Interval-based Non-dimensionalization Method (IBNM) and Its Application in PM2.5 Grade Prediction

TIANJIAO XU (mc05402@um.edu.mo)
University of Macau https://orcid.org/0000-0003-4268-0422
SHIHONG CHEN
Guangdong University of Foreign Studies
YAN YE
Guangdong Engineering Polytechic
BAIQI LI
Hong Kong Baptist University
HUAPING GUAN
Guangdong University of Foreign Studies

Research Article

Keywords: Interval division, Non-dimensionalization Method, Data processing, PM2.5 grade prediction

Posted Date: March 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1193946/v1

Interval-based Non-dimensionalization Method (IBNM) and Its Application in PM2.5 Grade Prediction

Tianjiao Xu[1†], Shihong Chen[2†], Yan Ye[3], Baiqi Li[4] and Huaping Guan[5*]

[1]Natural Language Processing Portuguese-Chinese Machine Translation Laboratory (NLP2CT), Faculty of Science and Technology, University of Macau, Macau, 999078, China.
[2]Laboratory of Language Engineering and Computing, School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, 510000, China.
[3]College of mechanical and electrical engineering, Guangdong Engineering Polytechnic, Guangzhou, 510000, China.
[4]School of communication, Hong Kong Baptist University, Hong Kong, China.
[5*]School of Economics and Trade, Guangdong University of Foreign Studies, Guangzhou, 510000, China.

*Corresponding author(s). E-mail(s): guan7911@163.com;
Contributing authors: mc05402@um.edu.mo; ibm255@126.com; bfredleaf@163.com; libaiqi@life.hkbu.edu.hk;
†These authors contributed equally to this work.

Abstract

Aiming at the disadvantages of rationality and adaptability of linear dimensionless method, as well as the complexity of constructing polyline and curve dimensionless method, this paper proposes an Interval-based Non-dimensionalization Method (IBNM). IBNM can be assembled in to polyline IBNM or curve IBNM based on the critical points formed by interval partition. Interval division can be classified based on the existing index data, which is scientific, reasonable, simple and practical. Taking example of the prediction of PM2.5 air quality grade

in Guangzhou, this paper constructs four models to predict air quality grade, such as, Support Vector Regression (SVR), Random Forest, Extremely Random Tree, and Gradient Lifting Regression Tree model. Meanwhile, the data is processed with extremum method, polyline IBNM and curve IBNM. The results show that the accuracy of polyline IBNM and curve IBNM is better than extremum method.

Keywords: Interval division; Non-dimensionalization Method; Data processing; PM2.5 grade prediction

1 Introduction

In recent decades, industries in China, India and other developing countries have developed fast; meanwhile, pressures are also brought to environmental governance. Accumulated for years, environmental pollution has become more and more serious, including air pollution. Due to backward conception of environmental protection, governance measures usually lag behind pollution. In recent twenty years, hazy weather in Beijing, Shanghai and Bombay is aggravating obviously, which has affected people's life severely. PM2.5 refers to atmospheric particulate matter with an aerodynamic equivalent diameter of 2.5 microns or less. As the primary pollutant of air pollution, PM2.5 not only could affect the weather, resulting in more foggy days, but also could harm people's health seriously, causing diseases in respiratory system and cardiovascular diseases. As the prime culprit of fog and haze, the concentration of PM2.5 is a significant standard to evaluate air quality; therefore, it is of great significance to study the prediction model of PM2.5 concentration.

At present, the major prediction method of PM2.5 concentration is to analyze historical data and to predict with the method of machine learning. So far, plenty of scholars have done a lot of work to predict PM2.5 concentration and attempted many modeling methods, such as Markov[1], SVM[2], Neural Network[3][4], etc.. In recent years, with the surging of in-depth learning, deep learning models such as CNN[5], RNN[6], DBN[7] have been applied in prediction of PM2.5 concentration. Most studies follow the following 5 steps:

Step 1: Analysis and selection of evaluation factors;

Step 2: Data collection and collation;

Step 3: Data preprocessing;

Step 4: Model specification and training;

Step 5: Model test and application.

Most of the research results focus on Step 4, i.e., trying to improve the accuracy and speed of the algorithm, and neglect the work on Step 3. In order to keep the speed and stability of model training, most methods choose data normalization, and also some will adopt data dimensionality reduction[8]. This paper will focus on step 3 data preprocessing. Some scholars choose the concentration value of PM2.5 as the prediction target when predicting PM2.5,

which could actually cause misunderstanding of the errors. For example, two actual values and prediction values (5,40), (400,600) respectively, which group has the larger error? It is obviously the latter is larger than the former according to the absolute error, but the actual value and the prediction value of the latter are in the same grade according to PM2.5 grade classification in Table 1, but those of the former are of two different grades. Therefore, it is obviously unreasonable to evaluate with absolute error and adopt relative error for the prediction, which is unfavorable to establishing prediction model. However, most the scholars neglect this problem.

Moreover, it is unfavorable for the public to read if the prediction result is concentration value. Nowadays, many countries have definite classification of PM2.5 air quality grades and the public are used to take countermeasures according to the grade prediction results. But if people do not understand the numerical range of specific pollution grades, value prediction would lose its meaning of information guidance. Instead, it is of more guidance significance to predict simply the pollution grades. So, some scholars also take prediction of PM2.5 air quality grades as their targets[9]. However, there are also problems in error judgment of grade prediction, which converts the predicted regression problem into a classification problem, i.e., the continuous value of PM2.5 concentration is changed into scattered grades values, which would discount the value of the error information.

