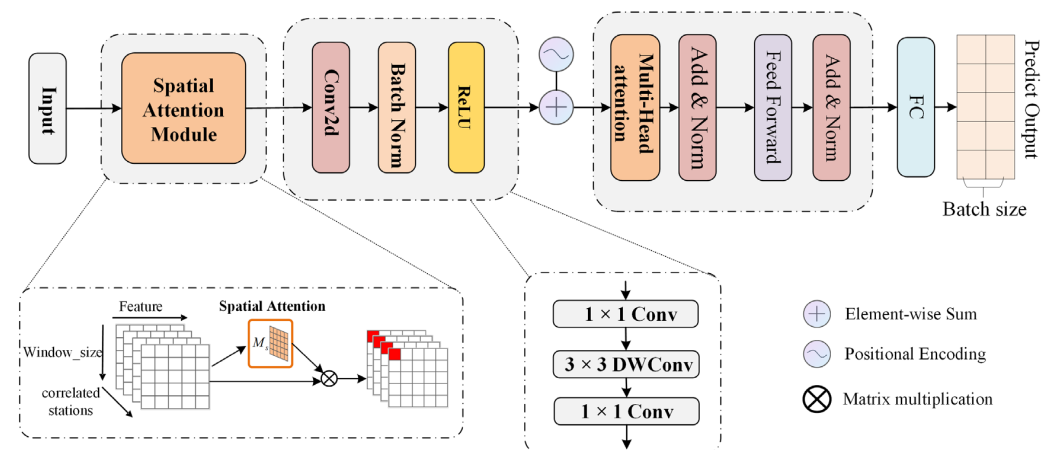


Figure 3. SAM–CNN–Transformer network architecture.

Firstly, the 3D strong correlation matrix $X_* \in R^{K \times D \times L}$ serves as input to the spatial attention module (SAM), enabling the extraction of spatial distribution features among the strong stations by computing feature and weight maps in the spatial dimension. The specific workflow is as follows: Feature maps and weight maps are computed by SAM using the input matrix. The feature map reflects the feature relationships at different spatial locations, and the weight map indicates the degree of feature attention between different spatial locations. Specifically, the SAM applies average pooling and maximum pooling operations along the channel of each feature point X_* , stacks the results, and obtains a spatial attention weight distribution $M_s(X_*)$. Finally, this weight distribution is multiplied by the input features, yielding the weighted feature layer X'_* . This allows the SAM to focus more on important spatially distributed features, which improves the model's ability to understand the input features. The calculation formula for this process is presented below:

$$M_s(X_*) = \sigma(\text{Conv}([\text{AvgPool}(X_*); \text{MaxPool}(X_*)])) \quad (5)$$

$$X'_* = M_s(X_*) \otimes X_* \quad (6)$$

where $M_s(X_*)$ denotes the spatial attention weights, $\text{AvgPool}(\cdot)$ denotes the average pooling operation, $\text{MaxPool}(\cdot)$ denotes the maximum pooling operation, σ denotes the activation function, $\text{Conv}(\cdot)$ denotes the convolutional layer, and X'_* denotes the final spatial feature map of the station obtained by multiplying the spatial attention weights with the weighted input feature layer.

Secondly, CNN is utilized in order to efficiently extract local spatial information between strongly correlated stations, and the extracted features are input to Transformer for further prediction. Specifically, a feature capture is further performed using local receptive fields using the CNN. The tensor is projected to a higher-dimensional space for more object representations through a 1×1 convolutional layer. Then, a 3×3 depth-wise convolution

is performed to process the feature maps for each channel. The number of channels in the feature maps is reduced using a 1×1 convolutional layer to obtain a two-dimensional matrix $X_*'' \in R^{D \times L}$ after dimensionality reduction.

Finally, the spatially extracted 2D feature matrix $X_*'' \in R^{D \times L}$ is input to the Transformer model as a time series. When addressing time series forecasting, the Transformer captures dependencies among different locations in the input series through self-attention. This enhances the model's understanding of the long-term dependencies in the series. A critical component of the Transformer is the multi-head self-attention mechanism, illustrated in Figure 4. Specifically, for each position in the input sequence, the self-attention mechanism calculates correlation weights with other positions, applying them to obtain a representation of that position. The formula for the self-attention mechanism is as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

where Q , K , and V denote the query vector, key vector, and value vector, respectively; d_k denotes the dimensionality of the key; and $\sqrt{d_k}$ is used to scale the dot product and is the activation function that maps the input to the interval $[0, 1]$. The self-attention mechanism obtains the attention weights by calculating the similarity between the query and the keys and obtains the final representation by weighted summation.

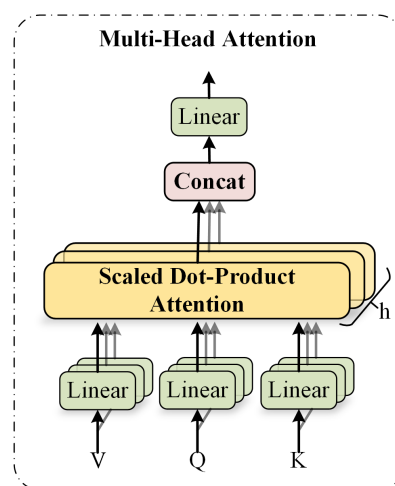


Figure 4. The multinomial self-attention mechanism.

Compared to the original Transformer structure, the model omits the need to calculate the probabilities through *Softmax* in the end. Instead, by mapping the resulting feature maps to the output values, the final predicted value of the target pollutant concentration at the station at time t is obtained.

4. Results

4.1. Data and Study Area

The city of Haikou, China, serves as the study area, with a dataset comprising POI data, air quality data, and meteorological data. Among these, the POI data, obtained through Baidu's open API, cover 14 different types. Each POI point contains six attributes, where first-level classification, longitude, and latitude are selected for the application, resulting in a total of 92,108 POI points. Air quality data include hourly air pollution concentration data from 95 air quality monitoring stations in Haikou City, spanning from 26 May 2021 to 11 March 2023, along with corresponding meteorological data. The distribution of monitoring stations is illustrated in Figure 5. In this study, S9 is chosen as the target station, marked in red in Figure 5. S9, located in the city center and surrounded by numerous stores, offers a better response to the influence of spatial-temporal correlation on the model's

prediction and is representative to a certain extent. In addition, S6, S25, and S41 are used in the third group of experiments to further validate the generalization ability of the proposed approach, and their geographic locations are shown in Figure 5. The data are explicitly described in Table 1. Data collection may encounter issues such as equipment damage and transmission errors, leading to outliers and missing values. Therefore, preprocessing is essential for air quality predictive learning. Three different approaches are utilized to handle missing values based on the varying durations of the missing values. Specifically, the forward-filling method is used for missing values in the short time periods (e.g., within 4 h) of the original data, the multiple interpolation method is used to fill in the missing values in the medium and long time periods (e.g., more than 4 h and less than 72 h), and the missing values in the long time periods (e.g., more than 72 h) are directly deleted and processed.

