sustainability

MDPI

Article

Assessing the Effects of Urban Morphology Parameters on $PM_{2.5}$ Distribution in Northeast China Based on Gradient Boosted Regression Trees Method

Peng Cui *, Chunyu Dai, Jun Zhang and Tingting Li

School of Landscape Architecture, Northeast Forestry University, Harbin 150040, China; daichunyu0629@nefu.edu.cn (C.D.); zhangjun-nefu@nefu.edu.cn (J.Z.); may_lting@nefu.edu.cn (T.L.)
* Correspondence: cuipeng@nefu.edu.cn; Tel.: +86-133-5180-6160

Abstract: The dispersion of urban pollutants is affected by the urban morphology parameters. The objective of this study was to investigate the correlation between $PM_{2.5}$ distribution and urban morphology parameters in a cold-climate city in China. Field measurements were performed to record the $PM_{2.5}$ concentration and microclimate parameters at 25 points in a 10 km^2 urban area in Harbin, China. It was found that the maximum difference of $PM_{2.5}$ concentration among the measuring points at the same time could be up to 69.03 $\mu g/m^3$. In this study, a geographic information system (GIS) was used to extract and screen the urban morphology parameter data under reasonable buffer radius, the gradient boosted regression trees model (GBRT) was used to carry out the prediction experiment of $PM_{2.5}$ concentration and explore the nonlinear influence of urban morphology factors on $PM_{2.5}$ concentration. In addition, random forest (RF), decision trees (DT), and multiple linear regression (MLR) models were selected to compare the prediction accuracy of the GBRT model. The results show that the GBRT model has the highest accuracy, with R^2 reaching 0.981; building density (57%) and average building height (49%) were the two most significant factors affecting $PM_{2.5}$ concentration.

Keywords: urban morphology parameters; $PM_{2.5}$ distribution; gradient boosted regression trees (GBRT) model; northeast China

check for updates

Citation: Cui, P.; Dai, C.; Zhang, J.; Li, T. Assessing the Effects of Urban Morphology Parameters on $PM_{2.5}$ Distribution in Northeast China Based on Gradient Boosted Regression Trees Method. *Sustainability* **2022**, *14*, 2618. https://doi.org/10.3390/su14052618

Academic Editor: Elena Cristina Rada

Received: 15 January 2022
Accepted: 18 February 2022
Published: 24 February 2022

1. Introduction

$PM_{2.5}$ refers to the particulate matter in the atmosphere with a diameter of 2.5 μm or less, often called lungable particulate matter or fine particulate matter. Due to its small particle size and a large number of toxic and harmful substances, $PM_{2.5}$ can easily cause health damage like respiratory diseases and pulmonary fibrosis to the human body [1]. Therefore, $PM_{2.5}$ has become one of the most important targets for environmental pollution prevention and control in the world. There are differences in the $PM_{2.5}$ situation in different regions because factors such as climate and urban morphology will have an impact on the formation and dispersion of $PM_{2.5}$ [2]. It is very important to understand the spatial distribution and related dynamic changes of $PM_{2.5}$, which is conducive to formulating effective measures to reduce and control the harm caused by $PM_{2.5}$ combined with the actual situation. $PM_{2.5}$ pollution is very serious in cold-climate cities of northeast China. In addition to common forms of urban air pollution such as long-distance transportation of pollutants and automobile exhaust emissions, winter heating and inversion layer aggravate the problem of declining urban air quality and frequent haze weather [3,4]. It is worth noting that the demand for wind and cold protection of cold-climate cities also forces the urban morphology design to be relatively simple and closed, which is not conducive to the dispersion of air pollutants [5]. Therefore, it is of great practical significance to study the distribution of $PM_{2.5}$ in cold-climate cities of northeast China.

The average wind speed and static wind frequency are the main factors affecting $PM_{2.5}$ dispersion [6,7]. The influence of urban morphology on $PM_{2.5}$ is mainly reflected in two aspects: first, block layout affects the change of temperature and humidity inside the area, which indirectly affects the condensation and precipitation of air particles [8]. On the other hand, block layout also affects the wind environment inside the region, which directly affects the flow and dispersion of air pollutants [9]. Longley I. D. [10] points out that wind speed and the relative direction of the street are decisive factors affecting the spatial distribution of $PM_{2.5}$. When the wind direction is parallel to the street, it is conducive to the dispersion of $PM_{2.5}$. By studying the volume relationship between blocks and buildings, Oke [11] found that wind pressure generated under different block layouts had different influences on the dispersion of particulate matter and that different street aspect ratios would produce different spatial vortex structures in the street valley, thus forming different $PM_{2.5}$ dispersion conditions. Kaplan [12] took a small-scale block as an example to simulate the distribution of particulate matter, which confirmed the scientific nature of the combination of data simulation technology and field monitoring method. Chan, L. Y. [13] took the long and narrow streets of Hong Kong as an example to study the influence of its spatial volume on the concentration of $PM_{2.5}$ and other particulates and concluded that the aspect ratio of street space is positively correlated with the concentration of particulates in street valleys.

At present, there are two main prediction methods for $PM_{2.5}$ concentration: deterministic models and empirical models. Deterministic models are represented by weather research and forecasting (WRF) and community multiscale air quality (CMAQ). However, deterministic models have limited the analysis of air quality at micro scales. Among empirical approaches, linear regression models and machine learning methods have received more attention. Multiple regression is to establish a regression model about the predicted object through several influencing factors such as meteorological factors, pollution sources, and land use environment. As a traditional prediction method of air pollutants, this method can be used to fit and predict pollutant mass concentration or pollution index through regression modeling. For example, Ziomasic et al. [14] established a multiple linear regression model based on seven meteorological factors to predict the maximum mass concentration of NO_2 in Athens, Greece. Machine learning uses multiple disciplines, such as probability theory and mathematical statistics, approximation theory, convex analysis, and algorithm complexity, to extract certain rules or patterns from raw data and then output prediction information which is widely used in recent years. Kukkonen J. et al. [15] used the neural network model to predict the concentrations of PM_{10} and NO_2 at two points in Helsinki, Finland, by taking traffic flow and meteorological factors as predictors. Mckendry I. [16] used the artificial neural network model to predict the daily maximum and average values of O_3, PM_{10}, and $PM_{2.5}$ mass concentrations by taking meteorological factors and pollutant mass concentrations as predictors. On the basis of having accurate meteorological parameters as the input data, all the above studies have achieved good prediction results. In general, the traditional multiple linear regression model is simple and intuitive, which can quickly analyze the linear relationship between each parameter, and determine the influence degree of the influence factor on the predicted object through correlation. However, in a real urban situation, the prediction environment of air pollutants is very complex, and there may be a strong nonlinear relationship between air pollutants and the predictors, which leads to great limitations of the multiple linear regression model to predict results. Machine learning algorithms obviously show great superiority in solving nonlinear model problems [17] and support vector machine [18], multi-layer perceptron [19], and sequence learning [20] have been applied to air pollution research and proved to perform well, but they cannot rank the influencing variables based on their importance, which cannot provide a basis for further pollution control and prevention.

The decision tree model (DT) is resistant to this potential problem, it can not only learn decision rules from data features to predict the value of target variables [21] but also can identify the relationship between response variables and predictive variables [22].

The gradient boosted regression tree (GBRT) model was developed on the basis of the DT model, further enhancing the stability and accuracy of prediction. It is widely used in big data mining research due to its own certain interpretability, accuracy, and efficiency. The model also shows stronger robustness and generalization ability when dealing with complex related variables [23].

In summary, a large number of research scholars focused their research on the macro level of the entire city and concluded that the concentration of $PM_{2.5}$ is mainly affected by various pollution sources and meteorological conditions [24,25]. In addition, fixed-point monitoring is widely used in the world to obtain the $PM_{2.5}$ pollution status [15]. However, the observation results of each monitoring point can only represent the $PM_{2.5}$ concentration within a certain radius around the monitoring point, while the monitoring points in the city are sparse, which can only reveal the $PM_{2.5}$ pollution level within a small space, and cannot represent the $PM_{2.5}$ pollution status and spatial difference of the whole city. In order to facilitate the public to understand the local air quality and help the government to take measures to prevent and control $PM_{2.5}$ pollution, it is necessary to analyze and predict $PM_{2.5}$ concentration at the block scale.

The problem of winter haze pollution in cold-climate cities of northeast China is very serious. Therefore, Harbin, a typical cold-climate city, ranked among the top ten cities with the worst air quality in China, was taken as an experimental case. This study plans to achieve the following goals: (1) to illustrate the spatial and temporal distribution of $PM_{2.5}$ concentration in block scale; (2) to analyze the influence of urban microclimate on $PM_{2.5}$ concentration and the influence radius of urban morphology parameters; (3) to establish a prediction model of $PM_{2.5}$ concentration in urban blocks of cold climate based on the gradient boosted regression trees model and verify its effectiveness; (4) to study the influence degree of different urban morphologies on $PM_{2.5}$ concentration and give advice on urban planning for a better environment.