Error problems analyzed above would affect seriously the efficacy of the loss function of the prediction model. To solve this problem, starting from data preprocessing, this essay puts forward Interval-based Non-dimensionalization Method (IBNM) and applied it in prediction of PM2.5 air quality grades. The experiment results show that IBNM could depict effectively the internal rules of the data and help improve the correctness of the prediction.

2 Common non-dimensionalization methods

There are usually different dimensions between various indicators of the sample data in statistical analysis and machine learning, i.e., the measurement units and the order of magnitude differ obviously, so that ideal comparison result between them could not be achieved, which require Non-dimensionalization treatment of these indicator data[10][11]. Non-dimensionalization is the method to transfer the data into the preset scope according to certain rule, for example, [0, 1] and [1, 1], with the aim to eliminate the dimension influence of the primary data. It is also the basic method to treat data in decision making and machine learning[12]. Data characteristics are various, so there are many Non-dimensionalization methods. Dimensions to eliminate data shall follow the principles of being scientific, rational, simple and practical. Common Non-dimensionalization methods include standardized approaches based on the mean and standard deviation, extremum method of converting in proportion according to maximum and minimum values scope, method of specific gravity of the ratio between individual data and the total sum, as well as grade

analysis (AHP) method, and etc. These common methods could be divided, in a word, linear, polyline and curve Non-dimensionalization method[13][14].

2.1 Linear non-dimensionalization method

Linear Non-dimensionalization method is to achieve dimensionless effect through specific linear transformation of the original data of the index[15]. There are various linear transformations methods, such as extremum method, standardization method (Z-score) and method of specific gravity. After the transformation, the data will retain the initial relative order.

2.1.1 Extremum method

Extremum method is mainly to examine the relative location between the indicator data xi and the maximum and minimum values, as shown in the following formula:

$$x_i' = \frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

Formula (1) is just one of the forms of extremum methods the others of which will not be discussed in this essay. Extremum method simply maps index data to a specified range in proportion. After the transformation, the sequence and change rule of the data do not change in essence. However, it would cause obvious interference data transformation if the maximum and minimum data are abnormal.

2.1.2 Standardization method

This method is to standardize the data of the average value and the standard deviation of the initial data. The conversion function is:

$$x_i' = \frac{x_i - \mu}{\sigma} \tag{2}$$

Where, μ is the mean value, and σ is the standard deviation. The mean value after conversion is 0 and the standard deviation is 1, in the other word, to have the data show standard normal distribution. Standardization method is suitable to the condition when the maximum and minimum values of the indicator data are unclear or there is divergence value; but it requires the distribution points of the indicator data conform to or are near normal distribution; otherwise, the conversion effect would be very poor.

2.1.3 Method of specific gravity

Method of specific gravity is the specific gravity of individual data in the overall data sequence, the formula of which is:

$$x_i' = \frac{x_i}{s} \tag{3}$$

Where, the value of s is $\sum x_i$, or $\sqrt{\sum x_i^2}$. Results of such conversion are that the maximum and minimum values of all attributes are different, which could generate different metric scales and is unfavorable for horizontal comparison between multi-attributes[12].

2.2 Polyline non-dimensionalization method

Sometimes, changes of indicator data in different regions have different influences on the evaluation results. For example, increase or decrease of 1000 dollar in one's monthly income, the influences on poor people and rich people are quite different. The same scale difference could not reflect the same evaluation result; similarly, method of specific gravity also has the same problem. In this case, segmental processing method could be adopted, i.e., polyline non-dimensionalization method.

Suppose x_m is the indicator value of the turning point, x'_m is the evaluation of x_m, and then fold line conversion formula is:

$$x'_i = \begin{cases} \dfrac{x_i}{x_m} x'_m & x_i \leq x_m \\ x'_m + \dfrac{x_i - x_m}{max(x)}(1 - x'_m) & x_i > x_m \end{cases} \tag{4}$$

If there are multiple turning points, the conversion function could also be extended to multi-fold function. Obviously, the folding method adopts different conversion coefficients for different ranges of data, which is theoretically more consistent with the change law of data in different ranges[16]. However, reasonable setting of the folds as well as the fold turning points requires deep understanding of law of the data. Literature[14] puts forward a kind of quantitative indicator Non-dimensionalization method aiming at big sample, which in essence is polyline Non-dimensionalization method which determines the folds and turning points with expertise knowledge and experiences.

2.3 Curve Non-dimensionalization method

Sometimes the rule change of index data is progressive, that is, the rule change of data in different intervals is obviously different but the critical point is not obvious. In this case, polyline Non-dimensionalization method could not determine the folds and the turning points, so it is more rational to adopt curve Non-dimensionalization method[17]. Curve Non-dimensionalization method, for example:

$$x'_i = \begin{cases} 0 & 0 \leq x_i \leq x_m \\ 1 - e^{-k(x_i - x_m)^2} & x_i > x_m \end{cases} \tag{5}$$

Formula (5) is semi-normal distribution curve. There are many choices of curves[14][18][19], but no curve has good universality, which has caused difficulties to select curve Non-dimensionalization formula. Also, it is the key point of utilizing reasonably curve Non-dimensionalization method to select

proper curve function under the premise of deep understanding of data properties and laws. Literature [20] put forward a non-linear Non-dimensionalization fuzzy treatment method based on fuzzy theory, which is in essence a curve Non-dimensionalization function.

