Front. Earth Sci. 2022, 16(3): 798−808
https://doi.org/10.1007/s11707-021-0948-z

RESEARCH ARTICLE

Comparison and correction of IDW based wind speed interpolation methods in urbanized Shenzhen

Wei ZHAO[1], Yuping ZHONG[1], Qinglan LI (✉)[1], Minghua LI[2], Jia LIU[2], Li TANG[2]

1 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
2 Shenzhen Meteorological Bureau, Shenzhen 518040, China

Abstract Based on the 2-min average wind speed observations at 100 automatic weather stations in Shenzhen from January 2008 to December 2018, this study tries to explore the ways to improve wind interpolation quality over the Shenzhen region. Three IDW based methods, i.e., traditional inverse distance weight (IDW), modified inverse distance weight (MIDW), and gradient inverse distance weight (GIDW) are used to interpolate the near surface wind field in Shenzhen. In addition, the gradient boosted regression trees (GBRT) model is used to correct the wind interpolation results based on the three IDW based methods. The results show that among the three methods, GIDW has better interpolation effects than the other two in the case of stratified sampling. The MSE and R^2 for the GIDW's in different months are in the range of 1.096–1.605 m/s and 0.340–0.419, respectively. However, in the case of leave-one-group-out cross-validation, GIDW has no advantage over the other two methods. For the stratified sampling, GBRT effectively corrects the interpolated results by the three IDW based methods. MSE decreases to the range of 0.778–0.923 m/s, and R^2 increases to the range of 0.530–0.671. In the non-station area, the correction effect of GBRT is still robust, even though the elevation frequency distribution of the non-station area is different from that of the stations' area. The correction performance of GBRT mainly comes from its consideration of the nonlinear relationship between wind speed and the elevation, and the combination of historical and current observation data.

Keywords wind interpolation, Shenzhen, inverse distance weight, gradient boosted regression trees

1 Introduction

Wind is an essential meteorological element that has a major impact on the industry and human life. It is one of the most important renewable energy resources for industry (Palutikof et al., 1987; Masseran et al., 2012; Wang et al., 2016a), and it plays a vital role in exchanging ocean-air material, regulating the regional air pollution, and influencing outdoor human comfort (Arain et al., 2007; Kleerekoper et al., 2012; Li et al., 2014; Wanninkhof, 2014; Coccolo et al., 2017; Yang et al., 2020). Therefore, a comprehensive understanding of the regional wind condition is indispensable for the region's long-term sustainable development.

To have a panorama of the wind condition in a region, a suitable interpolation is needed for estimating the wind condition at the points where no wind observation stations are available, or missing values occur at some points. The inverse distance weight (IDW) method is one of the most commonly used spatial interpolation methods (Mitas and Mitasova, 1999; Li et al., 2014). The advantage of IDW is that it is easy to understand and implement, but it has low accuracy when dealing with the wind field which is sensitive to elevation. To improve accuracy, Ozelkan et al. (2016) applied the wind profile model to convert the wind speed at different elevations to the same horizontal plane, and the Hellman coefficient of the wind profile is determined by the nature of the land cover. This method is called the modified inverse distance weight (MIDW) method. Nalder and Wein (1998) first proposed the gradient-plus-inverse distance squared (GIDS) method to interpolate temperature and precipitation in western Canada. This method showed better performance in interpolating monthly mean temperature and rainfall than the traditional IDW method. Peng (2017) proved that GIDS was also suitable for wind speed interpolation. Considering its similarity to IDW, Peng (2017) called it the gradient inverse distance weight (GIDW) method.

MIDW and GIDW have improved the interpolation accuracy by considering the influence of location altitude

Received April 6, 2021; accepted September 9, 2021

E-mail: ql.li@siat.ac.cn

based on the IDW method. However, the improvement is not significant. Previous studies have shown that, besides the elevation, the slope and orientation of the terrain are also important factors affecting the wind speed near the surface (McCutchan and Fox, 1986; Johnson, 1999; Curry et al., 2012; Yuan et al., 2015; Winstral et al., 2017; Li, 2019). Daly et al. (2008) have proved that considering the influence of terrain features may further improve the interpolation accuracy when facing a complex terrain environment.