2. Methodology

2.1. Study Area

Harbin (125°42′–130°44′ E longitude, 10°04′–46°40′ N latitude) is the capital of Heilongjiang Province, China with long winters, short and dry summers, and relatively short spring and autumn seasons. The special climate results in a heating period that lasts for half a year and huge consumption of fossil fuels. With the improvement of residents' quality of life, the consumption of fossil energy and the number of motor vehicles in Harbin have been increasing in recent years. In addition, a series of factors, such as excessive and substandard emissions of coal-fired exhaust gas, automobile exhaust emissions, straw burning, and long-distance transportation of pollutants all have led to the decline of air quality and frequent haze weather [26]. According to the data released by the local meteorological department, during the heating period, coal burning and industrial and secondary sources are important sources of $PM_{2.5}$ in Harbin, accounting for 25%, 20%, and 19% respectively, followed by traffic, dust, and biomass burning, as shown in Figure 1.

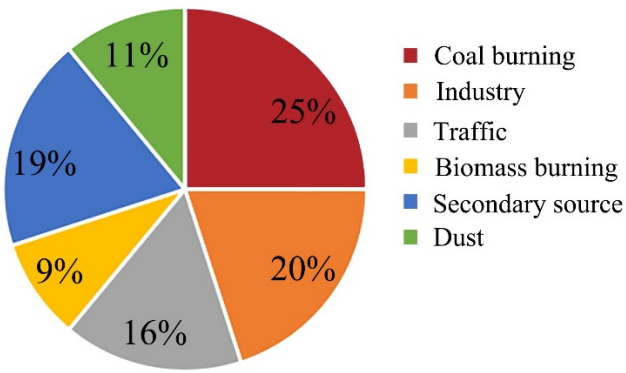

Figure 1. Sources of $PM_{2.5}$ pollution during heating period in Harbin.

The changes in $PM_{2.5}$ concentration in Harbin from January 2019 to January 2021 are shown in Figure 2. According to the requirements of China's Environmental Air Quality Standard (GB3095-2012), residential and commercial mixed areas, residential areas, etc., should meet the second level of $PM_{2.5}$ concentration limit, that is, the 24 h average concentration is below 75 μg/m^3. However, in January, February, and December 2019, the average daily $PM_{2.5}$ concentration exceeded 75 μg/m^3 for 9, 13, and 13 days respectively. In January, February, and December 2020, the number of exceeded days was 26, 5, and 7 days respectively. In January 2021, the number of exceeded days was 17 days. In addition, the monthly average $PM_{2.5}$ concentration increased significantly in April 2020, which was caused by straw burning in the surrounding countryside. It can be found that the months with excessive $PM_{2.5}$ concentration were mainly concentrated in the heating season of every year. Therefore, the $PM_{2.5}$ pollution situation in Harbin in January, February, and December was selected for this study.

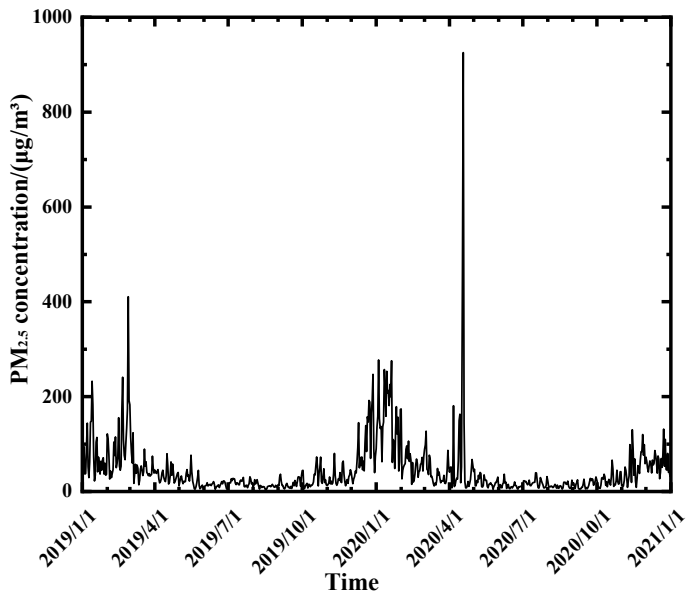

Figure 2. Changes in $PM_{2.5}$ concentration in Harbin in recent years.

In order to analyze the impact of urban morphology factors on the dispersion of $PM_{2.5}$ at the block scale, a study area with diverse spatial attributes should be selected. As shown in Figure 3, Harbin Central Street covers an area of 10.1 km^2 with a perimeter of 6.5 km and adopts an open block layout was selected as the research field. It covers pedestrian streets, shopping malls, squares, residential areas, small parks, and other urban activity spaces. Moreover, the building density in the region is high, the vegetation coverage is moderate, and the block types are diverse, including multi-story and high-rise buildings, which are suitable for research.

2.2. Measurement of $PM_{2.5}$ and Microclimate Parameter

Different building forms and complex road networks result in great differences in $PM_{2.5}$ concentration [27]. The micro-scale spatial variability of $PM_{2.5}$ concentration cannot be effectively observed by the fixed air quality monitoring stations in Harbin, so more intensive monitoring points need to be manually arranged. According to previous studies, there are differences in the arrangement of measuring points, and there is no strict method to determine the number of measuring points, so the principle of measuring points should be as comprehensive and abundant as possible. Combined with the actual situation and the demand of influence radius, 25 monitoring sites were selected for synchronous measurement with a monitoring density of 0.4 km^2. The arrangement of measuring points is shown in Figure 3.

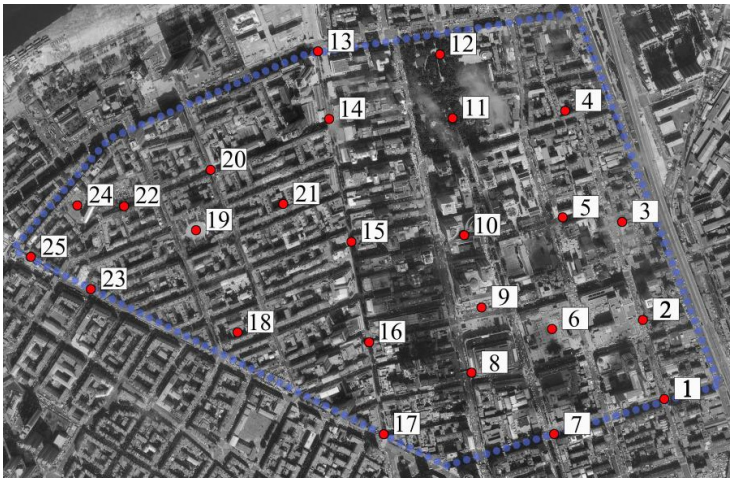

Figure 3. The layout of the study area and measurement points.

The measured parameters include the $PM_{2.5}$ concentration at each measurement point and the microclimate parameters including temperature (T), humidity (RH), and wind speed (WIND). Twenty-five sets of portable monitors were used to detect pollutants at different monitoring sites. Each set contains a DylosDC1700 particle detector, an NK5500 weather station, and a tripod. Recent literature on measurements has confirmed that the DylosDC1700 particle detector and NK5500 weather station perform well after reasonable calibration to investigate the small-scale spatial variability of $PM_{2.5}$ personal exposure and assess the effect of environmental features [8,28,29]. Related information such as instrument precision is listed in Table 1. The measuring instruments were placed on a tripod with a height of 1.5 m to obtain pedestrian height data. As shown in Figure 4, measurement point No.6, located in the center of St. Sophia Cathedral Square, is selected to show the field measurement.

The measurement was carried out from December 2020 to February 2021. In order to eliminate the interference of snow and other factors, the experiments were carried out in clear days and abnormal data were removed. Finally, a total of 21 days were selected for research. The chosen days contain different air pollution conditions of light, moderate, and heavy haze issued by the Meteorological Observatory. The specific selected test dates and their morning, middle, and evening meteorological conditions are listed in Table 2. In this study, round-the-clock monitoring was conducted and hourly data were recorded at each site. According to different research purposes, the measured data are processed, which are mainly divided into the following three parts:

(1) Observe the temporal and spatial variation of $PM_{2.5}$ concentration. The hourly $PM_{2.5}$ concentration data of each measuring point for 21 days were collected, and then the hourly average was calculated to observe the temporal distribution characteristics of $PM_{2.5}$ concentration. The $PM_{2.5}$ concentration data of each measuring point at 10:00 and 22:00 for 21 days were collated, and then the mean value of these two times was calculated to observe the spatial distribution characteristics of $PM_{2.5}$ concentration.
(2) Observe the influence of urban microclimate on $PM_{2.5}$ concentration. According to the temporal distribution characteristics of $PM_{2.5}$ concentration, the typical moments when $PM_{2.5}$ concentration changes were selected. The $PM_{2.5}$ concentration and microclimate data at the corresponding moments of each measuring point for 21 days were collected, and then the average value at the corresponding moments was calculated to observe the influence of microclimate change on $PM_{2.5}$ concentration.
(3) Collect data for predictive model training and validation. The hourly $PM_{2.5}$ concentration and microclimate data of each measuring point for 21 days were collected and combined with the subsequent urban morphology and other related data, finally, 12,600 sets of data were obtained, and then the training and verification of the prediction model was carried out.