To sum up, linear Non-dimensionalization method is simple and easy to implement, but it is hard to satisfy indicator data with complex changing rules and quite different intervals. While polyline and curve Non-dimensionalization methods have more complex changing ability, which could adapt to indicator data with more complex changing rule. Nevertheless, how to select conversion function and set appropriate parameters require profound professional knowledge and experience. Therefore, most scholars tend to choose the linear Non-dimensionalization method to process data when they are not confident enough.

3 Interval-based Non-dimensionalization method

Data have different change rule in different intervals, so there shall be different Non-dimensionalization methods. Take daily rainfall as an example (Unit: millimeter): Differences between 5, 35 and 160, 190 are same, but their influences on evaluation differ greatly, for the latter two are in the scope of big storm while the former two differ in small rain or medium rain. However, it is a hard and uncontrollable job to deal with numeric differences between different intervals. Literature [21] put forward Interval-based error evaluation method based on interval division(IEEM) and literature[22] constructed cost-sensitive lost function used for machine learning based on IEEM. Reasonable interval division is helpful for deep understanding and ration judgment of quantitative values.

This essay puts forward a kind of Non-dimensionalization method based on interval division, which construct Non-dimensionalization conversion function taking the advantage of the critical point of interval division, including polyline and curve conversion functions. In terms of complex interval division, existing data grade classification methods are used fully in this essay, such as ladder classification of rainfall grades, disaster grades, risk grades, income grades as well as water and electricity consumption. Data usually contains same change rule in the same grade scope. These grade divisions or summary of people's long term production and life experience, or scientific measurement and verification, all could provide convenient, scientific and reliable basis for interval division.

3.1 Interval division and threshold point

Supposed the Non-dimensionalization data is divided into m intervals. X_i is the value range of the i interval, then,

$$X_i = [\underline{x}_i, \bar{x}_i) = \{x \in R \,|\, \underline{x}_i \leq x \leq \bar{x}_i, i = 1, 2, ..., m\} \tag{6}$$

If the Non-dimensionalization data is continuous, then $\bar{x}_i = \underline{x}_{i+1}$.

Suppose $d_i = f(\bar{x}_i)$, then the point formed by set $(\bar{x}_i, d_i)$ is called the threshold point, as shown in Figure 1. Reflect the indicator data within the same interval to the same range, i.e., $f(X_i) \to [d_{i-1}, d_i)$. Since our object is data Non-dimensionalization, to covert the data to [0, m]. Suppose $d_i = i$, then the data of the i-th grade would be converted to the scope of [i-1, i).

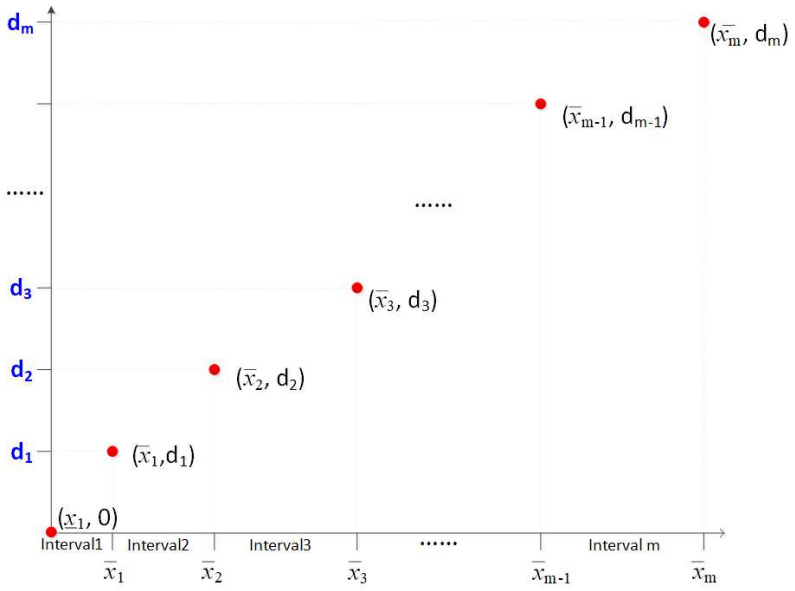

Fig. 1 Schematic plot of threshold points

3.2 Polyline IBNM

Interval-based Non-dimensionalization method is, according to the division rule of grades, to set grade intervals using the threshold point of grade division and adopt different conversion function in different grade intervals. The conversion function is actually a piecewise function:

$$f(x) = \begin{cases} f_1(x),\ x \in X_1 \\ f_2(x),\ x \in X_2 \\ \quad \cdots \\ f_i(x),\ x \in X_i \\ \quad \cdots \\ f_m(x),\ x \in X_m \end{cases} \tag{7}$$

If f in formula 7 adopts piecewise linear function, it is polyline IBNM, denoted by f_p, which in essence is polyline Non-dimensionalization function

with m folds. Suppose, f_p reflects the indicator data value to the scope [0, m), and the maximum difference within the same grade is 1, i.e.,$d_i = i$, then:

$$f_p(x) = \begin{cases} \dfrac{x - \underline{x}_1}{\bar{x_1} - \underline{x}_1}, x \in X_1 \\ \dfrac{x - \underline{x}_2}{\bar{x_2} - \underline{x}_2} + 1, x \in X_2 \\ \cdots \\ \dfrac{x - \underline{x}_i}{\bar{x_i} - \underline{x}_i} + i - 1, x \in X_i \\ \cdots \\ \dfrac{x - \underline{x}_m}{\bar{x_m} - \underline{x}_m} + m - 1, x \in X_m \end{cases} \tag{8}$$

The conversion result in formula 8 also brings convenience to grade judgment, i.e., round up to an integer of the conversion result of $f_p(x)$, and its corresponding grade could be obtained. If $f_p(x) = 3.45$, the corresponding grade of x is 4.