Considering the nonlinear influence of terrain on the wind field, we choose the gradient boosted regression trees (GBRT) model to analyze the relationship between topography and wind speed. GBRT incorporates important advantages of tree-based methods, handling different types of predictor variables and accommodating missing data. GBRT does not need prior data transformation or elimination of outliers, which can fit complex nonlinear relationships and automatically handle interaction effects between predictors (Elith et al., 2008). GBRT can produce competitive, highly robust, interpretable procedures for both regression and classification. It has been proved to be an effective tool that handles imperfect data, such as existing effects of wide tails and outliers in the distribution of the input variables (Friedman, 2001).

Shenzhen is one of the most prosperous and densely populated areas globally, located in the south China coast area. Meanwhile, Shenzhen is ranked as the 9th most competitive financial center in the world (between San Francisco and Zürich) and 6th most competitive in Asia (after Shanghai, Tokyo, Hong Kong, Singapore, and Beijing) in the 2020 Global Financial Centres Index (Morris et al., 2020). For such an important megacity, exploring the wind condition in the city is vital for regional sustainable development. Shenzhen has a variety of topography, most of which are hilly areas with plains around them (Chang et al., 2013). Its complex terrain makes wind speed interpolation difficult. Unfortunately, tropical cyclones, strong convection, and cold waves occur frequently in Shenzhen, bringing strong winds to the area and seriously threatening people's lives and property (Liu et al., 2020). Therefore, a thorough understanding of the city's wind condition is crucial, and accurate wind interpolation in Shenzhen is essential.

Based on the IDW based methods, this study tries to revise the interpolation error using the GBRT model. The terrain features, such as elevation, slope, and orientation, are imported into GBRT to improve the accuracy of the near surface wind speed interpolation in Shenzhen. As Shenzhen is a coastal city, its wind variation is affected by the sea-land breeze circulation and monsoon. Therefore, the day in a year and the hour in a day are also considered in the GBRT model. The importance level of the variables from the terrain-related features and time-related features in the GBRT model is analyzed to measure the impact weight of different factors on the wind speed field in the region. The remainder of this paper is organized as follows. The data and methodologies are described in Section 2. The results and discussion are shown in Section 3, and the conclusions are summarized in Section 4.

2 Data and methods

2.1 Data

This study focuses on Shenzhen, China (Fig. 1) to explore a novel way to interpolate the wind speed in a region. The data used in this study are the hourly wind speed data of 175 automatic weather stations in Shenzhen from January 2008 to December 2018, obtained from the Shenzhen Meteorological Bureau. The hourly wind speed data are the 2-min average wind speed calculated right on the hour. For example, the 2-min average wind speed at 12 o'clock is calculated by the average wind speed over the period from 11:58 to 12:00.

Before processing the above-mentioned observation data, quality control is necessary. Specific steps are as follows: 1) set the data with wind speed greater than 50 or less than 0 m/s to a missing value; 2) exclude the stations with a proportion of effective data less than 85%. Finally, the qualified observation data of 100 stations are kept for further research. The locations of these 100 meteorological observation stations are shown in Fig. 1. The average resolution of these stations is 4.44 km.

The topography data, including the slope, aspect, and height over the Shenzhen region, are derived from the global digital elevation model (DEM), produced by the Consultative Group for International Agriculture Research Consortium for Spatial Information Shuttle Radar Topographic Mission version 4.1 (CGIAR-CSI-SRTM v4.1) (Jarvis et al., 2008), based on the calculation of ArcGIS with a resolution of 90 m.

2.2 IDW based interpolation methods

2.2.1 IDW method

Inverse distance weight interpolation (IDW) is a traditional interpolation method based on the first law of geography which means that the closer the distance of two points, the higher the similarity of their attribute values (Tobler, 1970). The formula for IDW is

$$z(x_e) = \frac{\sum_{i=1}^{m} \frac{z(x_i)}{d_i^n}}{\sum_{i=1}^{m} \frac{1}{d_i^n}}, \quad (1)$$