Table 1. Technical parameters of measuring instruments.

Name	Usage	Technical Parameter
NK5500 weather station	Wind speed, Temperature, Humidity	Wind speed measurement range is 0.6–60 m/s, accuracy is ±3%, 1 inch | 25 mm diameter impeller with precision axle and low-friction Zytel® bearings; Temperature measurement range is −29–70 °C, accuracy ±0.5 °C, platinum resistance temperature sensor; Humidity measurement range is 0–100%, accuracy is ±2%, polymeric capacitance humidity sensor. The measurement range is the number of particles in the air per 0.01 cubic feet of volume. The unit is μg/m^3. Laser scattering method.
DylosDC1700 particle detector	$PM_{2.5}$ concentration	Two kinds of particles of 0.5 μm and 2.5 μm can be detected. This value divided by 100 is the mass concentration of $PM_{2.5}$, commonly used in China.

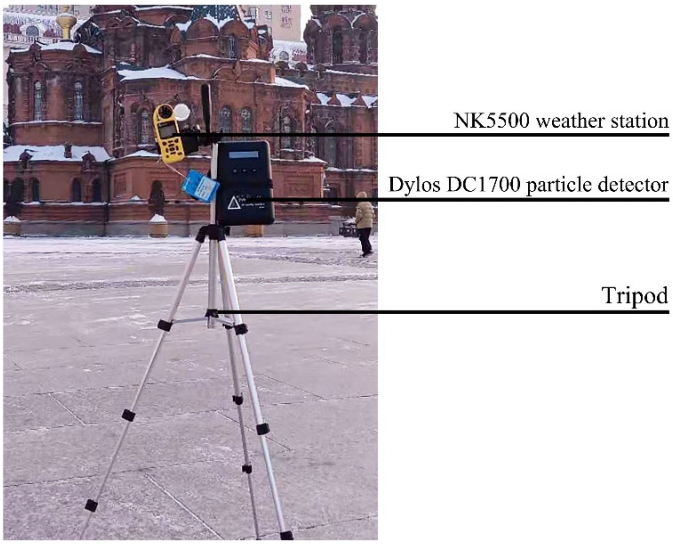

Figure 4. On-site measuring instrument installation.

Table 2. Average weather conditions of each period on the test day.

	8:00–10:00	12:00–15:00	19:00–22:00
	WeaT (°C)/WeaRH (%)/WeaWIN (m/s)/WeaPM$_{2.5}$ (μg/m^3)		
1 December 2020	−12.6/70/2.8/112.7	−9.2/54.2/3/86.8	−12.4/72/2.6/117
2 December 2020	−13.1/69.3/3.3/79	−10.2/61.8/2.9/70.3	−13.1/74/2.5/98.5
9 December 2020	−8.6/53.3/5.2/66	−5.4/60.5/5.3/90	−7.9/67.3/2.6/115
16 December 2020	−20.6/63.7/3/78	−16.4/51.5/3.9/65.5	−17.9/52.8/4.5/77.8
22 December 2020	−7/72.7/6.2/115	−5.7/62.5/4.7/134.3	−13.1/88.5/0.7/147.3
24 December 2020	−17/71.7/2.5/114.7	−14.1/59.5/2.6/139	−19.4/86/0.73/149.3
1 January 2021	−23.2/72/2.4/62.3	−18.6/57.5/2.5/95	−22.8/65/1.6/83.5
4 January 2021	−20.9/65.7/3.3/77.3	−17.4/55.3/2.9/92.3	−20.9/73.3/2.2/65.5
5 January 2021	−21.5/65.7/2.7/56.3	−17.5/55.3/3/77.5	−19.8/64.8/2.4/99
9 January 2021	−22.7/68/2.4/99	−17.9/54/3.2/155.8	−19.8/65.5/2/135.5
11 January 2021	−15.2/67.3/3.7/101.3	−12.1/57/3.4/90.3	−15.1/78.8/1.3/74.5
12 January 2021	−15.7/84.3/1.6/102.3	−8.9/73.8/2.7/96.5	−8.6/92.5/2/91.3
13 January 2021	−17/79.3/4/83.7	−14.3/70.8/3.8/104.5	−16.4/84/1.9/96.8
14 January 2021	−19.3/75.7/2.3/112.3	−16.4/66.3/1.9/99.3	−18.7/77.5/1.4/53.8
20 January 2021	−13.9/83/1.7/68	−4.2/85.5/4.8/99	−6.4/83.3/2.9/68.8
21 January 2021	−15.3/82.7/1.2/107.7	−8.9/56.8/3/198.8	−15.9/71.3/2.4/70
23 January 2021	−16/77/1/142.3	−7.6/52.8/1.3/162.3	−13.8/80.5/1.3/214.5

Table 2. *Cont.*

	8:00–10:00	12:00–15:00	19:00–22:00
	WeaT (°C)/WeaRH (%)/WeaWIN (m/s)/WeaPM$_{2.5}$ (μg/m^3)		
24 January 2021	−11.2/83.3/0.7/263.3	−2.5/56.3/1.3/210.5	−13.8/88.3/1.1/62
8 February 2021	−10.5/78/2.1/118	−3.2/63/2.3/116	−14.5/82/2.4/121
14 February 2021	−8.4/62/3.2/89	−2.3/54/3.6/78	−11.6/76/3.5/95
15 February 2021	−7.6/79/2.8/91	−1.9/69/3.4/88	−10.8/86/2.6/111

2.3. Urban Morphology Parameters Analysis

2.3.1. Urban Morphology Parameters Selection and Computation

Existing studies have shown that air pollution was affected by the traffic conditions, topographic features, economic development, population density, and local weather in the area [25] This study focuses on the impact of urban morphology on PM$_{2.5}$ of the urban canopy. Therefore, transport emissions, building morphology, climate, and local PM$_{2.5}$ concentration are taken as research carriers. The impact of social factors like economic development and population density on PM$_{2.5}$ is controversial, so it is not within this study. Considering the large difference of traffic networks in Harbin and the limited condition of obtaining traffic flow data, the road density was selected as the quantitative index of traffic pollution factor. In addition, due to the special climatic conditions in a cold-climate city, the leaves of most local trees have withered, and the impact on PM$_{2.5}$ is very weak, so it will not be studied here. Finally, referring to previous studies, the selection of urban morphology parameters should meet the following four criteria:

(1) The parameters should significantly affect PM$_{2.5}$ concentration.
(2) The parameters should be easy to extract and calculate.
(3) The parameters affect the design.
(4) Parameter redundancy should be avoided.

Finally, 4 meteorological indicators and 7 urban morphology indicators are selected. Meteorological indicators include hourly wind speed (WeaWIND), hourly humidity (WeaRH), hourly temperature (WeaT), and hourly PM$_{2.5}$ concentration (WeaPM$_{2.5}$) released by the Observatory. Urban morphology indicators include road density (RD), frontal area index (FAI), building volume density (BVD), building density (BD), plot ratio (PR), average building height (AH), and the standard deviation of building height (SDBH). Each urban morphology index is obtained by a geographic information system (GIS), and its research significance and calculation method are listed in Table 3.

Table 3. Selected urban morphology factors.