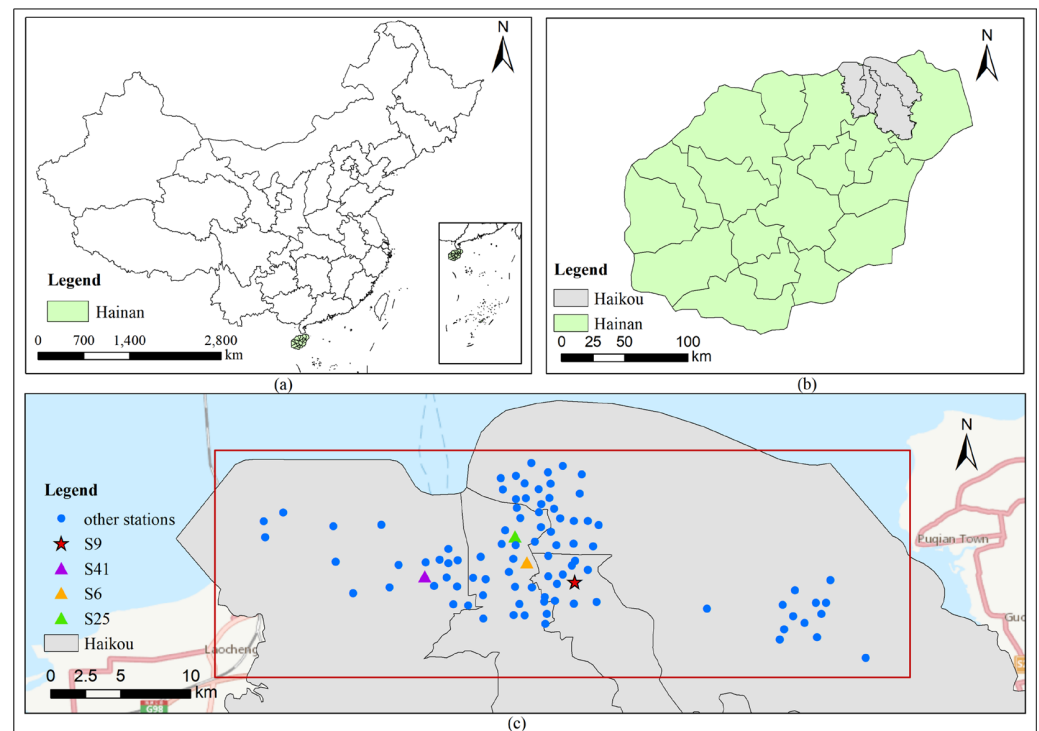


Figure 5. Study area and spatial distribution of air monitoring stations (the base of the map is from ESRI (<https://hub.arcgis.com/maps/0c539fdb47d34b17bd1452f6b9f49e97/explore>, accessed on 25 January 2024)): (a) The green part is the boundary map of Hainan Province. (b) The green part is the boundary map of Hainan Province, and the gray part is the boundary map of Haikou city. (c) Distribution of air monitoring stations in Haikou city.

Table 1. Dataset description.

Type		Variable	Unit
POI data	POI	First-level classification	-
		Longitude	-
		Latitude	-
Air quality data	Particulate pollutant	PM _{2.5}	µg/m ³
		PM ₁₀	µg/m ³
	Gaseous pollutant	CO	µg/m ³
		NO ₂	µg/m ³
		SO ₂	µg/m ³
		O ₃	µg/m ³

Table 1. Cont.

Type		Variable	Unit
Meteorological data	Meteorological factor	Pressure	hPa
		Humidity	%
		Temperature	°C
		Wind_direction	-
		Wind_speed	km/h

4.2. Evaluation Metrics

To comprehensively evaluate the performance of the proposed approach in this study, three evaluation metrics as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-Square (R^2) are used. These metrics are calculated as shown in Equations (8)–(10). RMSE measures the deviation between the predicted value and actual value, while MAE provides a better reflection of prediction errors. The smaller values of RMSE and MAE indicate the higher model accuracy. R^2 assesses the fitting ability of the model, and the value closer to 1 indicates better fitting of the predictive learning result.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^N (y_t - \hat{y}_t)^2} \quad (8)$$

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^N |y_t - \hat{y}_t| \quad (9)$$

$$R^2 = 1 - \frac{\sum_{t=0}^N (y_t - \hat{y}_t)^2}{\sum_{t=0}^N (y_t - \bar{y})^2} \quad (10)$$

where N denotes the total number of samples, y_t denotes the true value, \hat{y}_t denotes the predicted value, and \bar{y} denotes the mean value.

4.3. The Software and Hardware Details

In this study, the proposed deep learning and baseline models are implemented using Python packages (including PyTorch, an open-source deep learning framework that can accelerate the training process using GPUs and distributed computing; Pandas and Numpy are used for data processing and analysis; Scikit-gstat is used for spatial variability analysis; and Matplotlib and Seaborn are used to visualize the results). Heavier workloads are run on a computer equipped with Intel(R) Core (TM) i9-10900X CPU @ 3.70 GHz 3.70 GHz 64.0 GB RAM. The rest of the models are conducted on a server equipped with Intel(R) Core (TM) i75600U CPU @ 2.60 GHz 3.2 GHz 16.0 GB RAM.

4.4. Hyperparameter Tuning Based on Bayesian Optimization

Hyperparameter tuning plays a crucial role in optimizing model performance, and Bayesian optimization serves as a valuable tool for selecting hyperparameters during the model training process. It is a global optimization method grounded in the principles of Bayesian inference, featuring a core algorithm that includes the probabilistic agent model and the sampling function. First, the probabilistic agent model employs Gaussian process regression to approximate the objective function, computing the mean and variance of function values at each point to establish a likely distribution. Second, a sampling function determines points for sampling in the current iteration, aiming to balance the exploration of unknown regions with the utilization of known high-performance regions. By leveraging both prior and posterior models, Bayesian optimization theoretically ensures eventual convergence, finding the input parameter combination that optimizes the objective function within a finite number of iterations.

In this study, the model's hyperparameters are categorized into two groups. One group defines the model and its structure, and the other group is for the objective function and the optimization algorithm (solver). The group of parameters defining the model and structure includes num_layers, num_heads, d_model, d_ff, and dropout. Specifically, num_layers indicates the stacked encoder and decoder layers in the Transformer, num_heads represents the amount of multi-head attention, d_model denotes the model's dimensionality, d_ff corresponds to the hidden layer's dimensionality in the feedforward neural network, and dropout signifies the proportion of randomly dropped nodes in each iteration. The group of hyperparameters required for the objective function and optimization algorithm includes batch size, learning rate, time step, and optimizer. Specifically, batch size denotes the size of the data input to the model for training in each iteration, learning rate determines the step size for each iteration, time step represents the size of the historical time series window, and optimizer signifies the adaptive learning rate optimization algorithm. In the Bayesian optimization process, RMSE serves as the objective function. The scikit-optimize library's Bayesian optimization algorithm is employed to adjust these parameters. The ultimately optimized values for hyperparameters are presented in Table 2.

Table 2. Hyperparameter values after Bayesian-based optimization.

Hyperparameter	Value
num_layers	2
num_heads	8
d_model	512
d_ff	2048
dropout	0.05
batch size	1
learning rate	1×10^{-4}
time step	11
optimizer	Adam

4.5. Correlated Station Selection

4.5.1. Analysis of Spatial Anisotropy

This study employs the variational function to analyze the spatial change process of PM_{2.5} concentrations at all monitoring stations in the study area. The variational function is computed using an open-source Python library, scikit-gstat (<https://scikit-gstat.readthedocs.io>, accessed on 28 February 2024)). This approach included the calculation of the semi-variance function, spatial interpolation, analysis of orientation dependence, and visualization of results. The spatial dependence of PM_{2.5} concentration variations in different directions is demonstrated in Figure 6. Overall, the spatial variation of PM_{2.5} concentrations in different directions is very similar. Specifically, Figure 6 shows that over the initial 10 km distance, the shape of the variograms is very similar in each direction. At this range, the anisotropy is not significant. However, the effective range of the variograms in the north–south and southeast–northwest directions extends only up to about 10 km, suggesting that observations become statistically more independent at greater distances in these directions. In contrast, the east–west variograms demonstrate a significantly larger effective range, indicating more substantial correlation lengths in this orientation. Such differences in spatial variation can be attributed to the distribution of pollution sources, meteorological conditions, etc. The analysis of spatial anisotropy provides essential clues for the subsequent selection of strongly correlated stations.