where $z(x_e)$ is the interpolation result of point x_e, and $z(x_i)$

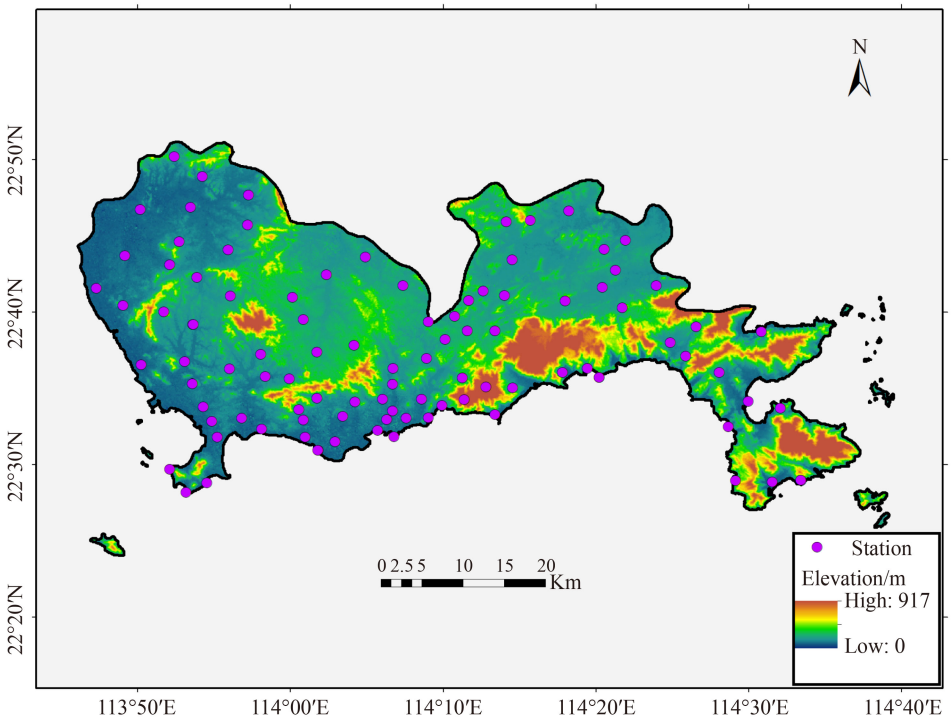

Fig. 1 Location of the 100 automatic weather stations (AWSs) and the geographical features of the study area.

is the observation value of point x_i, and d_i^n is the nth power of the distance from the point x_i to point x_e.

2.2.2 MIDW method

Observations of wind speed at weather stations are affected by varying characteristics of local topography, surface roughness, and land cover. To estimate an unknown wind speed of a specific point, all the observation data should be normalized by considering the elevation's effects before the interpolation process can be accurately conducted in spatial interpolation (Ozelkan et al., 2016). The first step of MIDW is transforming the known wind speeds from different heights of the points to a unified reference height by the wind profile model. The power law, described by Eq. (2), is one of the most common methods in wind engineering for expressing the relationship between wind speed and height above ground (Bañuelos-Ruedas et al., 2010; Kikumoto et al., 2017):

$$z' = z\left(\frac{h}{h_0}\right)^{\alpha}, \tag{2}$$

where z', z are the wind speeds at the heights of h and h_0, respectively. h_0 is the elevation of the reference plane given manually, which is set to 30 m in this study, close to the mean elevation of the 100 meteorological observation stations in Shenzhen. α is the Hellman coefficient. It signifies the surface roughness and land-cover types and can be determined by the Normalized Difference Vegetation Index (NDVI) as in Ozelkan's work (Ozelkan et al., 2016). Since Ozelkan worked in a more heterogeneous area, different α values were used for each station (Ozelkan et al., 2016). On the other hand, a single α value was preferred in this study because the study area is an urbanized city. Shenzhen is a highly urbanized city located along the Guangdong coast. A previous study found that the mean Hellmann coefficients in several Guangdong coastal cities varied from 0.131 to 0.208 (Zhi et al., 2001). In addition, a study about the wind speed interpolation under complex terrain conditions showed that the appropriate mean Hellmann coefficient for the built-up area was 0.16 (Xu et al., 2012). Therefore, we use 0.16 as the approximate Hellmann coefficient in Shenzhen in this study for simplicity.

The second step is estimating the wind speed of the interpolated points at the reference height by the IDW method with the transformed wind speed of known points. The detailed computing process is described by Eq. (3):

$$z'(x_e) = \frac{\sum_{i=1}^{m} \frac{z'(x_i)}{d_i^n}}{\sum_{i=1}^{m} \frac{1}{d_i^n}}, \tag{3}$$

where the winds speed $z'(x_i)$ of point x_i is calculated by Eq. (2). The form of Eq. (3) is the same as that of IDW, but use $z'(x_e)$ and $z'(x_i)$ instead of $z(x_e)$ and $z(x_i)$ in the IDW method.