Taking PM2.5 air quality level as an example, the folded IBNM function f_p based on PM2.5 level is constructed. PM2.5 refers to particulate matters in the atmosphere whose diameter is less than or equals 2.5 micron, also known as particulate matter that could enter the lung[23]. Various states have various standards for PM2.5. This essay adopts American PM2.5 daily average concentration to divide the air grades, as shown in Table 1.

Table 1 American air grades corresponding to daily average PM2.5 concentration

Air quality grade	PM2.5 daily average concentration ($\mu g/m^3$)
Good (grade 1)	0-12
Medium (grade 2)	12-35
Unhealthy to sensitive people (grade 3)	35-55
Unhealthy (grade 4)	55-150
Very unhealthy (grade 5)	150-250
Poisonous (grade 6)	$\geq$ 250

Take PM2.5 air quality grades as the y-axis, draw PM2.5 grades and the threshold points of its scope and connect the threshold points with m fold lines, as shown in Figure 2(a).

It could be seen from Figure 2(a) the conversion function is monotonic function in the definition domain, the value range of which after Non-dimensionalization is [0,6). At the tail of the piecewise function, the gradient of the fold lines is relatively lower, which could inhibit the sensitivity of the data change; while it is the reverse at the front.

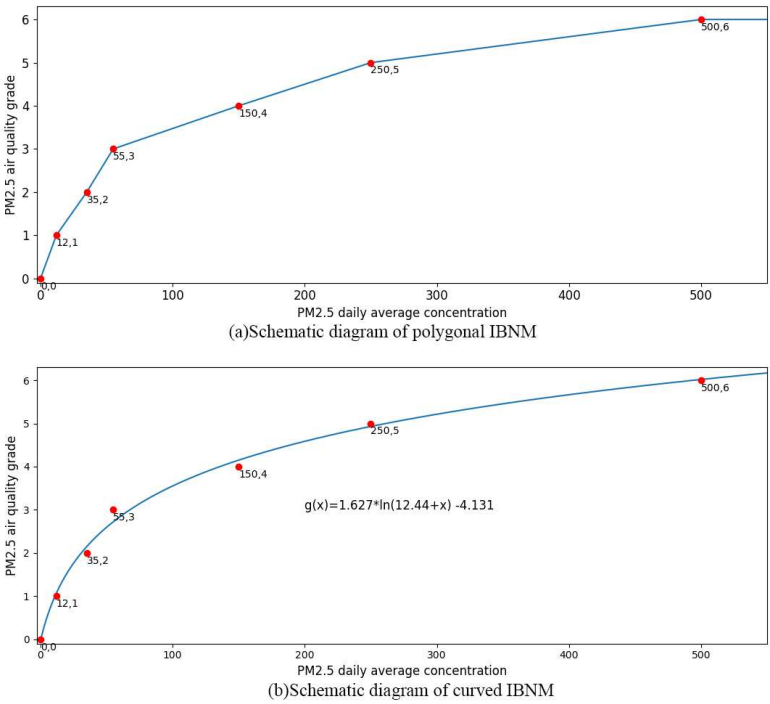

Fig. 2 Schematic diagram of IBNM transformation function

3.3 Curve IBNM

Curve Non-dimensionalization method (curve IBNM): Based on the threshold point of the interval division, a curve function is fitted and constructed, which is known as curve IBNM conversion function, denoted by $f_c(x)$.

Polyline IBNM is piecewise function, which has larger difficulty in treating non-piecewise function. Therefore, it is necessary to put forward a curve IBNM conversion function which is simple in form and could meet the above Non-dimensionalization requirements. Derivable log function and exponential function are all good choices. The method of converting function is also simple, i.e., to fit and construct a curve function based on the threshold points in Figure 2(a). Python, Matlab and other software could provide quick and convenient function fitting tools.

$$f_c(x) = a * ln(b + x) + c \tag{9}$$

Fit function curve_fit in Python Scipy database is adopted. According to the coordinate of 7 threshold points in Figure 2(b), fit formula 9,and obtain a=1.627,b=12.44,c=-4.131. Now the fitted RMSE=0.1827,R2=0.9952,which suggests the fitting effect is very good. The fit effect is as shown in Figure 2(b).

So, the fitted log function is adopted as the curve interval Non-dimensionalization conversion function, i.e.:

$$f_c(x) = 1.627 * ln(12.44 + x) - 4.131,\ x \geq 0 \tag{10}$$

Curve IBNM converts piecewise polyline IBNM into derivable log function. Based on the threshold points of grade division, it is easy to determine the coefficient of this log loss function, which is the loss function obtained by fitting threshold points in intervals. It could be log function, exponential function or monotonous compound function.

It is a fuzzy process from people's sense to the size of values to sensitivity analysis of differences between interval values, so the method of fuzzy mathematics could be used for the analysis. No matter the polyline IBNM or the curve IBNM, all could be considered as fuzzy function. Fuzzification of Non-dimensionalization with this kind of fuzzy function could help rational conversion between interval-based data and eliminate the dimension of the indicator data.