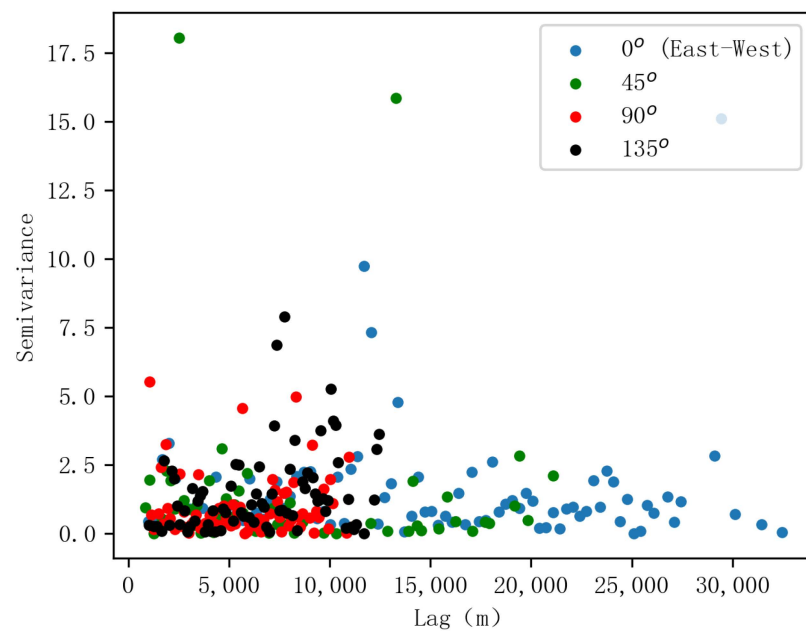


Figure 6. The result of spatial anisotropy.

4.5.2. Spatial Clustering

By utilizing the hierarchical clustering method and incorporating POI data, all 95 monitoring stations are clustered, resulting in four distinct clusters, as shown in Figure 7. Among these, stations in Cluster 1 and Cluster 3 are situated in the city center, characterized by dense architectural facilities, including prominent urban structures like commercial buildings, cultural institutions, and shopping centers. Stations in Cluster 2 are situated in the peripheral zone, primarily surrounded by schools and parks. Stations in Cluster 3 are situated in comprehensive areas, featuring a convergence of various activity facilities in the vicinity. Based on the clustering results, it is observed that the target monitoring station (S9) belongs to Cluster 1. Combined with the spatial variability analysis, the stations within Cluster 1 are further filtered to identify the correlated stations in the east–west direction of the target station. The final selections of strongly correlated stations including S2, S3, S8, S39, and S45 are shown in Figure 7.

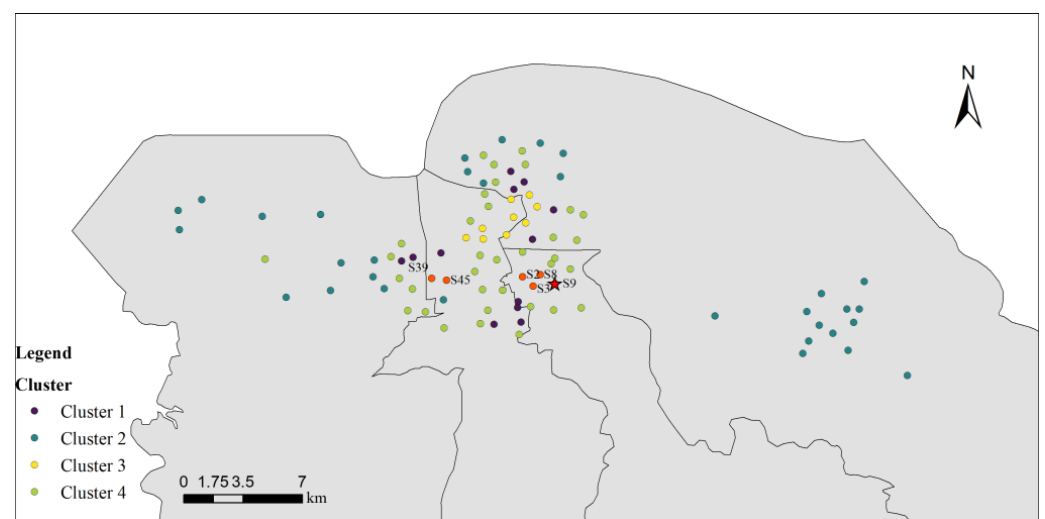


Figure 7. The spatial distribution of selected strongly correlated stations.

4.6. Implementation Details and Comparative Analysis

To verify the effectiveness and robustness of the proposed approach, Table 3 outlines four sets of comparison experiments. The first set of experiments analyzes the predictive performance of six baseline models (ARIMA, SVR, GRU, LSTM, TCN, and Informer) in comparison with the proposed approach, validating its effectiveness. The second set of experiments conducts ablation experiments, evaluating the impact of incorporating meteorological factors and integrating the Third Law of Geography with SAM–CNN–Transformer for air quality predictive learning. The third set of experiments randomly selects stations S6, S25, and S41 from each of the remaining three clusters to examine the predictive performance of the proposed approach for the stations in different clusters. The fourth set of experiments further validates the effectiveness of each module of the proposed model.

Table 3. Model description of the four groups of experiments.

Experiments	Station	Model Notation	Model Description
		Ours	The proposed approach
Experiment I: verifying the effectiveness of the proposed approach	S9	model1	ARIMA
		model2	SVR
		model3	GRU
		model4	LSTM
		model5	TCN
		model6	Informer
Experiment II: testing the predictive ability of various exogenous variables	S9	model7	The proposed approach, disregarding meteorological influencing factors
		model8	The proposed approach, disregarding regional influencing factors
Experiment III: verifying the different modules of the proposed approach	S9	model9	Transformer
		model10	CNN–Transformer
		model11	SAM–Transformer
Experiment IV: verifying the generalization ability of the proposed approach	S6, S25, S41	model1	ARIMA
		model2	SVR
		model3	GRU
		model4	LSTM
		model5	TCN
		model6	Informer

4.6.1. Experiment I: Verifying The Effectiveness of The Proposed Approach

To validate the effectiveness of the proposed approach in Experiment I, it is compared with several baseline models, namely ARIMA, SVR, GRU, LSTM, TCN, and Informer, denoted as model1 to model6. The target station selected for the experiment is S9 in Cluster 1. Table 4 gives the prediction results of the proposed model and the baseline model for

PM_{2.5} and PM₁₀ on the test dataset of the station. The proposed model achieved better performance for both PM_{2.5} and PM₁₀, with the lowest prediction errors for RMSE and MAE and the highest R² score. Taking PM_{2.5} as an example, the values of RMSE, MAE, and R² of the proposed model are 2.168, 1.454, and 0.953, respectively. Compared with the baseline model, the proposed model reduces RMSE and MAE by 15.8% and 12.23% on average, while improving R² by 2.6% on average. In particular, the Informer (model6) in the baseline model is a modification based on the Transformer, which achieves good predictive performance on ETT (Electricity Transformer Temperature) data, outperforming the Transformer [49]. However, in conducting air quality predictive learning experiments, the Transformer-based predictive learning model proposed in this study is significantly better than the Informer-based model.

Table 4. Experimental results of the proposed approach and the baseline models in Experiment I.