The third step is transforming the values of the interpolated points from the reference height back to their real heights by the following formula:

$$z(x_e) = z'(x_e) / \left(\frac{h}{h_0}\right)^{\alpha}. \quad (4)$$

2.2.3 GIDW method

Gradient inverse distance weight interpolation (GIDW) is an interpolation method that considers the anisotropy of meteorological elements in different directions. Its weights are defined by the gradient in different directions instead of the traditional inverse distance in the IDW method, and the calculation formula is as following (Nalder and Wein, 1998):

$$z(x_e) = \frac{\sum_{i=1}^{m} \frac{z(x_i) + (X_e - X_i) \times C_x + (Y_e - Y_i) \times C_y + (H_e - H_i) \times C_h}{d_i^n}}{\sum_{i=1}^{m} \frac{1}{d_i^n}}, \quad (5)$$

where X_e and X_i are the X coordinates of the interpolation point and observation station 'i', respectively; Y_e and Y_i are the Y coordinates of the interpolation point and observation station 'i', respectively. H_e and H_i are the elevations of the interpolation point and observation station 'i', and C_x, C_y, and C_h are the regression coefficients for X, Y, and elevation, respectively.

2.3 Interpolation correction by GBRT

Gradient boosted regression trees model (GBRT) is an ensemble method realized by concatenating a series of decision trees. Its main idea is to build a new regression tree in the descending direction of the loss function gradient based on the results of the last regression tree (Wang et al., 2016b). The GBRT model has the advantage of high robustness of ensemble learning for model training (Zhang et al., 2020), so it is very suitable to deal with the problem of interpolation, which has a large number of non-observed data (Friedman, 2002).

In this study, GBRT is used to deal with the stations' wind speed regression. Here, m denotes the number of total samples; T is the maximum iterations and L is the loss function. The core steps of the GBRT algorithm are as follows (Wang and Tang, 2019).

A) Initialize $f_0(x) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, c)$.

B) For $t = 1$ to T.

a) For $i=1,2,\ldots,m$ compute $r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)}$.

b) Fit a regression tree to targets r_{it} giving terminal regions $R_{mj}, j = 1,2,...,J_t$.

c) For $j = 1,2,\ldots,J_t$ compute $c_{jt} = \underset{c}{\arg\min} \sum_{x_i \in R_{jt}} L(y_i, f_{t-1}(x_i) + c)$.

d) Update $f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{jt} I(x \in R_{jt})$.

e) Output $\hat{f}(x) = f_T(x) = \sum_{t=1}^{T} \sum_{j=1}^{J} c_{jt} I(x \in R_{jt})$.

Further information about GBRT algorithm is available at Scikit-learn.org website.

Table 1 shows the input features and regression target of GBRT. The super parameters of the model are obtained by the Bayesian hyperparametric optimization algorithm. The corrected results of the three interpolation methods are named GBRIDW, GBRMIDW, and GBRGIDW, respectively.

Table 1 Input features and regression target of GBRT

Input feature		Target
Interpolation result	Results of IDW, MIDW, or GIDW	Observation value
Terrain-related features	DEM Slope Aspect	
Time-related features	Day of year Hour of day	

2.4 Test method

This study uses the mean square error (MSE) to evaluate the error between the interpolated result and the observation value. In addition, the coefficient of determination, R^2, is used when comparing the different interpolation effects of different stations. Their computing formulas are as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{Z}_i - Z_i\right)^2, \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(\hat{Z}_i - Z_i\right)^2}{\sum_{i=1}^{n} \left(\hat{Z}_i - \bar{Z}\right)^2}, \quad (7)$$

where $\hat{Z}_i$ is the interpolation result, Z_i is the observation value, and $\bar{Z}$ is the mean of observation values.

Two different testing sets are selected to evaluate the performance of the wind interpolation in two different situations. The stratified sampling (Ghosh, 1958) is used for the situation that the interpolation results are used to fill in the missing values of the weather stations. This selecting method ensures that the probability distribution of each feature (i.e., terrain-related features and time-related features) of the testing set is approximate to that of the training set. The leave-one-group-out cross-validation (LOGO-CV) method is used for evaluating the interpolation results over an area that has never been observed. The specific steps of LOGO-CV are as follows: 1) remove all historical observations of one group of stations from the training set randomly; 2) use the remaining data set to train a GBRT model; 3) the removed observation record is used as the testing set to evaluate the accuracy of the trained GBRT; 4) put back the previously removed observations, and remove the next group's observations, and the previous steps are repeated.