Urban Morphology Factor	Unit	Equation of Calculation	Theoretical Meaning
RD	%	$\mathrm{RD} = \frac{\sum L_i}{A}$	Traffic pollution intensity
FAI	%	$\mathrm{FAI} = \frac{F}{A}$	The blocking effect of the buildings in the plot on the airflow
BVD	%	$\mathrm{BVD} = \frac{\sum_{i=1}^{n} S_i H_i}{H_{max} A}$	The spatial density of the buildings in the plot
BD	%	$\mathrm{BD} = \frac{\sum_{i=1}^{n} S_i}{A}$	The level of building density in the horizontal direction within the plot
PR	-	$\mathrm{PR} = \frac{\sum_{i=1}^{n} S_i h_i}{A}$	The overall volume and development intensity of the buildings in the plot
AH	m	$\mathrm{AH} = \frac{1}{n} \sum_{i=1}^{n} H_i$	Vertical building development intensity
SDBH	-	$\mathrm{SDBH} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (H - AH)^2}$	The degree of difference and dislocation of the vertical building height within the plot

Note: H_i is the height of each building in the buffer area; H_{max} is the height of the tallest building in the buffer area; h_i is the number of floors of each building in the buffer area; S_i is the bottom area of each building in the buffer area; R is the total floor area of vehicles in the buffer area; F is the sum of the windward area of the building in the direction of the incoming wind (the incoming wind is from the northwest, which is the dominant wind direction of Harbin in winter); L_i is the road length in the buffer area; A is the total area of the buffer area.

2.3.2. Determination of Influence Radius of Urban Morphology Parameters

Diverse urban morphology factors have different degrees of impact on $PM_{2.5}$ concentration under different buffer zones [30]. As shown in Figure 5, in order to obtain the urban morphology factors that can explain the change of $PM_{2.5}$ concentration to the maximum extent, we established buffer zones of different sizes with each measuring point as the center of the circle. According to the existing literature [8,30], we set the radii to 50 m, 100 m, 200 m, 300 m, 400 m, and 500 m respectively.

Figure 5. Changes in architectural spatial morphology under different buffer radii.

The correlation analysis between urban morphology factors with different buffer radii and the $PM_{2.5}$ concentrations of corresponding measurement points was carried out to obtain the buffer radius which can highest interpretation of $PM_{2.5}$ concentration, and the correlation coefficient R^2 and significance Sig. are calculated for comparison. If R^2 were the largest and Sig. (2-tailed) were less than 0.05, then urban morphological parameters under this buffer radius will apply for further analysis.

2.4. Gradient Boosted Regression (GBRT) Trees Model

2.4.1. Model Construction Principle

Gradient boosted regression trees (GBRT) model is derived from the ensemble learning boosting algorithm and have improved on it. Boosting is an integrated method for improving model accuracy. The idea is to combine many "weak learners" into one "strong learner" [31]. It is a numerical optimization technique in which predictors are successively added to the set, and each predictor modifies its predecessor. The gradient descent method is used to minimize the loss expectation function. This sequential process focuses on residuals and continues to iterate until the model meets the observations with minimal residuals. The workflow of the GBRT algorithm is shown in Figure 6.

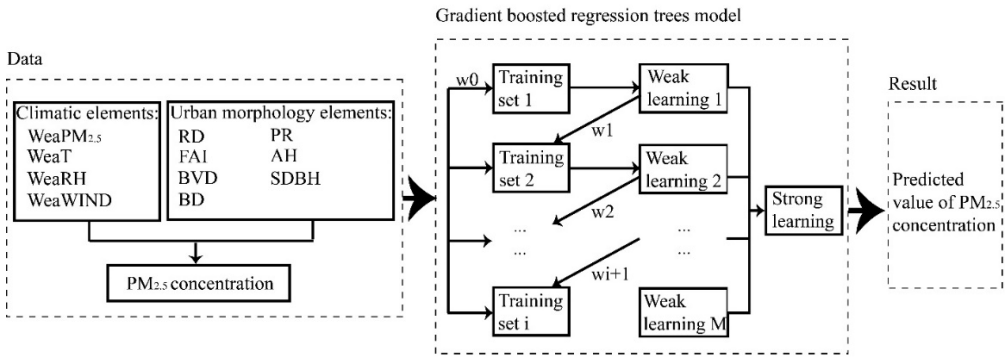

Figure 6. GBRT model workflow.

The main process of GBRT model establishment is as follows:
Let training set sample $T = \{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$,
Determine loss function:

$$L(y, f(x)) = (y - f(x))^2 \tag{1}$$

Step 1. Initialize the first weak learner:

$$f_0(\mathrm{x}) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, c) \tag{2}$$

Step 2. Let the number of iterations $m = 1, 2 \ldots, M$

(a) For $i = 1, 2, \ldots, N$. The negative gradient direction of the loss function was calculated, and the predicted value of the model was obtained, which was used as the prediction residual. The negative gradient of the i-th training data is as follows:

$$r_{mi} = -\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{3}$$

(b) Build a regression tree on the basis of r_{mi}, and obtain the leaf node area R_{mj} of the m-th tree. Predict the leaf node area of the decision tree to obtain an approximate value of the fitting residual.

(c) For $j = 1, 2, \ldots, J$. Linear search is used to obtain the value in the range of leaf nodes. Minimize the loss function. The best residual fitting value of each blade is as follows:

$$\mathrm{c}_m = \underset{\mathrm{c}}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + \mathrm{c}) \tag{4}$$

(d) Update the regression tree:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I\left(x \in R_{mj}\right) \tag{5}$$

Step 3. Get the final model:

$$f(x) = f_M(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I\left(x \in R_{mj}\right) \tag{6}$$

2.4.2. Model Construction and Comparative Validation

Finally, 11 factors including urban morphological variables were selected. $PM_{2.5}$ concentration and climate variables released by the Meteorological Observatory were used as input variables. The $PM_{2.5}$ concentration of each measurement point recorded every hour was used as the output variables of the model. Among them, 70% were divided into training data and 30% test data. Before calculation, grid search (GS) was used to adjust the model parameters, and the Z-score algorithm was used for dimensionless standardization of all data.

In order to verify the prediction accuracy of the GBRT model, decision tree (DT), random forest (RF), and multiple linear regression (MLR) were selected to complete the contrast experiment. Among them, the MLR model is one of the traditional regression methods, while DT and RF models belong to machine learning, which are among currently popular forecasting research methods. The coefficient of determination (R^2), mean square

error (MSE), and mean absolute error (MAE) are selected as the model evaluation indicators. The formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{7}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

Note: y_i is the actual value of $PM_{2.5}$ concentration; $\overline{y}_i$ is the average value of $PM_{2.5}$ concentration; $\hat{y}_i$ is the predicted value of $PM_{2.5}$ concentration; n is the total amount of experimental data.

3. Results and Analysis

3.1. Temporal and Spatial Distribution of $PM_{2.5}$ at Urban Block Scale

We performed 24 h simultaneous monitoring of 25 measurement points for 21 days, collated hourly average data and day–night average data of each measuring point, and then observed the time and spatial distribution characteristics of $PM_{2.5}$ concentration. In this measurement, it was found that there was a big difference in $PM_{2.5}$ concentration among different measurement points, and the maximum difference can reach 69.03μg/m^3. The temporal distribution of $PM_{2.5}$ concentration is shown in Figure 7. As shown in Figure 7, the daily variation of $PM_{2.5}$ concentration presents a bimodal distribution. The $PM_{2.5}$ concentration's first peak appears between 9:00 and 10:00 in the morning. After 10:00, the concentration begins to decline rapidly, and it drops to a minimum between 15:00 and 16:00 when a valley appears. After that, the concentration gradually increased again, reaching a second peak around 21:00–22:00. After 22:00, the concentration decreased slowly, reaching the second valley around 5:00, and then rising again.

In general, there are low wind speeds near the ground and a strong inversion layer at night in winter in Harbin, which is not conducive to the horizontal and vertical dispersion of pollutants. With the increase of surface temperature in the daytime, the inversion layer weakens or disappears, while the effects of near-surface wind and turbulence are strengthened. On the other hand, the temperature in the northern winter decreases significantly at night, the demand for coal increases compared with the daytime, therefore, smoke and dust emissions reach the maximum of the day. After 22:00, human activity gradually ceased and the amount of coal burned at night was greatly reduced. In addition, around 8:00 and 18:00 correspond to the peak commuting period, and the traffic flow during this period increases rapidly and exhaust emissions are the largest. It is worth noting that the peak at night appears after 20:00, 1–2 h behind the off-duty peak, indicating that there is a process of accumulation of pollutants. It can be found that during the day, traffic flow is larger than that at night, the production activities are concentrated, and the emissions are high. However, the change in $PM_{2.5}$ concentration shows that $PM_{2.5}$ concentration at night is higher than that during the day and decreases more slowly. The pollution at night is more serious than that during the day. It shows that the dispersion effect of meteorological factors is more significant than the impact of human activities on the change of $PM_{2.5}$ concentration.

The spatial distribution of $PM_{2.5}$ concentration is shown in Figure 8. As shown in Figure 8, the average $PM_{2.5}$ concentration data of each measurement point during the day (10:00) and at night (22:00) of 21 days were selected. The maximum concentration difference was 62.2 μg/m^3 and 55.5 μg/m^3, respectively.