Model	PM _{2.5}			PM ₁₀		
	RMSE	MAE	R ²	RMSE	MAE	R ²
model1	3.638	1.553	0.882	5.521	2.698	0.883
model2	2.382	1.913	0.928	3.318	2.376	0.939
model3	2.353	1.645	0.939	3.521	2.429	0.941
model4	2.244	1.524	0.946	4.583	3.365	0.911
model5	2.336	1.583	0.942	3.451	2.340	0.945
model6	2.501	1.722	0.935	4.436	3.166	0.914
Ours	2.168	1.454	0.953	3.331	2.230	0.948

4.6.2. Experiment II: Testing The Predictive Ability of Various Exogenous Variables

Various factors usually affect air pollutant concentrations, making it difficult to achieve highly accurate predictions by relying only on air pollutant data from a single monitoring station. Most current research uses meteorological data and information from surrounding stations as influencing factors for predictive learning. However, simply incorporating information about surrounding stations into model inputs can be problematic. Namely, the spatial patterns of geographical phenomena and their variations are often influenced by uncontrollable factors, resulting in no correlation between the target station and the selected stations, thus introducing redundant information and negatively affecting the prediction performance. Therefore, when considering the selection of correlated stations in this study, the results based on the Third Law of Geography as well as spatial variability analysis are used to support the selection of the surrounding stations that are closely correlated to the target station (S9), specifically S2, S3, S8, S39, and S45, as shown in Figure 7. Figure 8 demonstrates the correlation between air pollutants, meteorological factors, and the correlated stations. As can be seen from the figure, the selected correlated stations have a strong correlation with PM_{2.5} concentrations.

In Experiment II, model7 and model8 are established to explore in depth the influence of meteorological factors and consideration of the Third Law of Geography on the predictive performance of the proposed approach. Specifically, in model7, the air pollutant concentrations at the target station and the air pollutant concentrations at the surrounding stations associated with it are used as input features to verify the effect of meteorological factors on the prediction performance. In model8, only air pollutant concentrations and meteorological factors at the target station are considered as input features, and information from the remaining stations that have synergistic variations with the target station is not included in the modeling based on the Third Law of Geography.

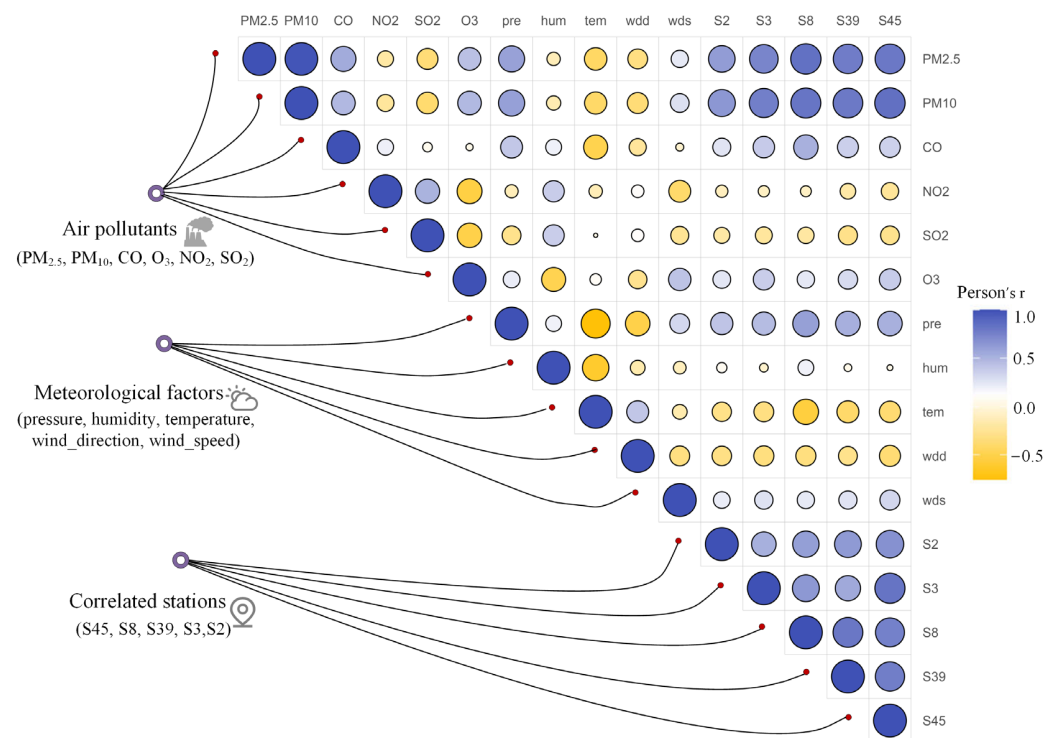


Figure 8. Correlations between the influencing factors.

The results obtained from the experiment are shown in Table 5. The prediction performance of the proposed approach (ours) is significantly better than that of model7 and model8. Taking the PM_{2.5} prediction results as an example, the RMSE and MAE are reduced by 7.7% and 11.23% on average, respectively, and the R² is improved by 1.27%. In particular, the results of model7 are better than those of model8, which indicates that utilizing the selected stations as features to aid prediction with a full consideration of the environmental similarity theory is superior to utilizing meteorological factors as features.

Table 5. Experimental results of the proposed approach and the baselines in Experiment II.

Model	PM _{2.5}			PM ₁₀		
	RMSE	MAE	R ²	RMSE	MAE	R ²
model7	2.193	1.507	0.946	3.455	2.351	0.939
model8	2.505	1.769	0.936	4.122	2.909	0.930
Ours	2.168	1.454	0.953	3.331	2.230	0.948

Figure 9 plots the prediction results of the proposed model, model7, and model8 for PM_{2.5}. Notably, there are relatively significant changes and frequent fluctuations in PM_{2.5} concentrations due to various environmental factors. Generally, all three methods can produce accurate overall trend predictions. However, for the sudden increase in the PM_{2.5} concentration, there may be a slight accuracy bias near the trend change point and a time lag phenomenon. By comparing the ability of the proposed model (ours), model7, and model8 in dealing with abrupt changes, it can be seen that the prediction trend of the proposed approach is more consistent with the actual observed shapes. The time lag phenomenon is weaker compared with the other two models, which indicates that the integrated consideration of meteorological factors and regional influences can deal with abrupt changes more effectively.