4 IBNM efficiency analysis

It is a fuzzy process from people's sense to the size of values to sensitivity analysis of differences between interval values, so the method of fuzzy mathematics could be used for the analysis. No matter the polyline IBNM or the curve IBNM, all could be considered as fuzzy function. Fuzzification of Non-dimensionalization with this kind of fuzzy function could help rational conversion between interval-based data and eliminate the dimension of the indicator data.

To uncover further the function and efficiency of IBNM data preprocess put forward in this essay, this section will analyze it deeply from two aspects, Non-dimensionalization and error awareness.

4.1 Non-dimensionalization effect

The function of Non-dimensionalization is to eliminate the influences of dimension of the primary data without changing the sequence of the data, that is, nondimensional function is strictly monotonic.

Theorem 1: The transfer function of IBNM is monotonic function in definitional domain.

4.1.1 Polyline IBNM is monotonic function.

Proof: Suppose a, b are two arbitrary numbers in the definition domain, and $a > b$. If a, b are within the same grade, and [min, max) is the range of this level. Then:

$$linIBNM(a) = (a - min)/(max - min) \tag{11}$$

$$linIBNM(b) = (b - min)/(max - min) \tag{12}$$

$$\rightarrow linIBNM(a) > linIBNM(b) \tag{13}$$

Suppose a, b are not in the same grade, according to formula 8, $linIBNM(a) > linIBNM(b)$.

4.1.2 Curve IBNM is monotonic function

Proof: For the curve IBN fitted based on the threshold points, monotonic exponential function, log function or is selected or reasonable monotonous compound function is constructed in the definition domain; so, curve IBNM is also monotonic function in definition domain.

So, polyline IBNM is monotonic function.

4.2 Error evaluation

As is mentioned above, at present many machine learning algorithms do not take error sensitivity into consideration, which differs in different intervals. Another objective of IBNM is to adjust the error to provide rational loss computation method for machine learning algorithm training.

Generally, supervised learning of machine learning training could be regarded to minimize the following target function:

$$\theta^* = argmin(\omega) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{y}_i) + \lambda R(\theta) \tag{14}$$

$\hat{y}_i$ indicates the true value of i-th sample, while y_i indicates the observed value of i-th sample or the output predicted by the learning machine. Taking the common loss function L which is squared loss function as an example:

$$L(y, \hat{y}) = (y - \hat{y})^2 \tag{15}$$

It could be seen that loss function has no limit on maximum error value nor does any treatment on the observed error sensitivity.

IBNM loss function is defined in this study as the loss function of the predicted value y and the true value $\hat{y}$ which have been treated in advance by IBNM Non-dimensionalization.

$$L_{IBNM} = L(IBNM(y), IBNM(\hat{y})) \tag{16}$$

Taking squared loss function as the example, If y and $\hat{y}$ have been treated in advanced by IBNM nodimensionalization, that is:

$$L_{IBNM} = (IBNM(y) - IBNM(\hat{y}))^2 \tag{17}$$

IBNM has two functions on error evaluation and loss computation. One is to limit the maximum error in the definition domain; moreover, polyline IBNM also limits the maximum error in the intervals. The other function is to regulate the sensitivity of the error. As shown in Figure 2(a) and Figure 2(b), in the grade intervals with larger dereferencing scope, the IBNM gradient is

smaller, which could restrict the error sensitivity effectively and be close to people's true sense of error with more convincing computation results.

5 Experimental evaluations

To verify the effect of IBNM, in this study, Guangzhou PM2.5 air quality grade forecast is taken as the example to establish various prediction models, among which there are traditional machine learning algorithm, such as Supporting Vector Machine (SVM), Random Forest, Extremely Random Tree and Gradient Lifting Regression Tree. In the following, brief introduction will be given to these algorithms.

Taking the PM2.5 air quality grade prediction of air quality assessment as an example, the commonly used extreme value method and two kinds of IBNM are used to preprocess the evaluation factor data, and the application is tested in the same evaluation model to test the effect. Most scholars predict a 24-hour mean concentration of PM2.5. However, in view of the vague concept of concentration values among ordinary people, this paper selects the PM2.5 air quality grade that is more instructive to the public as the prediction result. As mentioned earlier, the classification problem actually discretizes the true and predicted values when calculating the error, which inactivates the value of the error. Different from the classification method in [9], this paper still uses the regression algorithm. The result of the model prediction is still the concentration value of PM2.5, or the value of PM2.5 converted by IBNM, and the value of error calculation to the training model is exerted. Finally, the prediction results are converted into PM2.5 pollution levels. If the predicted and actual PM2.5 air quality grades are the same, then the prediction is considered correct, otherwise the prediction is wrong.

5.1 Data preprocessing

The current literature provides long-term forecasts (cycles for the year and month), medium-term forecasts (cycles for the day), and short-term forecasts in hours for the PM2.5 forecast. Among them, the forecasting application in daily is the most widely used, that is, the 24-hour average concentration of PM2.5 is the predicted target. This article takes Guangzhou, the third largest city in China, as the research object. The previous two days of data predicts the PM2.5 grade on the third day.