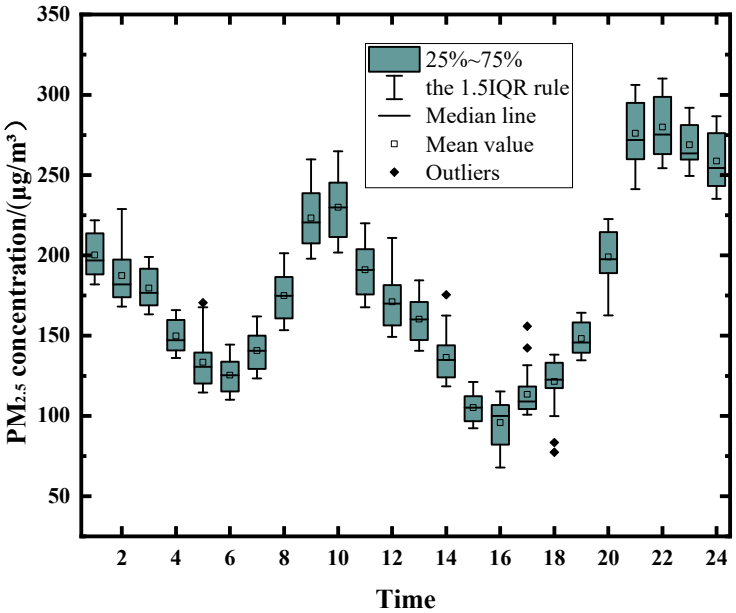

Figure 7. $PM_{2.5}$ concentration's time distribution.

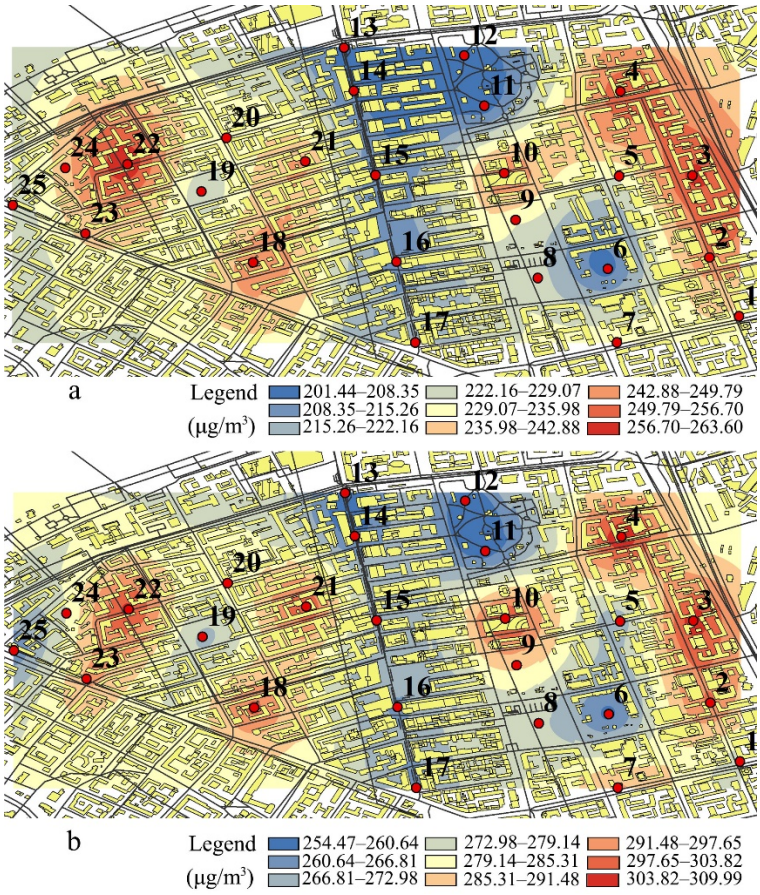

Figure 8. Spatial distribution of $PM_{2.5}$ concentration: (**a**) the day (10:00); (**b**) the night (22:00).

It can be found that the $PM_{2.5}$ concentrations at points 3 and 22 are higher than other measurement points. These points are densely built with high BD, BVD, and PR, which are not conducive to the dispersion of $PM_{2.5}$. The intensity of traffic flow is also relatively large, and the increase in RD greatly increases $PM_{2.5}$ pollution. FAI is higher at point 23 near 22, which is conducive to $PM_{2.5}$ dispersion, so $PM_{2.5}$ concentration at point 23 is lower. SDBH

at point 1 is higher than that at point 7, which is conducive to $PM_{2.5}$ diffusion, so the $PM_{2.5}$ concentration at point 1 is lower. At points 6, 11, and 12, the $PM_{2.5}$ concentrations are lower than other measurement points because these points are in parks or squares, with low BD and RD and far from the road, which is conducive to the spread of $PM_{2.5}$. In addition, contrary to daytime, the value of point 5 at night is lower than point 19, and the values of points 25 and 14 level off. Points 25 and 5 are located on the streets with heavy traffic, and as the traffic flow at night decreases, the $PM_{2.5}$ concentration decreases accordingly.

3.2. Correlation Analysis of $PM_{2.5}$ Concentration and Microclimate

In the study of the time distribution of $PM_{2.5}$ concentration in Section 3.1, we found that $PM_{2.5}$ concentration reached its minimum value at 5:00 (No.1) and 16:00 (No.3) and reached its maximum value at 10:00 (No.2) and 22:00 (No.4). Therefore, we selected the microclimate and $PM_{2.5}$ concentration measured data at these moments to combine with linear regression analysis for correlation research.

As shown in Figure 9a, there is a significant negative correlation between air temperature and $PM_{2.5}$ concentration. With the rise of temperature, $PM_{2.5}$ concentration shows a trend of gradual decline. The measured data show that the variation law of the correlation between $PM_{2.5}$ concentration and temperature is different from that of meteorological temperature (No.1: WeaT = −17.6 °C, R^2 = 0.88; No.2: WeaT = −14.1 °C, R^2 = 0.82; No.3: WeaT = −9.2 °C, R^2 = 0.79; No.4: WeaT = −15.9 °C, R^2 = 0.74), indicating that meteorological temperature has little influence on the correlation between $PM_{2.5}$ concentration and temperature.

As shown in Figure 9b, wind speed has a significant negative correlation with $PM_{2.5}$ concentration. With the increase of wind speed, $PM_{2.5}$ concentration shows a trend of gradual decline. The measured data showed that the correlation between $PM_{2.5}$ concentration and wind speed increased with the increase of meteorological wind speed. Among them, No.3 has the largest meteorological wind speed, and the correlation between $PM_{2.5}$ concentration and wind speed is the strongest (No.1: WeaWIN = 1.4 m/s, R^2 = 0.78; No.2: WeaWIN = 2.8 m/s, R^2 = 0.87; No.3: WeaWIN = 3.5 m/s, R^2 = 0.91; No.4: WeaWIN = 2.3 m/s, R^2 = 0.83), indicating that the higher the meteorological wind speed, the more significant the correlation between $PM_{2.5}$ concentration and wind speed.

As shown in Figure 9c, there is a significant positive correlation between relative humidity and $PM_{2.5}$ concentration. With the rise of relative humidity, $PM_{2.5}$ concentration shows a trend of gradual increase. Among them, No.1 has the largest meteorological relative humidity, and the correlation between $PM_{2.5}$ concentration and relative humidity is the strongest (No.1: WeaRH = 87.6%, R^2 = 0.86; No.2: WeaRH = 70.6%, R^2 = 0.81; No.3: WeaRH = 59.8%, R^2 = 0.76; No.4: WeaRH = 77.6%, R^2 = 0.84), indicating that the higher the meteorological relative humidity, the more significant the correlation between $PM_{2.5}$ concentration and relative humidity.

In summary, urban microclimate has an obvious effect on $PM_{2.5}$ concentration. The increase in temperature and wind speed are conducive to the dispersion of $PM_{2.5}$. Microclimate also has spatial variability and is related to urban morphology factors [32]. Therefore, the influence of urban morphology factors on microclimate should also be considered.

3.3. Model Analysis and Comparison of Validation Results

The urban morphology parameters of each measuring point under the buffer radius of 50 m, 100 m, 200 m, 300 m, 400 m, and 500 m are extracted by GIS and divided into six groups for research. In each group, we successively analyzed the correlation between each urban morphology parameter and its corresponding $PM_{2.5}$ concentration at different times and at different measuring points. Among them, the $PM_{2.5}$ concentration data comes from the 24 h continuous monitoring of each measuring point for 21 days. Finally, the correlation analysis results of different urban morphology parameters and $PM_{2.5}$ concentration in each group are obtained, as shown in Table 4. It can be found that BVD, BD, FAI, and RD reach the maximum when the buffer radius is 300 m. PR and SDBH reach the maximum

when the buffer radius is 200 m. AH reaches the maximum at a buffer radius of 500 m. The urban morphology factors with the highest correlation were selected for further analysis.