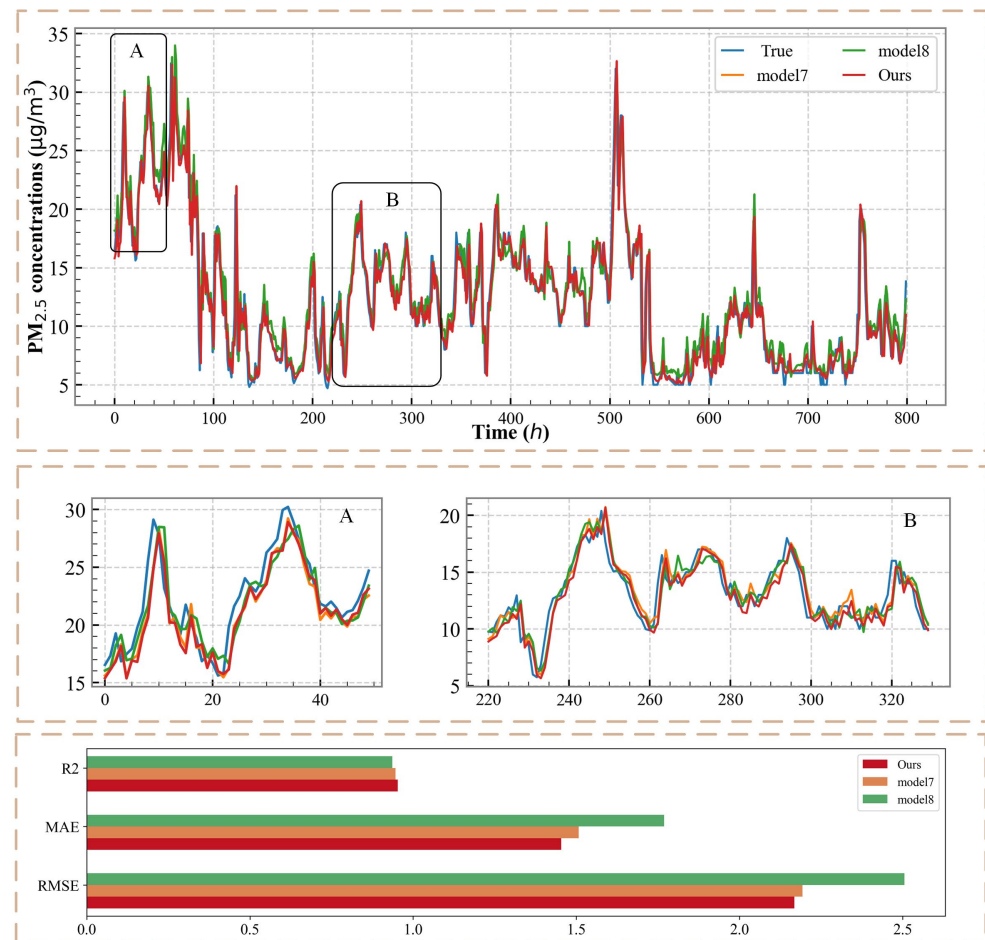


Figure 9. The comparison of the proposed method and the baselines in Experiment II. A and B are zoomed-in plots of each model's predicted results at time periods A and B.

4.6.3. Experiment III: Verifying The Different Modules of The Proposed Approach

The employed network modules such as spatial attention mechanisms, convolutional neural networks, and Transformer are thoroughly tested to assess their effectiveness in extracting input features. The experimental results are shown in Table 6. The results show that combining CNN with Transformer and SAM with Transformer improves the predictive learning performance compared to relying only on Transformer for predictive learning. The fusion of these three modules, i.e., the hybrid model proposed in this study, significantly outperforms model9 to model11. This proves that the proposed approach has a high degree of fitting ability with significant advantageous performance in air quality predictive learning verification.

Table 6. Experimental results of the different modules of the proposed approach in Experiment III.

Model	PM _{2.5}			PM ₁₀		
	RMSE	MAE	R ²	RMSE	MAE	R ²
model9	3.432	2.292	0.945	3.431	2.255	0.945
model10	2.173	1.458	0.950	3.426	2.286	0.945
model11	2.128	1.416	0.952	3.436	2.263	0.947
Ours	2.168	1.454	0.953	3.331	2.230	0.948

4.6.4. Experiment IV: Verifying The Generalization Ability of The Proposed Approach

To further evaluate the generalization ability of the proposed approach, one station is selected from Cluster 2, Cluster 3, and Cluster 4, respectively, for a series of experiments.

The datasets for the three stations are independent of each other. To maintain the experiment consistency, the same setup as in Experiment I is maintained in terms of data preprocessing, feature selection, model input, and output structure. It is clear from Table 7 that the two error indicators, RMSE and MAE, are relatively small for the three stations in the prediction of $PM_{2.5}$. The value of the R^2 indicator, which represents the accuracy, basically reaches results more than 90% (0.934, 0.933, 0.939), and the proposed model has obvious advantages compared with the other baseline models.

Table 7. Experimental results of the proposed approach and the baselines in Experiment IV.

Station	Model	$PM_{2.5}$			PM_{10}		
		RMSE	MAE	R^2	RMSE	MAE	R^2
S6	model1	5.749	2.452	0.814	9.137	5.849	0.801
	model2	6.405	5.597	0.638	8.909	5.613	0.811
	model3	4.113	2.561	0.895	5.439	3.314	0.927
	model4	3.670	2.241	0.918	5.207	2.991	0.932
	model5	4.102	2.417	0.862	5.303	3.414	0.921
	model6	4.461	2.588	0.871	7.647	4.850	0.868
	Ours	3.435	1.959	0.934	5.124	2.877	0.935
S25	model1	2.625	1.282	0.831	5.192	2.026	0.837
	model2	2.781	2.299	0.674	4.203	2.725	0.873
	model3	1.693	1.101	0.904	3.389	2.447	0.905
	model4	1.663	1.096	0.929	3.383	2.201	0.927
	model5	1.471	0.872	0.932	3.949	2.599	0.881
	model6	1.843	1.158	0.901	4.389	2.850	0.864
	Ours	1.444	0.853	0.933	2.897	1.687	0.939
S41	model1	3.121	2.368	0.879	4.237	4.237	0.797
	model2	3.584	2.739	0.832	11.567	7.540	0.654
	model3	2.879	2.018	0.893	10.099	5.188	0.634
	model4	2.417	1.623	0.935	9.073	4.697	0.746
	model5	2.363	1.518	0.922	10.004	5.674	0.638
	model6	2.698	1.825	0.907	9.282	5.080	0.739
	Ours	2.410	1.617	0.939	8.536	4.367	0.788

4.7. Shapley's Analysis

Shapley's analysis is a model interpretation method grounded in game theory principles. It assesses the importance of different features by examining their average contribution in all possible combinations. This analysis quantifies the importance of features into specific values, allowing us to gain insight into the relative contribution of each feature to the model's predictive learning. Compared to other interpretation methods, Shapley's analysis provides more accurate results by calculating feature weights from both local and global perspectives. In this study, Shapley's analysis is employed to comprehensively assess the importance of 15 characteristic variables including air pollutant concentrations, meteorological factors, and the correlated station factors in predicting $PM_{2.5}$ at the S9 station.

Figure 10 provides a localized view of Shapley's analysis for a single dataset sample. The vertical axis represents the input features, including PM_{10} , wind speed, S8, etc., and the horizontal coordinate represents the time window 1–11, where 1 represents the first 1 h of the current moment, and 11 represents the first 11 h. In the figure, the region corresponding to (x, y) reflects the impact of feature y on the model output in the first x hours; red indicates that the specified input feature positively affects the model output, while purple indicates the magnitude of the negative impact, with darker colors indicating larger impact values. The figure shows that PM_{10} , wind_speed, and S8 positively affect the model predictions, while O_3 negatively impacts the $PM_{2.5}$ concentration. In contrast, the predictive effect of CO, pressure, and S3 on the target is weak and can be ignored. Concurrently, as time passes,

the influence of features from periods farther apart gradually diminishes for the current moment. This indicates the diminishing impact of air quality conditions in past periods on the current moment.

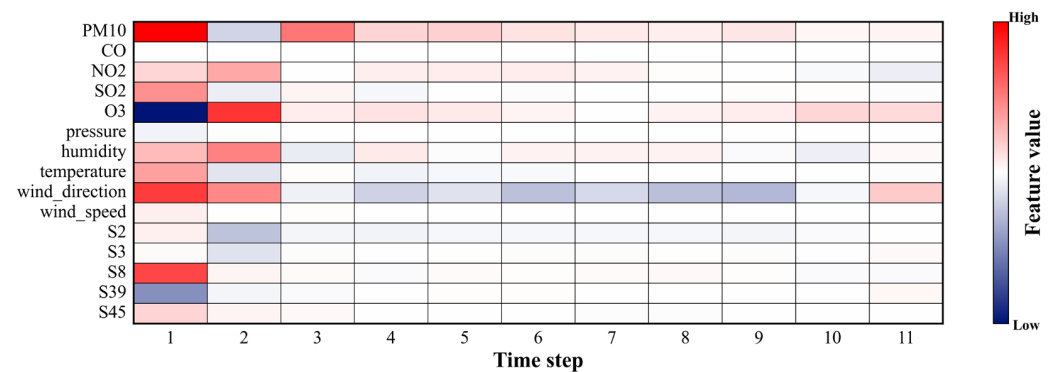


Figure 10. Influence of input features on $PM_{2.5}$ prediction (from local perspective).