There are many factors affecting the concentration of PM2.5, but the fluctuations are mainly related to the weather factors. Some scholars believe that other air pollutants, such as sulfur dioxide and nitrogen dioxide, may also have a certain impact on PM2.5, but considering that data acquisition is not easy, these factors are discarded. Therefore, the evaluation factor only takes meteorological factors. The meteorological data includes: average wind speed, maximum wind speed, and wind direction of maximum wind speed, daily precipitation, average temperature, average air pressure, sunshine hours, and average relative humidity. Regarding the prediction model input, due to the

air quality of a certain day, in addition to the PM2.5 status and meteorological factors of the previous day, there is a strong correlation with the weather conditions on the forecast day, and the accuracy of the existing weather forecast is high, which could provide an effective reference for air quality prediction. The model input established in this paper is similar to the literature[24]. Each data is divided into three parts: the first part is the PM2.5 concentration value of one day, the second part is the weather data of the day; the third part is the second weather image of the weather forecast. The output of the model is the PM2.5 predicted value for the next day, which is then converted to the PM2.5 pollution grade. If it is the same as the actual grade, the prediction is considered believable, otherwise the prediction is inaccurate.

The PM2.5 data of Guangzhou City is derived from the monitoring data of the US Embassy from November 2011 to June 2017[1], and its time-sharing monitoring data is aggregated on a daily basis. The meteorological data comes from the National Meteorological Science Data Sharing Platform[2]. It is well known that the concentration of PM2.5 is highly correlated with the season. In order to eliminate the unnecessary seasonal factors, only the data of winter (1, 2, and 12 months) is selected for analysis, and after the records with incomplete data are removed, 482 sets of data are finally obtained. 48 groups were randomly selected as test data, and the rest were used as training data.

In order to verify the effect of interval-based Non-dimensionalization, we use four models, such as SVM and random forest to do three experiments. For the first time, the ordinary extremum method is used to normalize the data. For the second and third times, the folded IBNM method (Equation 8) and curvilinear IBNM method (Equation 10) were used to process PM2.5 data respectively. The data of other factors were still processed by extremum method.

5.2 Experimental implementation

Four prediction models are constructed, which are based on SVR, random forest, extreme random tree, gradient boost regression tree. The specific construction method and parameter settings are as follows.

The radial basis function (RBF) is chosen as the kernel function of the SVR. The two important parameters of the SVR, gamma and cost, need to be preset. The grid search cross-validation method is used to traverse all possible combinations of gamma and cost parameters in an exhaustive way to obtain optimal results. The search range of Cost is set to 6 numbers: (1, 5, 8, 10, 12, 20), and the search range of gamma is 10^i,i=-2,-1,0,1,2. The resulting optimal parameter results are: gamma=0.1, cost=5.

Random forests, extreme random trees, and gradient-enhanced regression trees are simple and easy to utilize. There are not many parameters to be adjusted, and the meaning of the parameters is intuitive and easy to understand. These parameters are either used to enhance the predictive power of

[1] http://www.stateair.net/web/post/1/3.html
[2] http://www.stateair.net/web/post/1/3.html

the model or to make the model faster. The libraries of the three methods provided by the Sklearn library, using the default hyperparameters of the library functions, usually produce a good prediction. Since the pivot point of the experiment is to verify the effect of data preprocessing, the parameter settings take the default values.

5.3 Analysis of experimental results

Figure 3, 4, 5 show the predicted results after processing PM2.5 data using the polar implant method, the polyline IBNM and the curve IBNM. Each sub graph in the figure has a section divided by a horizontal dotted line. If both the real value and the corresponding predicted value are in the same range, it means that the prediction grade is the same, and the prediction is regarded as correct. Table 2 shows the prediction accuracy of the four models in the case of three kinds of data prediction processing. It can be seen that the prediction results of various methods are acceptable. From the perspective of model methods, random forests and extremely random trees perform better than SVR and Gradient regression.

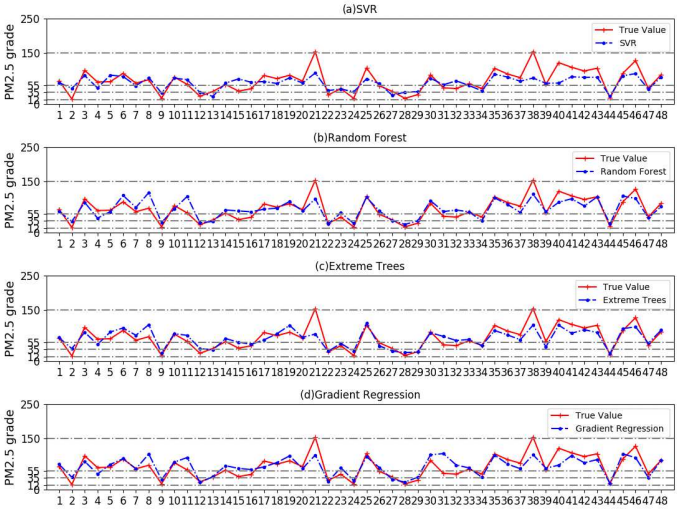

Fig. 3 Prediction results of four models after the data processing by the extremum non-dimensionalization method

From the perspective of dealing with the Non-dimensionalization method, the accuracy of the polyline IBNM and the curve IBNM increased by an average of 4.7% and 7.33%. The SVR model is the most obvious, which has an accuracy of 8.3% and 12.5%, respectively. This result shows that IBNM,