2.5 Importance measurement

To find valid information in historical data, it is necessary to evaluate the importance of each input feature of the regression tree (Casalicchio et al., 2018). The values of the mean decrease impurity importance (MDI) and permutation importance are most wildly used to measure the importance of the input features (Altmann et al., 2010; Louppe et al., 2013). MDI is computed by measuring how effective the feature is at reducing variance when creating regression trees (Li et al., 2020). The permutation importance is calculated by the decrease in a model score when a single input feature is randomly shuffled. Sklearn package in Python software can calculate MDI and permutation importance easily in tree-based models. More details can be found at Scikit-learn.org website.

3 Results and discussion

3.1 Performance of IDW based interpolation methods and corrected effect by GBRT

By rule of stratified random sampling, 70% of the sample data at each station is selected as the training set, and the remaining 30% of which is selected as the testing set. Table 2 shows the interpolation evaluation of IDW, MIDW, and GIDW. It can be seen that the MSE of the interpolation by the IDW based method is generally in the range of 1.081 to 1.700 m/s, and R^2 is in the range of 0.303 to 0.419. GIDW performed better than the other two interpolation methods in all months. Shenzhen is affected by the East Asian summer monsoon; therefore, the monthly prevailing wind direction and the mean wind speed are different. However, there is no significant difference in the error of interpolation results of the 2-min average wind speed in each month. Table 3 shows the correction effect of GBRT on these three methods. It can be seen that the MSE drops to the range of 0.778 to 0.923 m/s, and R^2 increases to the range of 0.530 to 0.671. After correction, the difference between the MSEs of the interpolation by the three interpolation methods, i.e., GBRIDW, GBRMIDW, GBRGIDW, decreases, as well as the difference between the R^2 for the interpolation the three revised models. This improvement may come from that GBRT has considered the nonlinear relationship between elevation and wind speed.

Spatially, Fig. 2 shows the R^2 and MSE of each station's interpolation result for the testing set by IDW, MIDW, and GIDW methods, respectively. It can be seen that the stations located in eastern Shenzhen with relatively high DEM have a larger MSE than other stations. The stations with good interpolation performance (i.e., high R^2 or low MSE) obtained by the three interpolation methods are almost consistent.

Figure 3 shows the interpolation improvement by GBRIDW, GBRMIDW, and GBRGIDW methods in terms of ΔR^2 and ΔMSE, comparing to the interpolations by IDW, MIDW, and GIDW methods, respectively. The figure shows that most of the stations have improvements of R^2 and decreases of MSE, which means GBRT significantly corrects the interpolation error.

3.2 Evaluating interpolation effect in the non-station area

For areas that have never been observed before, the LOGO-CV method is applied to evaluate the interpolation performance. Three groups, with each group containing all the historical data of ten stations, are used in the LOGO-CV. Each group will be removed once and be

Table 2 Interpolation evaluation of IDW, MIDW, and GIDW

Month	MSE			R^2		
	IDW	MIDW	GIDW	IDW	MIDW	GIDW
Jan.	1.487	1.514	1.440	0.351	0.339	0.371
Feb.	1.425	1.459	1.393	0.364	0.349	0.379
Mar.	1.364	1.375	1.302	0.364	0.359	0.393
Apr.	1.307	1.332	1.264	0.345	0.332	0.366
May	1.265	1.309	1.240	0.327	0.303	0.340
Jun.	1.271	1.350	1.267	0.358	0.319	0.361
Jul.	1.241	1.310	1.229	0.355	0.319	0.361
Aug.	1.081	1.137	1.096	0.411	0.381	0.404
Sep.	1.278	1.319	1.254	0.408	0.389	0.419
Oct.	1.368	1.410	1.334	0.347	0.327	0.363
Nov.	1.464	1.496	1.406	0.366	0.352	0.391
Dec.	1.664	1.700	1.605	0.389	0.376	0.411
All year	1.350	1.392	1.318	0.368	0.348	0.383