Figure 11 shows a global view of Shapley's analysis for all dataset samples. The vertical axis represents the 15 different input features, the horizontal axis represents the weight of each feature's influence on the model output, and the right vertical axis is colored with different colors to indicate the high or low feature values. The results show that PM_{10} has the most significant effect on the prediction of $PM_{2.5}$, probably because they both belong to atmospheric particulate matter and have similar trends. The correlation analysis in Figure 8 also shows a strong correlation between PM_{10} and $PM_{2.5}$. In addition, wind speed and direction, as well as the S8 station, significantly impact $PM_{2.5}$ predictions, with changes in wind speed affecting the transmission distance of $PM_{2.5}$ concentrations at surrounding stations. Higher wind speeds usually encourage a faster dispersion of pollutants, while wind direction changes the direction of $PM_{2.5}$ transport. Combining the above analysis, the predictive power of the features through Shapley's analysis illustrates the rationality of this study based on the Third Law of Geography and combined with the spatial variability analysis to select the surrounding correlated stations as the model feature inputs.

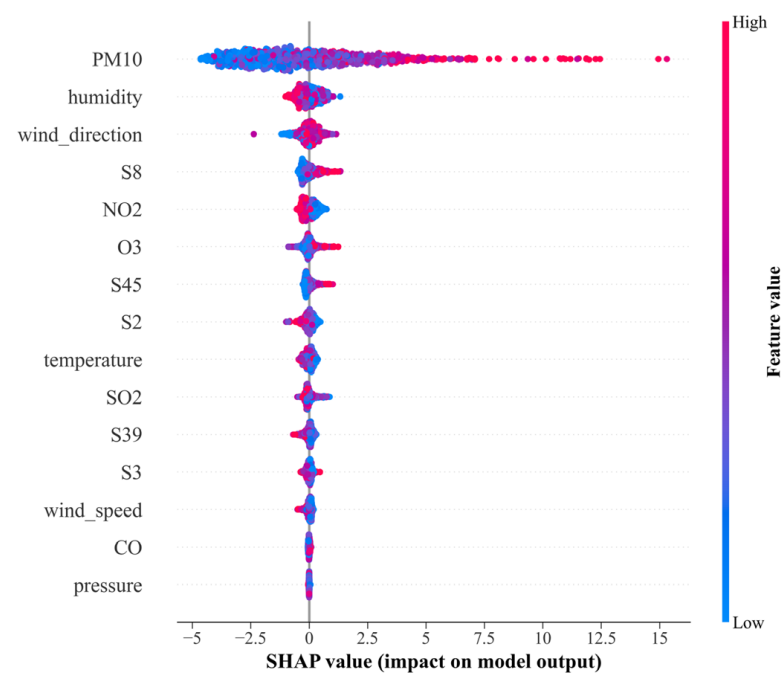


Figure 11. Influence of input features on $PM_{2.5}$ prediction (from global perspective).

5. Discussion

5.1. Analysis of The Impact of The Third Law of Geography in Predictive Learning

Sections 4.5.2 and 4.6 quantify the impact of multiple influencing factors on model predictive learning results through ablation experiments and Shapley's analysis, focusing on the role of the Third Law of Geography. The experimental results show that the stations selected based on this law, strongly correlated with the target pollutants, can effectively improve the prediction accuracy. Meanwhile, some meteorological factors also significantly affect the prediction of air pollutant concentrations. The experimental results align with common sense, as air pollutants are typically influenced by human activities, air circulation, and atmospheric circulation, leading to uncontrollable spatial variations and differences in pollutant concentration changes. The current experimental findings suggest that considering the Third Law of Geography and integrating multiple influencing factors positively contributes to the model's predictive performance. However, due to the current limitations of the data, this study still lacks universally applicable conclusions in the process of a spatial clustering analysis. Additionally, regarding the selection of target monitoring stations, the experiments in this study are conducted only for four stations (S9, S6, S25, S41). Further research on the remaining monitoring stations is needed, following the proposed approach in this study. How to conduct experiments on all monitoring stations and quantify the degree of influence of each influencing factor on the model prediction results for different types of monitoring stations is an important direction for future research in this study.

5.2. Impact of Different Clustering Algorithms

This study utilizes POI data and employs the hierarchical clustering method to categorize all monitoring stations into four distinct clusters, providing valuable information for selecting correlated stations. As an unsupervised learning algorithm, the hierarchical clustering method is based on grouping similar samples into clusters by measuring their similarity. Among the factors influencing the effectiveness of hierarchical clustering algorithms, the calculation method for distance and the metric used for similarity measurement are particularly crucial.

Different distance calculation methods will lead to different segmentation results in the subsequent similarity calculation. To consider the standardized differences in each dimension comprehensively, this study chooses the Normalized Euclidean distance as the distance metric to improve the robustness to noise and outliers in spatial data. Regarding the similarity measurement method, the average-linkage method is employed, which is relatively less sensitive to outliers than single-linkage and complete-linkage methods, which contributes to obtaining more practically interpretable clustering results.

Compared to other clustering algorithms, the K-means algorithm requires the pre-specification of the number of clusters to be divided as well as its sensitivity to the initial cluster centers, which may lead to instability in the results. DBSCAN is suitable for clustering with variable density and requires more tuning for its parameter selection and noise handling. Spectral clustering is suitable for dealing with graphical data and flow structures. Hierarchical clustering has the advantage of not requiring a pre-specified number of clusters, automatically forming a hierarchical structure of clusters, and being more robust to noise and outliers in the data. Therefore, the hierarchical clustering algorithm is ideal for the spatial clustering of air quality monitoring stations in this study.

5.3. Advantages of The Proposed Approach

In terms of modeling, the currently popular methods include models based on CNN and models based on LSTM. However, models based on CNN are limited to capturing features in localized areas, resulting in irregular topological information in pollutant monitoring networks that is difficult to fully mine [25]. Models based on LSTM still have limitations in modeling long-range dependencies. The SAM-CNN-Transformer model proposed in this study combines different deep learning models to overcome the

problems of the above models by utilizing the advantages of each model. A spatial attention mechanism (SAM) implements self-attention computation in the spatial dimension to extract spatial features embedded in the data. CNN is utilized to mine the local information between the relationships of each input feature, and the Transformer is employed to deeply capture the temporal dependency between long-distance time series. Compared with other deep learning models, this study introduces the spatial attention mechanism and Transformer to enhance the capability to extract spatial-temporal features effectively.

6. Conclusions and Future Directions

6.1. Summary of Experimental Results

In this work, particulate pollutant data, gaseous pollutant data, and meteorological data from the target stations are used as characterizing variables. In addition, according to the Third Law of Geography, information about the surrounding stations is introduced by spatial clustering and a spatial variability analysis of POI, which shows synergistic variations with the pollutant concentrations at the target stations. Meanwhile, an innovative air quality predictive learning method is proposed by integrating the advantages of SAM, CNN, and Transformer.

The proposed approach is implemented as follows: Firstly, in the aspect of model factor feature selection, based on the Third Law of Geography, the hierarchical clustering algorithm is used to cluster all monitoring stations into four classes, and air pollutants, meteorological factors, and pollutant concentrations of neighboring stations are used as characterization variables based on the clustering area where the target stations are located and the spatial anisotropy of air pollutant concentrations. Secondly, in terms of modeling, SAM, CNN, and Transformer technology are combined to fully explore the spatial and temporal dependence and complex relationship of air pollutant concentrations. Finally, regarding the model interpretability analysis, Shapley's analysis is used to analyze the importance of each influencing factor in model predictive learning, providing direction for further modeling.