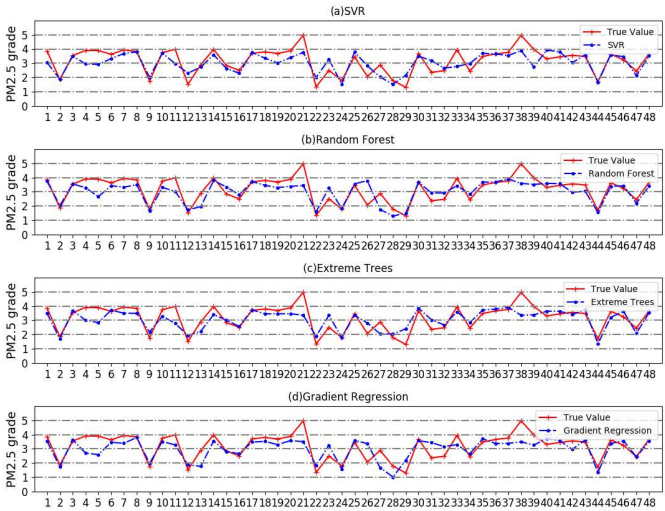

Fig. 4 Prediction results of four models after the data processing by the extremum non-dimensionalization method

Table 2 Accuracy of four models after processing data with three non-dimensionalization methods

Methods	SVR	Random forest	Extreme trees	Gradient regression
Extremum method	64.6%	77.1%	75.0%	70.8%
Polygonal IBNM	72.9%	79.2%	79.2%	75.0%
Curved IBNM	77.1%	79.2%	81.3%	79.2%

Table 3 Pearson correlation coefficient of four models after processing data by three non-dimensionization methods

Methods	SVR	Random forest	Extreme trees	Gradient regression
Extremum method	0.799	0.842	0.792	0.740
Polygonal IBNM	0.781	0.780	0.775	0.739
Curved IBNM	0.772	0.779	0.783	0.775

whether it is a polyline or a curve type, is significantly better than the extreme value method.

As can be seen in Table 3, the Pearson correlation coefficient between the predicted result and the true value, differences between the prediction performance of each model on the three Non-dimensionalization methods are trivial, indicating the almost the same prediction performance of the model. However, IBNM considers the division of the grade interval in the step of data

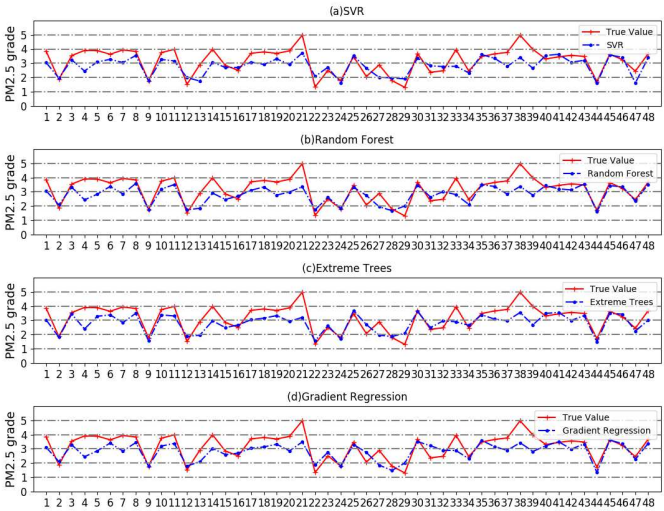

Fig. 5 Prediction results of four models after the data processing by the extremum non-dimensionalization method

preprocessing. Therefore, it is helpful for the model loss function to adjust, so that the predicted value is closer to the real value grade interval. Ultimately, IBNM is more accurate at the PM2.5 grade.

6 Summary and outlook

6.1 Summary

In statistical analysis, machine learning, etc., in order to truly reflect the substantive meaning of the indicator data, or accelerate the computer speed of the model and other goals, it is often necessary to process the data to be dimensionless. The commonly used linear dimensionless method is difficult to reflect the numerical difference of the index data in different intervals, and the polyline and curve methods often make it difficult to select the conversion function because it is difficult to grasp the variation law of the data. This paper proposes an Interval-based Non-dimensionalization Method (IBNM). The IBNM method can be combined into a polyline IBNM according to the critical point formed by the interval division, or can be fitted to generate a curve IBNM. Interval division can be based on the existing index data classification method, which is scientific, reasonable, simple and practical. By constructing the SVR, random forest, extreme trees and gradient regression to predict the PM2.5 air quality grade in Guangzhou City, verify that the results of IBNM are better than the commonly used extremum method.

Because IBNM relies on the scientific and rational division of data intervals, there are certain limitations in its use. The hierarchical method of data can provide a convenient, scientific and reliable basis for interval division. However, when there is no unwarranted grade division of data, scientific and reasonable division can be made by means of expert knowledge.

6.2 Outlook

The main goal of this paper is to verify the effectiveness of IBNM, so the traditional four machine learning methods are selected on the predictive model. In the future, maybe more models could be applied, such as deep learning algorithms. In terms of the evaluation factor, the easily accessible meteorological data was selected as PM2.5 data, and other air pollutants such as sulfur dioxide were not considered, and they could also be studied in future research. In order to eliminate the interference of seasonal factors, we filter the data of other seasons but winter. How to eliminate the interference of seasonal factors, is our further work objective as well. The status of PM2.5 varies widely, and it may be ineffective to change a model algorithm that is accurate in one place to another. Therefore, constructing a universal model algorithm is also a direction for efforts in the future.

In addition, in the prediction of PM2.5 air pollution grade, this paper does not take into account the imbalance of losses, that is, when severe pollution is incorrectly predicted as minor contamination, and minor pollution is incorrectly predicted as severe contamination, the former will bring greater losses. for the reason that unbalanced losses should be considered in the error assessment. Therefore, unbalanced loss should be considered in error assessment. How to adjust this imbalance loss is also an important content of our next work.