Table 3 Performance evaluation of the GBRT correction

Month	MSE			R^2		
	GBRIDW	GBRMIDW	GBRGIDW	GBRIDW	GBRMIDW	GBRGIDW
Jan.	0.880	0.878	0.878	0.615	0.617	0.617
Feb.	0.885	0.879	0.880	0.606	0.608	0.608
Mar.	0.881	0.874	0.877	0.589	0.592	0.591
Apr.	0.891	0.889	0.890	0.553	0.554	0.553
May	0.883	0.882	0.883	0.530	0.531	0.530
Jun.	0.922	0.920	0.923	0.535	0.536	0.534
Jul.	0.881	0.880	0.882	0.542	0.543	0.541
Aug.	0.779	0.778	0.780	0.576	0.577	0.576
Sep.	0.844	0.841	0.840	0.609	0.610	0.611
Oct.	0.837	0.836	0.835	0.600	0.601	0.601
Nov.	0.835	0.830	0.831	0.638	0.641	0.640
Dec.	0.901	0.897	0.899	0.669	0.671	0.670
All year	0.868	0.865	0.866	0.594	0.596	0.595

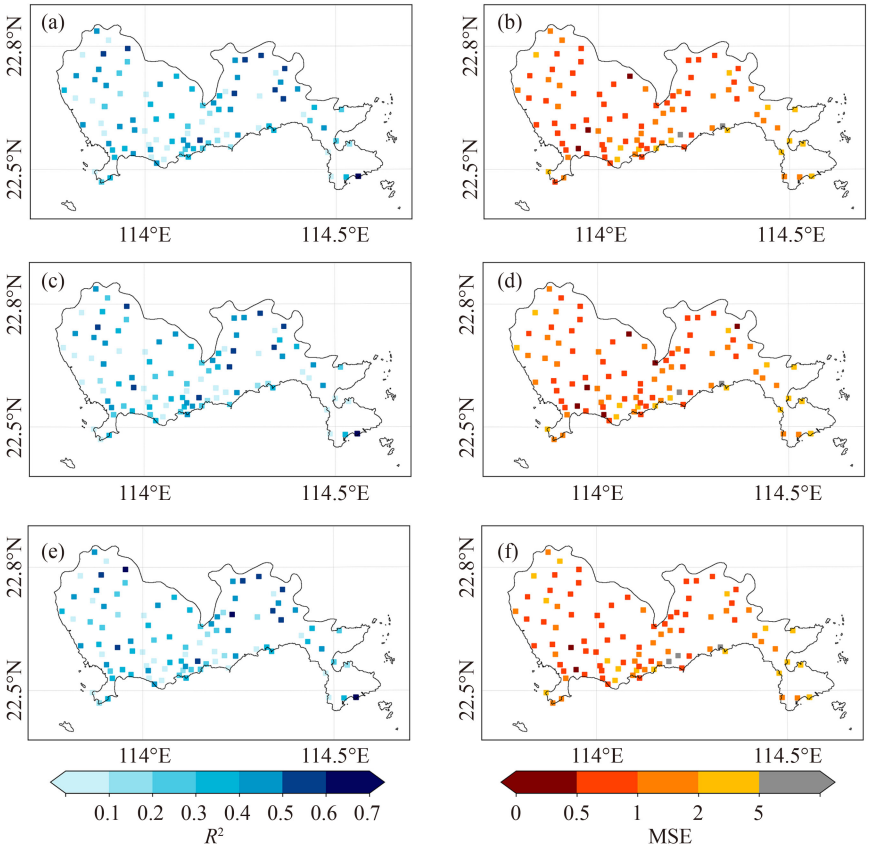

Fig. 2 The R^2 and MSE of each station's interpolation result for the testing set by IDW, MIDW, and GIDW methods. (a) (c) (e) for R^2, and (b) (d) (f) for MSE.

used to test the interpolation effect. Before the cross validation, the distribution consistency of the elevations and slope of the 100 stations and the natural terrain needs to be tested (Aguilar et al., 2005). The blue curves in Fig. 4 show the proportions of elevation and slope for Shenzhen's topography. The columns represent the proportion of the elevation and slope of 100 weather stations used in this study. Figure 4 shows that about 96% of the stations have an elevation below 100 m and about 83% of the stations have a slope below 5 degrees. By