In order to verify the performance of the proposed approach, Haikou City is selected as the study area, and four groups of comparative experiments are designed with RMSE, MAE, and R^2 as the evaluation indexes. The results of Experiment I and Experiment III show that the proposed approach achieves optimal accuracy and stability compared to the other baseline models, highlighting its effectiveness. The results of Experiment II verify that meteorological factors and the consideration of synergistic changes between regions based on the Third Law of Geography can effectively improve the model's predictive learning accuracy. Experiment IV trains the model on different stations with different data features, and the effective predictive learning obtained proves the model's generalizability.

In summary, the above experimental results and analysis demonstrate that the proposed approach performs better in the air quality predictive learning task with good prediction generalization ability and robustness, providing a new direction for air quality predictive learning.

6.2. Caveats and Future Directions

There are still some limitations in this study. Firstly, this study has only performed predictive modeling from a data-driven perspective without considering air pollutants' physical and chemical mechanisms. Therefore, combining a priori knowledge of domain experts with the proposed approach in this study is a promising research direction. In the follow-up work, we will address the above research directions, and the air quality predictive learning method currently proposed in this study will play an essential role in the follow-up work. Secondly, it is necessary to explore the impact of different clustering methods on spatial clustering results in depth and investigate a universally applicable spatial clustering method to improve the accuracy of monitoring station classification. Furthermore, conducting experiments on all monitoring stations in Haikou and other cities and quantifying the impact of each influencing factor on the model's predictive results for different types of monitoring stations is also an important direction.

Author Contributions: Conceptualization, X.L. and Y.Z.; methodology, Y.Z.; software, Y.Z.; formal analysis, S.L. and X.L.; investigation, X.Z. and X.F.; resources, S.L.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, S.L., X.L. and Q.M.; visualization, Y.Z. and X.F.; supervision, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key R&D Program, grant number 2020YFB2104400.

Data Availability Statement: The datasets generated and analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Acknowledgments: The authors would like to express sincere gratitude to the anonymous reviewers for their valuable feedback and constructive comments on our manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, X.; Han, L.; Wei, H.; Tan, X.; Zhou, W.; Li, W.; Qian, Y. Linking Urbanization and Air Quality Together: A Review and a Perspective on the Future Sustainable Urban Development. *J. Clean. Prod.* **2022**, *346*, 130988. [\[CrossRef\]](#)
2. Zhu, J.; Chen, L.; Liao, H. Multi-Pollutant Air Pollution and Associated Health Risks in China from 2014 to 2020. *Atmos. Environ.* **2022**, *268*, 118829. [\[CrossRef\]](#)
3. Duan, R.-R.; Hao, K.; Yang, T. Air Pollution and Chronic Obstructive Pulmonary Disease. *Chronic Dis. Transl. Med.* **2020**, *6*, 260–269. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Glencross, D.A.; Ho, T.-R.; Camiña, N.; Hawrylowicz, C.M.; Pfeffer, P.E. Air Pollution and Its Effects on the Immune System. *Free Radic. Biol. Med.* **2020**, *151*, 56–68. [\[CrossRef\]](#)
5. Zhang, J.; Li, S. Air Quality Index Forecast in Beijing Based on CNN-LSTM Multi-Model. *Chemosphere* **2022**, *308*, 136180. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Fang, W.; Zhu, R.; Lin, J.C.-W. An Air Quality Prediction Model Based on Improved Vanilla LSTM with Multichannel Input and Multiroute Output. *Expert Syst. Appl.* **2023**, *211*, 118422. [\[CrossRef\]](#)
7. Seng, D.; Zhang, Q.; Zhang, X.; Chen, G.; Chen, X. Spatiotemporal Prediction of Air Quality Based on LSTM Neural Network. *Alex. Eng. J.* **2021**, *60*, 2021–2032. [\[CrossRef\]](#)
8. Mao, W.; Jiao, L.; Wang, W.; Wang, J.; Tong, X.; Zhao, S. A Hybrid Integrated Deep Learning Model for Predicting Various Air Pollutants. *GIScience Remote Sens.* **2021**, *58*, 1395–1412. [\[CrossRef\]](#)
9. Carreño, G.; López-Cortés, X.A.; Marchant, C. Machine Learning Models to Predict Critical Episodes of Environmental Pollution for PM_{2.5} and PM₁₀ in Talca, Chile. *Mathematics* **2022**, *10*, 373. [\[CrossRef\]](#)
10. Wu, C.; He, H.; Song, R.; Zhu, X.; Peng, Z.; Fu, Q.; Pan, J. A Hybrid Deep Learning Model for Regional O₃ and NO₂ Concentrations Prediction Based on Spatiotemporal Dependencies in Air Quality Monitoring Network. *Environ. Pollut.* **2023**, *320*, 121075. [\[CrossRef\]](#)
11. Zhang, B.; Liu, Y.; Yong, R.; Zou, G.; Yang, R.; Pan, J.; Li, M. A Spatial Correlation Prediction Model of Urban PM_{2.5} Concentration Based on Deconvolution and LSTM. *Neurocomputing* **2023**, *544*, 126280. [\[CrossRef\]](#)
12. Tong, Y.; Luo, K.; Li, R.; Pei, L.; Li, A.; Yang, M.; Xu, Q. Association between Multi-Pollutant Mixtures Pollution and Daily Cardiovascular Mortality: An Exploration of Exposure-Response Relationship. *Atmos. Environ.* **2018**, *186*, 136–143. [\[CrossRef\]](#)
13. Lin, M.-D.; Liu, P.-Y.; Huang, C.-W.; Lin, Y.-H. The Application of Strategy Based on LSTM for the Short-Term Prediction of PM_{2.5} in City. *Sci. Total Environ.* **2024**, *906*, 167892. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Yu, Y.; Li, H.; Sun, S.; Li, Y. PM_{2.5} Concentration Forecasting through a Novel Multi-Scale Ensemble Learning Approach Considering Intercity Synergy. *Sustain. Cities Soc.* **2022**, *85*, 104049. [\[CrossRef\]](#)
15. Tobler, W.R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **1970**, *46*, 234–240. [\[CrossRef\]](#)
16. Zhu, A.-X.; Turner, M. How Is the Third Law of Geography Different? *Ann. GIS* **2022**, *28*, 57–67. [\[CrossRef\]](#)
17. Sui, S.; Han, Q. Multi-View Multi-Task Spatiotemporal Graph Convolutional Network for Air Quality Prediction. *Sci. Total Environ.* **2023**, *893*, 164699. [\[CrossRef\]](#)
18. Zhang, C.; Hu, Y.; Adams, M.D.; Liu, M.; Li, B.; Shi, T.; Li, C. Natural and Human Factors Influencing Urban Particulate Matter Concentrations in Central Heating Areas with Long-Term Wearable Monitoring Devices. *Environ. Res.* **2022**, *215*, 114393. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Nurcahyanto, H.; Prihatno, A.T.; Alam, M.M.; Rahman, M.H.; Jahan, I.; Shahjalal, M.; Jang, Y.M. Multilevel RNN-Based PM₁₀ Air Quality Prediction for Industrial Internet of Things Applications in Cleanroom Environment. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, e1874237. [\[CrossRef\]](#)
20. Zhu, J.; Deng, F.; Zhao, J.; Zheng, H. Attention-Based Parallel Networks (APNet) for PM_{2.5} Spatiotemporal Prediction. *Sci. Total Environ.* **2021**, *769*, 145082. [\[CrossRef\]](#)
21. Zou, X.; Zhao, J.; Zhao, D.; Sun, B.; He, Y.; Fuentes, S. Air Quality Prediction Based on a Spatiotemporal Attention Mechanism. *Mob. Inf. Syst.* **2021**, *2021*, e6630944. [\[CrossRef\]](#)