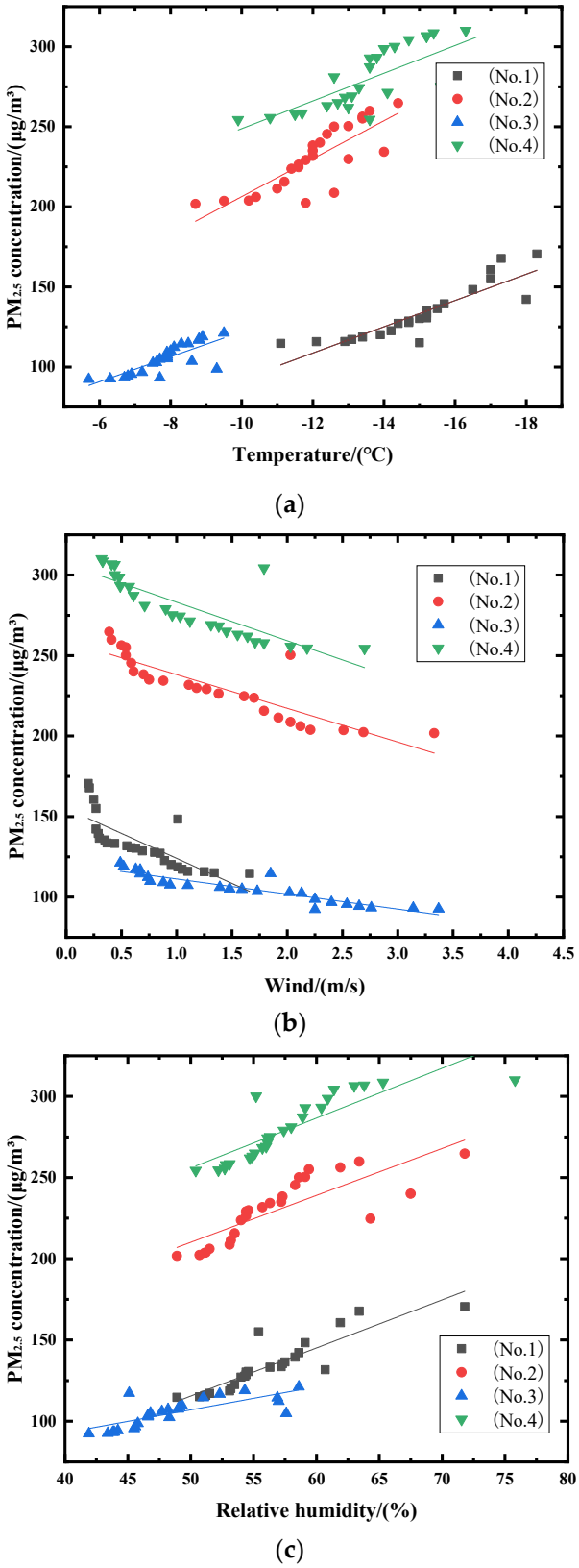

Figure 9. The correlation between $PM_{2.5}$ concentration and microclimate: (**a**) temperature; (**b**) wind; (**c**) relative humidity.

Table 4. Correlation analysis of urban spatial morphology factors and $PM_{2.5}$ concentration under different buffer radii.

Urban Morphology Factor	RD	FAI	BVD	BD	PR	AH	SDBH
No.1: R^2/sig (50 m)	0.696/0.0	0.633/0.5	0.766/0.0	0.580/0.0	0.635/0.0	0.663/0.09	0.731/0.06
No.2: R^2/sig (100 m)	0.754/0.0	0.685/0.0	0.829/0.0	0.628/0.0	0.605/0.0	0.718/0.0	0.792/0.0
No.3: R^2/sig (200 m)	0.792/0.0	0.720/0.01	0.87/0.0	0.660/0.02	0.794/0.0	0.754/0.0	0.890/0.0
No.4: R^2/sig (300 m)	0.895/0.0	0.814/0.0	0.915/0.03	0.846/0.0	0.750/0.0	0.752/0.02	0.840/0.0
No.5: R^2/sig (400 m)	0.625/0.0	0.568/0.0	0.688/0.0	0.521/0.0	0.753/0.0	0.795/0.0	0.656/0.0
No.6: R^2/sig (500 m)	0.533/0.1	0.485/0.0	0.586/0.0	0.444/0.0	0.640/0.0	0.852/0.01	0.560/0.2

As shown in Table 5, the coefficients of determination (R^2), mean square error (MSE), and mean absolute error (MAE) of decision tree (DT), random forest (RF), and multiple linear regression (MLR) models were calculated respectively. All the evaluation indexes of GBRT, DT, and RF models are higher than those of the MLR model, indicating that the machine learning model has a higher explanatory effect on the difference of $PM_{2.5}$ concentration. The reason is that it captures linear and nonlinear relationships between variables. Meanwhile, MAE and MSE of the GBRT model were 1.452 μg/m^3 and 3.246 μg/m^3, respectively, which were 26.3% and 31.5% lower than those of RF and DT models on average.

Table 5. The prediction accuracy of each model on the test set.

	GBRT	MLR	RF	DT
MAE (μg/m^3)	1.452	3.690	1.631	2.308
MSE (μg/m^3)	3.246	8.872	4.285	5.197
R^2	0.978	0.791	0.966	0.894

The comparison between the actual value and predictive value of the GBRT model is shown in Figure 10. From the results of the model, the GBRT model performs well, with an R^2 value of 0.978, indicating that the prediction performance of the GBRT model is stable during the whole research period. To sum up, the GBRT model has the minimum prediction error, the best fitting effect, and high prediction accuracy.

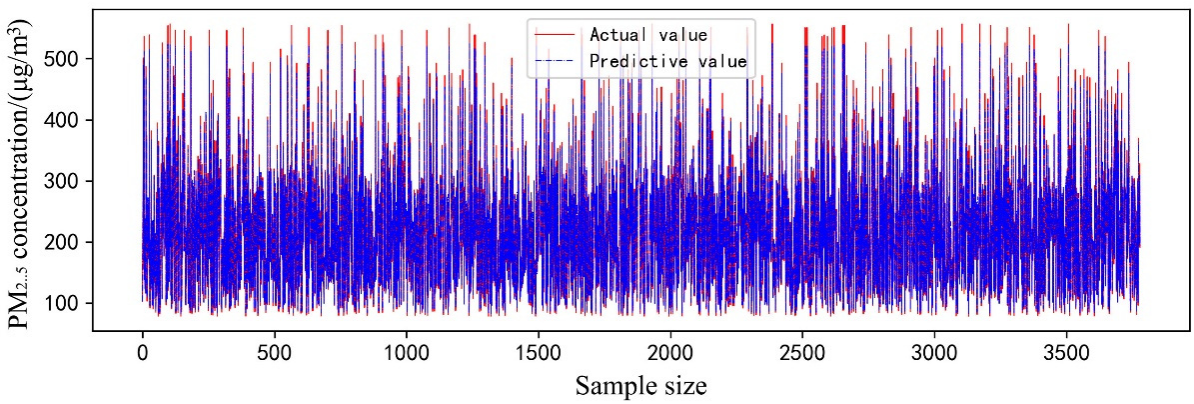

Figure 10. Comparison of the real value and the predicted value of the GBRT model.

3.4. The Influence of Urban Spatial Morphology on $PM_{2.5}$ Distribution

According to the above study, the GBRT model has high accuracy in predicting $PM_{2.5}$ concentration. Therefore, the "Feature_importances" command of the GBRT model is used to further study the influencing factors. The analysis of the contribution degree of each influencing factor is shown in Figure 11.

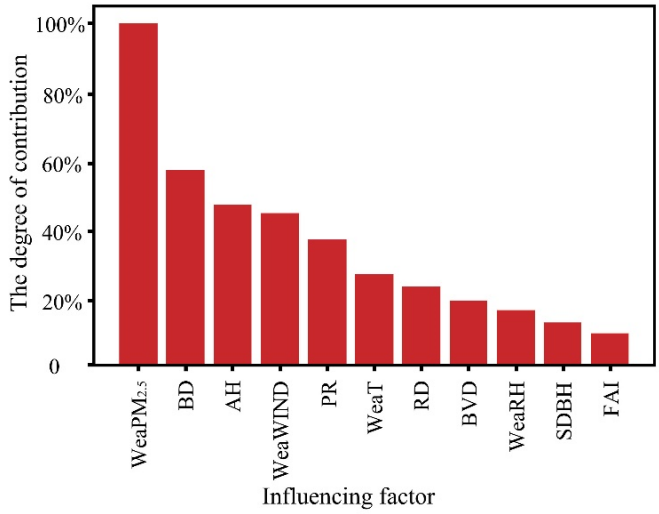

Figure 11. Ranking of factors affecting $PM_{2.5}$ concentration in GBRT model.