Declarations

Ethical approval This study does not contain any studies with human participants or animals performed by any of the authors.

Funding details This work was supported by National Natural Science Foundation of China (grant No. 61772146), the Colleges Innovation Project of Guangdong (grant No. 2016KTSCX036), Guangzhou program of Philosophy and Science Development for 13rd 5-Year Planning (grant No. 2018GZGJ40), Guangdong Key Lab of Ocean Remote Sensing (LORS). (2017B030301005, GDJ20154400004), and GuangDong University of Foreign Studies (17ss13).

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed Consent The authors declare no conflict of interest. The funders

had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Data transparency The datasets generated during and analysed during the current study are available through
http://www.stateair.net/web/post/1/3.html,
http://www.stateair.net/web/post/1/3.html.

Authorship contributions Software, investigation, formal analysis, writing—review and editing, T.X.; conceptualization, methodology, resources, writing—original draft preparation, S.C.; data curation, visualization, B.L.; project administration, Y.Y.; validation, supervision, H.G.; funding acquisition, S.C. and H.G. All authors have read and agreed to the published version of the manuscript.

References

[1] Dong, M., Yang, D., Kuang, Y., He, D., Erdal, S., Kenski, D.: Pm2. 5 concentration prediction using hidden semi-markov model-based times series data mining. Expert Systems with Applications **36**(5), 9046–9055 (2009)

[2] Sun, W., Sun, J.: Daily pm2. 5 concentration prediction based on principal component analysis and lssvm optimized by cuckoo search algorithm. Journal of environmental management **188**, 144–152 (2017)

[3] Chen, Y.: Prediction algorithm of pm2. 5 mass concentration based on adaptive bp neural network. Computing **100**(8), 825–838 (2018)

[4] Zhou, S., Li, W., Qiao, J.: Prediction of pm2. 5 concentration based on recurrent fuzzy neural network. In: 2017 36th Chinese Control Conference (CCC), pp. 3920–3924 (2017). IEEE

[5] Huang, C.-J., Kuo, P.-H.: A deep cnn-lstm model for particulate matter (pm2. 5) forecasting in smart cities. Sensors **18**(7), 2220 (2018)

[6] Ong, B.T., Sugiura, K., Zettsu, K.: Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting pm 2.5. Neural Computing and Applications **27**(6), 1553–1566 (2016)

[7] Cui, X., Xie, J., Zhang, F., Ding, L., Zengshun, L.I., Hao, Z., Liu, Y., Zhao, Q.: Establishment of pm2.5 forecasting model based on deep learning. Beijing Surveying and Mapping (2017)

[8] Gao, Y., Su, C., Li, H.-G.: A kind of deep belief networks based on nonlinear features extraction with application to pm2. 5 concentration

prediction and diagnosis. Acta Autom. Sin **44**, 318–329 (2018)

[9] Xia: Prediction and space distribution analysis of pm2.5 daily average pollution levels based on esn mode. China University of GeosciencesBeijing **02** (2015)

[10] Yajun, G., Pingtao, Y., *et al.*: Character analysis of linear dimensionless methods. Statistical Research **25**(2), 93–100 (2008)

[11] Gregory, A., Jackson, M.: Evaluation methodologies: a system for use. Journal of the Operational Research Society **43**(1), 19–28 (1992)

[12] Xiaoming, Z.: Comparative analysis of data nondimensionalization in decision analysis [j]. Journal of minjiang university **33**(5), 21–25 (2012)

[13] Chen, S.: Risk rating statistical methodology research. Statistics and Decision (4), 8–10 (2003)

[14] W. Jiang, D.L. H. Zhao, Wang, L.: Dimensionless method of quantitative index for large sample evaluation. Statistics and Decision (2012-17), 4–9 (2021)

[15] LiaoWenhe, C.Z.: An approach to evaluate the performance level of the information system applications in the manufacturing industry [j]. Journal of Nanjing University of Science and Technology **1** (2003)

[16] Qiu. Cai, G.L.: Pft-based comprehensive effectiveness evaluation for the space early warning system. Graduate School of National University of Defense Technology (5), 3 (2013)

[17] Mei, Z.: Research on satisfactory evaluation method and its application. PhD thesis, Southwest jiaotong University (2004)

[18] Xu, J., Louge, M.Y.: Statistical mechanics of unsaturated porous media. Physical Review E **92**(6), 062405 (2015)

[19] Ma, L.: Standardization of statistical data—dimensionless method. Beijing Statistics **34**, 35 (2003)

[20] Zhu, K.: Nonlinear dimensionless fuzzy handling of evaluation indexes. Systems Engineering **14**(6), 58–62 (1996)

[21] X. Liu, D.T.Y.Y. S. Chen: Interval error evaluation method (ieem) and it's application (2016-1), 84–86 (2021)

[22] Chen, S., Liu, X., Li, B.: A cost-sensitive loss function for machine learning. In: International Conference on Database Systems for Advanced Applications, pp. 255–268 (2018). Springer

[23] He, K., Yang, F., Ma, Y., Zhang, Q., Yao, X., Chan, C.K., Cadle, S., Chan, T., Mulawa, P.: The characteristics of pm2. 5 in beijing, china. Atmospheric Environment **35**(29), 4959–4970 (2001)

[24] Ting, D., Jianhui, Z., Yong, H.: Aqi levels prediction based on deep neural network with spatial and temporal optimizations. Comput. Eng. Appl **53**, 17–23 (2017)