22. Ng, W.; Minasny, B.; McBratney, A. Convolutional Neural Network for Soil Microplastic Contamination Screening Using Infrared Spectroscopy. *Sci. Total Environ.* **2020**, *702*, 134723. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Kabir, S.; Islam, R.U.; Hossain, M.S.; Andersson, K. An Integrated Approach of Belief Rule Base and Convolutional Neural Network to Monitor Air Quality in Shanghai. *Expert Syst. Appl.* **2022**, *206*, 117905. [\[CrossRef\]](#)
24. Qi, Y.; Li, Q.; Karimian, H.; Liu, D. A Hybrid Model for Spatiotemporal Forecasting of PM_{2.5} Based on Graph Convolutional Neural Network and Long Short-Term Memory. *Sci. Total Environ.* **2019**, *664*, 1–10. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Yan, R.; Liao, J.; Yang, J.; Sun, W.; Nong, M.; Li, F. Multi-Hour and Multi-Site Air Quality Index Forecasting in Beijing Using CNN, LSTM, CNN-LSTM, and Spatiotemporal Clustering. *Expert Syst. Appl.* **2021**, *169*, 114513. [\[CrossRef\]](#)
26. Wu, S.; Xiao, X.; Ding, Q.; Zhao, P.; Wei, Y.; Huang, J. Adversarial Sparse Transformer for Time Series Forecasting. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17105–17115.
27. Wang, J.; Song, G. A Deep Spatial-Temporal Ensemble Model for Air Quality Prediction. *Neurocomputing* **2018**, *314*, 198–206. [\[CrossRef\]](#)
28. Goodchild, M.F. GIScience, Geography, Form, and Process. *Ann. Assoc. Am. Geogr.* **2004**, *94*, 709–714. [\[CrossRef\]](#)
29. Zhao, F.-H.; Huang, J.; Zhu, A.-X. Spatial Prediction of Groundwater Level Change Based on the Third Law of Geography. *Int. J. Geogr. Inf. Sci.* **2023**, *37*, 2129–2149. [\[CrossRef\]](#)
30. Hou, J.; Wang, Y.; Hou, B.; Zhou, J.; Tian, Q. Spatial Simulation and Prediction of Air Temperature Based on CNN-LSTM. *Appl. Artif. Intell.* **2023**, *37*, 2166235. [\[CrossRef\]](#)
31. Mao, Y.; Lee, S. Deep Convolutional Neural Network for Air Quality Prediction. *J. Phys. Conf. Ser.* **2019**, *1302*, 032046. [\[CrossRef\]](#)
32. Xiao, X.; Jin, Z.; Wang, S.; Xu, J.; Peng, Z.; Wang, R.; Shao, W.; Hui, Y. A Dual-Path Dynamic Directed Graph Convolutional Network for Air Quality Prediction. *Sci. Total Environ.* **2022**, *827*, 154298. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Wang, Z.; Wu, F.; Yang, Y. Air Pollution Measurement Based on Hybrid Convolutional Neural Network with Spatial-and-Channel Attention Mechanism. *Expert Syst. Appl.* **2023**, *233*, 120921. [\[CrossRef\]](#)
34. Fong, I.H.; Li, T.; Fong, S.; Wong, R.K.; Tallón-Ballesteros, A.J. Predicting Concentration Levels of Air Pollutants by Transfer Learning and Recurrent Neural Network. *Knowl.-Based Syst.* **2020**, *192*, 105622. [\[CrossRef\]](#)
35. Wang, B.; Kong, W.; Zhao, P. An Air Quality Forecasting Model Based on Improved Convnet and RNN. *Soft Comput.* **2021**, *25*, 9209–9218. [\[CrossRef\]](#)
36. Chen, Y.-C.; Lei, T.-C.; Yao, S.; Wang, H.-P. PM_{2.5} Prediction Model Based on Combinational Hammerstein Recurrent Neural Networks. *Mathematics* **2020**, *8*, 2178. [\[CrossRef\]](#)
37. Liu, X.; Du, H.; Yu, J. A Forecasting Method for Non-Equal Interval Time Series Based on Recurrent Neural Network. *Neurocomputing* **2023**, *556*, 126648. [\[CrossRef\]](#)
38. Jin, X.-B.; Yang, N.-X.; Wang, X.-Y.; Bai, Y.-T.; Su, T.-L.; Kong, J.-L. Deep Hybrid Model Based on EMD with Classification by Frequency Characteristics for Long-Term Air Quality Prediction. *Mathematics* **2020**, *8*, 214. [\[CrossRef\]](#)
39. Liang, Y.; Xia, Y.; Ke, S.; Wang, Y.; Wen, Q.; Zhang, J.; Zheng, Y.; Zimmermann, R. AirFormer: Predicting Nationwide Air Quality in China with Transformers. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 14329–14337. [\[CrossRef\]](#)
40. Van Houdt, G.; Mosquera, C.; Nápoles, G. A Review on the Long Short-Term Memory Model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [\[CrossRef\]](#)
41. Zhou, X.; Xu, J.; Zeng, P.; Meng, X. Air Pollutant Concentration Prediction Based on GRU Method. *J. Phys. Conf. Ser.* **2019**, *1168*, 032058. [\[CrossRef\]](#)
42. Liu, Q.; Li, J.; Lu, Z. ST-Tran: Spatial-Temporal Transformer for Cellular Traffic Prediction. *IEEE Commun. Lett.* **2021**, *25*, 3325–3329. [\[CrossRef\]](#)
43. Reza, S.; Ferreira, M.C.; Machado, J.J.M.; Tavares, J.M.R.S. A Multi-Head Attention-Based Transformer Model for Traffic Flow Forecasting with a Comparative Analysis to Recurrent Neural Networks. *Expert Syst. Appl.* **2022**, *202*, 117275. [\[CrossRef\]](#)
44. Pundir, I.; Aggarwal, N.; Singh, S. Time-Series Based Prediction of Air Quality Index Using Various Machine Learning Models. In Proceedings of the Decision Intelligence Solutions; Hasteer, N., McLoone, S., Khari, M., Sharma, P., Eds.; Springer Nature: Singapore, 2023; pp. 61–70.
45. Bekkar, A.; Hssina, B.; Douzi, S.; Douzi, K. Air-Pollution Prediction in Smart City, Deep Learning Approach. *J. Big Data* **2021**, *8*, 161. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Wen, C.; Liu, S.; Yao, X.; Peng, L.; Li, X.; Hu, Y.; Chi, T. A Novel Spatiotemporal Convolutional Long Short-Term Neural Network for Air Pollution Prediction. *Sci. Total Environ.* **2019**, *654*, 1091–1099. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Liu, X.; Qin, M.; He, Y.; Mi, X.; Yu, C. A New Multi-Data-Driven Spatiotemporal PM_{2.5} Forecasting Model Based on an Ensemble Graph Reinforcement Learning Convolutional Network. *Atmos. Pollut. Res.* **2021**, *12*, 101197. [\[CrossRef\]](#)
48. Huang, Y.; Ying, J.J.-C.; Tseng, V.S. Spatio-Attention Embedded Recurrent Neural Network for Air Quality Prediction. *Knowl.-Based Syst.* **2021**, *233*, 107416. [\[CrossRef\]](#)
49. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.