It can be found that $WeaPM_{2.5}$ is the most significant factor affecting $PM_{2.5}$ concentration. In previous studies, the air monitoring station far away from the city was often selected to estimate $PM_{2.5}$ concentration [33]. However, it does not apply to the assessment of $PM_{2.5}$ concentration at the block scale. The monitoring site is located in Daoli District, Harbin. Therefore, the data of meteorological stations in this area were selected for research. The influence degree of urban morphology factors on $PM_{2.5}$ in descending order are: BD > AH > PR > RD > BVD > SDBH > FAI.

4. Discussion and Urban Design Recommendations

In recent years, the difference in $PM_{2.5}$ concentration and its relationship with urban morphology factors have attracted much attention. Based on previous studies, this paper considers the influence of potential factors such as microclimate and urban morphology on $PM_{2.5}$ concentration. All of these variables were measured synchronously at high-density measuring points. Compared with previous single studies on buildings, green space, roads, and water bodies at block scale [34,35], this study focuses more on comprehensive consideration of various influencing factors, which is helpful to understand the mode and degree of influence of urban space on $PM_{2.5}$ concentration.

In terms of urban morphology, it is found that building density, average building height, and road density all have an impact on $PM_{2.5}$ concentration which is consistent with previous research but with some differences, for example, Gao Y. [17] proposed that traffic land and $PM_{2.5}$ concentration have a strong correlation, and the correlation is more than that of other urban morphology factors. However, in this study, although road density has a high correlation with $PM_{2.5}$ concentration, it is lower than building density, average building height, and other influencing factors. The main reason for this difference is that cities in different regions have different sources of $PM_{2.5}$ pollution. Pollution in southern cities is still dominated by vehicle emissions even in winter, while in Harbin, a cold-climate city, the heating emissions are even greater in winter. In terms of urban microclimate, temperature and wind speed are strongly negatively correlated with $PM_{2.5}$ concentration, while relative humidity is strongly positively correlated with $PM_{2.5}$ concentration, which is basically consistent with previous studies [4]. In terms of selecting research methods for $PM_{2.5}$ concentration prediction, some scholars have conducted prediction research on air pollution. As shown in Table 6, the following methods are common, and each has its own advantages and disadvantages.

Table 6. The advantages and disadvantages of existing research results.

		Model	Advantage	Disadvantage
Empirical model	Linear regression model	Land use regression (LUR)	Fast calculation speed	Failed to capture the nonlinear relationships
		Multiple linear regression (MLR)	Fast calculation speed	Failed to capture the nonlinear relationships
	Machine learning method	Decision tree (DT)	Capture the nonlinear relationships	Low prediction accuracy
		Random forest (RF)	Capture the nonlinear relationships; Rank the influencing variables based on their importance	-
		Gradient boosted regression trees (GBRT)	Capture the nonlinear relationships; Rank the influencing variables based on their importance	-
		Support vector machine (SVM)	Capture the nonlinear relationships	Cannot rank the influencing variables based on their importance
		Multi-layer perceptron	Capture the nonlinear relationships	Cannot rank the influencing variables based on their importance
		Sequence learning	Capture the nonlinear relationships	Cannot rank the influencing variables based on their importance
Deterministic model	-	Weather research and forecasting (WRF)	Applicable to macroscale	Limited the analysis of air quality at microscales
		Community multiscale air quality (CMAQ)	Applicable to macroscal	Limited the analysis of air quality at microscales

In this study, the GBRT model was used to further analyze the influencing factors of $PM_{2.5}$ concentration in near-surface cities, and design suggestions for promoting urban air pollutant dispersion were put forward as follows:

(1) Horizontal layout of buildings: Building density is the urban morphology factor that has the greatest impact on $PM_{2.5}$ concentration, with an impact degree of 57%; plot ratio and building volume density have an impact degree of 33% and 22% respectively. Therefore, building density parameters should be given priority.
(2) Vertical layout of buildings: the influence degree of average building height and standard deviation of building height is 49% and 12% respectively, so it is necessary to make reasonable restrictions on building height. Attention should also be paid to the diversity of building height.
(3) Existing buildings: it is unrealistic to demolish buildings on a large scale, but the existing urban spatial form can be improved. The impact degree of frontal area index and road density is 11% and 23% respectively. The essence of the impact of road density on $PM_{2.5}$ concentration comes from automobile exhaust emissions. Based on this, removing part of the windward wall and controlling street vehicles is a practical solution.

In conclusion, designers and relevant departments should comprehensively consider the design scheme according to the actual situation. It is worth noting that the actual built environment is very complex, and the generation and dispersion of air pollution is the result of a combination of various factors. This study mainly focuses on the relationship between urban morphological characteristics and $PM_{2.5}$ concentration. However, such as the location distribution of pollution sources, wind direction, turbulence, heat and momentum fluxes, surface temperature, solar radiation for shadowing effects, seasonality and others are also important factors affecting the distribution of $PM_{2.5}$ in cities. Therefore,

pollution sources should be included in future research. At the same time, the accuracy of GBRT model prediction is closely related to sample data. When the training samples can represent the characteristics of the predicted problems, the learning efficiency, and prediction accuracy of the model will be better. On the contrary, the GBRT model will learn a lot of useless experience, which will greatly reduce the prediction rate and affect the prediction accuracy. Therefore, in future research, it is necessary to improve the analysis of sample data to make the prediction research more accurate.

5. Conclusions

In this study, a machine learning method was introduced to establish a prediction model of $PM_{2.5}$ concentration in cold-climate cities. At the same time, this model is used to analyze the influencing factors of $PM_{2.5}$ concentration, providing theoretical reference and technical support for relevant workers in the design and governance of urban blocks. The main conclusions are as follows:

(1) There are significant temporal and spatial differences in $PM_{2.5}$ concentration. The temporal difference indicates that the daily variations in $PM_{2.5}$ concentration are influenced by human activities and meteorological factors. The curves of the average daily variations of $PM_{2.5}$ concentration are similar, with two peaks. The spatial difference indicates that the variation in $PM_{2.5}$ concentration is influenced by urban morphology factors, and $PM_{2.5}$ concentration is different under different urban morphology.
(2) There is a significant linear relationship between microclimate and $PM_{2.5}$ concentration. Wind speed and temperature are negatively correlated with $PM_{2.5}$ concentration, while humidity is positively correlated with $PM_{2.5}$ concentration. However, both microclimate and $PM_{2.5}$ concentrations are affected by urban morphology, indicating that urban morphology, microclimate, and $PM_{2.5}$ concentration interact with each other.
(3) Compared with other models, it is found that the gradient boosted regression trees (GBRT) prediction model has higher prediction accuracy and stability. The GBRT model was used to rank the influencing factors, and it was found that, except for the local $PM_{2.5}$ concentration and climate data released by meteorological stations, urban morphology factors contributed significantly to the change of $PM_{2.5}$ concentration. The highest influence degree is building density and average building height, followed by plot ratio, road density, building volume density, and finally standard deviation of building height and frontal area index.

Author Contributions: P.C. and J.Z. conceived and designed the experiments; C.D. and T.L. performed the experiments; C.D. and P.C. analyzed the data; C.D. and P.C. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Fundamental Research Funds for the Central Universities] grant number [2572021BK02].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

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

Nomenclature

RD	Road density (%)
FAI	Frontal area index (%)
BVD	Building volume density (%)
BD	Building density (%)
PR	Plot ratio
AH	Average building height (m)
SDBH	Standard deviation of building height (m)
T	Measured hourly temperature (°C)
WIND	Measured hourly wind speed (m/s)
RH	Measured hourly humidity (%)
WeaT	Hourly temperature released by the Meteorological Observatory (°C)
WeaWIND	Hourly wind speed released by the Meteorological Observatory (m/s)
WeaRH	Hourly humidity released by the Meteorological Observatory (%)
$WeaPM_{2.5}$	Hourly $PM_{2.5}$ concentration released by the Meteorological Observatory (μg/m^3)

References

1. Ma, K.; Li, C.; Xu, J.; Ren, F.; Xu, X.; Liu, C.; Niu, B.; Li, F. LncRNA Gm16410 regulates $PM_{2.5}$-induced lung Endothelial-Mesenchymal Transition via the TGF-β1/Smad3/p-Smad3 pathway. *Ecotoxicol. Environ. Saf.* **2020**, *205*, 111327. [CrossRef]
2. Zhang, Q.; Shen, Z.; Zhang, T.; Kong, S.; Lei, Y.; Wang, Q.; Tao, J.; Zhang, R.; Wei, P.; Wei, C.; et al. Spatial distribution and sources of winter black carbon and brown carbon in six Chinese megacities. *Sci. Total Environ.* **2021**, *762*, 143075. [CrossRef] [PubMed]
3. Chen, J.; Shan, M.; Xia, J.; Jiang, Y. Effects of space heating on the pollutant emission intensities in "2+26" cities. *Build. Environ.* **2020**, *175*, 106817. [CrossRef]
4. Luo, Y.; Liu, S.; Che, L.; Yu, Y. Analysis of temporal spatial distribution characteristics of $PM_{2.5}$ pollution and the influential meteorological factors using Big Data in Harbin, China. *J. Air Waste Manag. Assoc.* **2021**, *71*, 964–973. [CrossRef] [PubMed]
5. Liu, Z.; Jin, Y.; Jin, H. The Effects of Different Space Forms in Residential Areas on Outdoor Thermal Comfort in Severe Cold Regions of China. *Int. J. Environ. Res. Public Health* **2019**, *16*, 3960. [CrossRef] [PubMed]
6. Fang, G.-C.; Huang, W.-J.; Chen, H.-L.; Chang, M.-C.; Chen, Y.; Huang, C.-Y. Concentrations of particulates and metallic elements in slow wind (average 1.5 m/s) in the winter season. *Environ. Forensics* **2017**, *18*, 188–196. [CrossRef]
7. Yoshie, R.; Jiang, G.; Shirasawa, T.; Chung, J. CFD simulations of gas dispersion around high-rise building in non-isothermal boundary layer. *J. Wind Eng. Ind. Aerodyn.* **2011**, *99*, 279–288. [CrossRef]
8. Zhang, J.; Cui, P.; Song, H. Impact of urban morphology on outdoor air temperature and microclimate optimization strategy base on Pareto optimality in Northeast China. *Build. Environ.* **2020**, *180*, 107035. [CrossRef]
9. Huang, H.; Akutsu, Y.; Arai, M.; Tamura, M. A two-dimensional air quality model in an urban street canyon: Evaluation and sensitivity analysis. *Atmos. Environ.* **2000**, *34*, 689–698. [CrossRef]
10. Longley, I.D.; Gallagher, M.W.; Dorsey, J.R.; Flynn, M.; Allan, J.D.; Alfarra, M.R.; Inglis, D. A case study of aerosol (4.6 nm < Dp < 10 μm) number and mass size distribution measurements in a busy street canyon in Manchester, UK. *Atmos. Environ.* **2003**, *37*, 1563–1571. [CrossRef]
11. Oke, T.R. Street design and urban canopy layer climate. *Energy Build.* **1988**, *11*, 103–113. [CrossRef]
12. Kaplan, H.; Dinar, N. A lagrangian dispersion model for calculating concentration distribution within a built-up domain. *Atmos. Environ.* **1996**, *30*, 4197–4207. [CrossRef]
13. Chan, L.Y.; Kwok, W.S. Vertical dispersion of suspended particulates in urban area of Hong Kong. *Atmos. Environ.* **2000**, *34*, 4403–4412. [CrossRef]
14. Ziomas, I.C.; Melas, D.; Zerefos, C.S.; Bais, A.F.; Paliatsos, A.G. Forecasting peak pollutant levels from meteorological variables. *Atmos. Environ.* **1995**, *29*, 3703–3711. [CrossRef]
15. Kukkonen, J.; Partanen, L.; Karppinen, A.; Ruuskanen, J.; Junninen, H.; Kolehmainen, M.; Niska, H.; Dorling, S.; Chatterton, T.; Foxall, R.; et al. Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmos. Environ.* **2003**, *37*, 4539–4550. [CrossRef]
16. McKendry, I.G. Evaluation of Artificial Neural Networks for Fine Particulate Pollution (PM_{10} and $PM_{2.5}$) Forecasting. *J. Air Waste Manag. Assoc.* **2002**, *52*, 1096–1101. [CrossRef]
17. Gao, Y.; Wang, Z.; Li, C.-Y.; Zheng, T.; Peng, Z.-R. Assessing neighborhood variations in ozone and $PM_{2.5}$ concentrations using decision tree method. *Build. Environ.* **2021**, *188*, 107479. [CrossRef]
18. Lu, W.-Z.; Wang, D. Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. *Sci. Total Environ.* **2008**, *395*, 109–116. [CrossRef]
19. Lu, W.-Z.; Wang, D. Learning machines: Rationale and application in ground-level ozone prediction. *Appl. Soft Comput.* **2014**, *24*, 135–141. [CrossRef]
20. Wang, H.-W.; Li, X.; Wang, D.; Zhao, J.; He, H.-D.; Peng, Z.-R. Regional prediction of ground-level ozone using a hybrid sequence-to-sequence deep learning approach. *J. Clean. Prod.* **2019**, *253*, 119841. [CrossRef]

21. Pach, F.P.; Abonyi, J. Association rule and decision tree based methods for fuzzy rule base generation. *World Acad. Sci. Eng. Technol.* **2006**, *13*, 45–50.
22. Sachdeva, K.; Hanmandlu, M.; Kμmar, A. Real life applications of fuzzy decision tree. *Int. J. Comput. Appl.* **2012**, *42*, 24–28. [CrossRef]
23. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *45*, 1189–1232. Available online: https://www.jstor.org/stable/2699986 (accessed on 14 January 2022). [CrossRef]
24. Shi, K.; Shen, J.; Wang, L.; Ma, M.; Cui, Y. A multiscale analysis of the effect of urban expansion on $PM_{2.5}$ concentrations in China: Evidence from multisource remote sensing and statistical data. *Build. Environ.* **2020**, *174*, 106778. [CrossRef]
25. Liu, C.; Henderson, B.H.; Wang, D.; Yang, X.; Peng, Z.-R. A land use regression application into assessing spatial variation of intra-urban fine particulate matter ($PM_{2.5}$) and nitrogen dioxide (NO_2) concentrations in City of Shanghai, China. *Sci. Total Environ.* **2016**, *565*, 607–615. [CrossRef] [PubMed]
26. Cheng, Y.; Yu, Q.-Q.; Liu, J.-M.; Zhu, S.; Zhang, M.; Zhang, H.; Zheng, B.; He, K.-B. Model vs. observation discrepancy in aerosol characteristics during a half-year long campaign in Northeast China: The role of biomass burning. *Environ. Pollut.* **2021**, *269*, 116167. [CrossRef] [PubMed]
27. de Kok, T.M.; Driece, H.A.; Hogervorst, J.G.; Briedé, J. Toxicological assessment of ambient and traffic-related particulate matter: A review of recent studies. *Mutat. Res. Mutat. Res.* **2006**, *613*, 103–122. [CrossRef]
28. Han, I.; Symanski, E.; Stock, T.H. Feasibility of using low-cost portable particle monitors for measurement of fine and coarse particulate matter in urban ambient air. *J. Air Waste Manag. Assoc.* **2017**, *67*, 330–340. [CrossRef]
29. Liu, C.; Xu, N.; Song, J.; Hu, S. Research on visitors' thermal sensation and space choices in an urban forest park. *Acta Ecol. Sin.* **2017**, *37*, 3561–3569. [CrossRef]
30. Shi, Y.; Xie, X.; Fung, J.C.-H.; Ng, E. Identifying critical building morphological design factors of street-level air pollution dispersion in high-density built environment using mobile monitoring. *Build. Environ.* **2018**, *128*, 248–259. [CrossRef]
31. Franklin, J. The elements of statistical learning: Data mining, inference and prediction. *Math. Intell.* **2005**, *27*, 83–85. [CrossRef]
32. Cao, Q.; Luan, Q.; Liu, Y.; Wang, R. The effects of 2D and 3D building morphology on urban environments: A multi-scale analysis in the Beijing metropolitan region. *Build. Environ.* **2021**, *192*, 107635. [CrossRef]
33. Mo, L.; Yu, X.X.; Zhao, Y.; Sun, F.B.; Mo, N.; Xia, H.L. Correlation analysis between urbanization and particle pollution in Beijing. *Ecol. Environ. Sci.* **2014**, *23*, 806–811.
34. Shi, Y.; Lau, K.K.-L.; Ng, E. Developing Street-Level $PM_{2.5}$ and PM_{10} Land Use Regression Models in High-Density Hong Kong with Urban Morphological Factors. *Environ. Sci. Technol.* **2016**, *50*, 8178–8187. [CrossRef]
35. Wang, Z.; Zhong, S.; He, H.-D.; Peng, Z.-R.; Cai, M. Fine-scale variations in $PM_{2.5}$ and black carbon concentrations and corresponding influential factors at an urban road intersection. *Build. Environ.* **2018**, *141*, 215–225. [CrossRef]