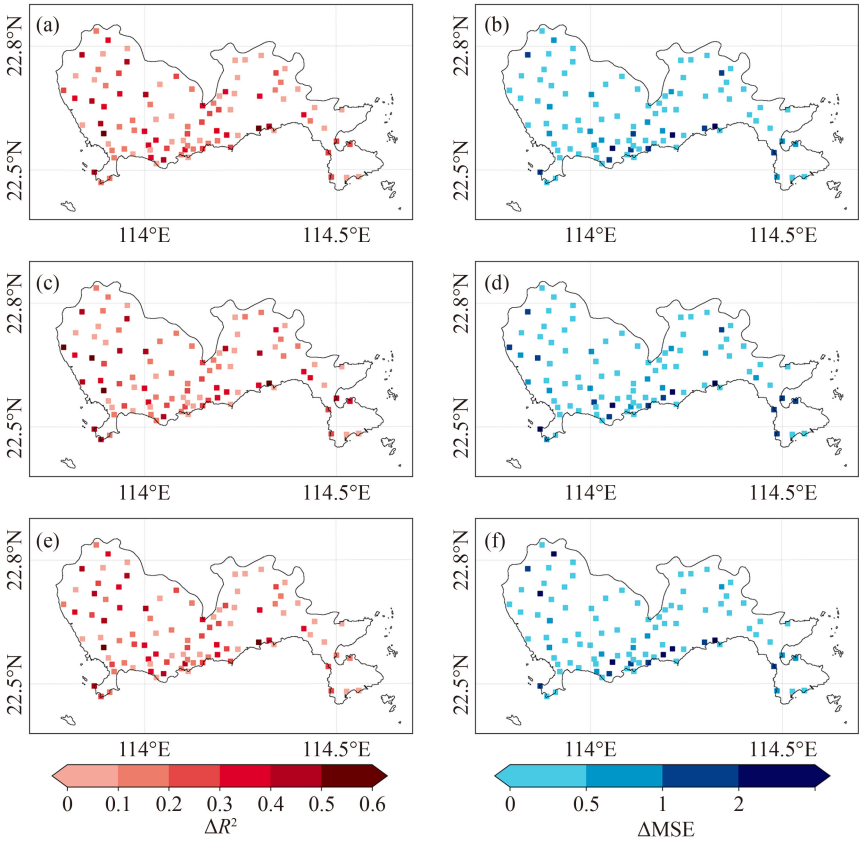

Fig. 3 Interpolation improvement by GBRIDW, GBRMIDW, and GBRGIDW methods in terms of ΔR^2 and ΔMSE, comparing to the interpolations by IDW, MIDW, and GIDW methods.

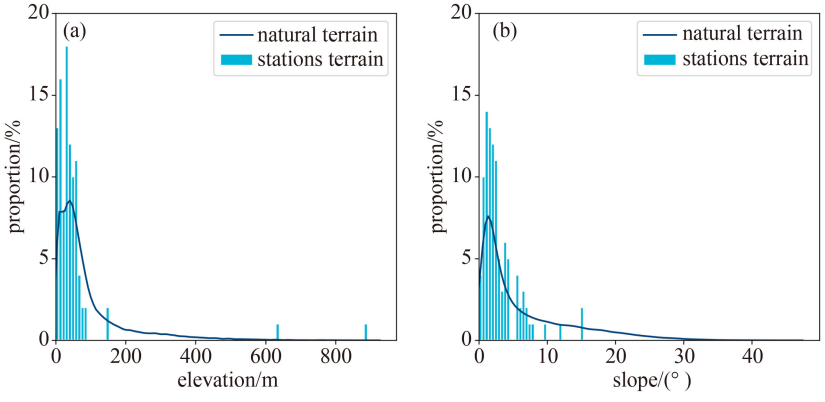

Fig. 4 Comparison of (a) elevation and (b) slope distribution between the 100 stations and the natural terrain in Shenzhen.

comparison, about 75% of the region in Shenzhen is below 100 m and about 60% of the Shenzhen area has a slope below 5 degrees. The comparison of curves and columns shows that the elevation proportion distribution of stations is generally consistent with that of the actual terrain; however, there are few representative stations with large elevations. In terms of the slope, the proportion distribution of the stations is also similar to that of the actual terrain; however, there is no representative station with a large slope of more than 20 degrees.

Figure 5 shows the locations of the stations included in the three groups. The blue square, orange triangle and

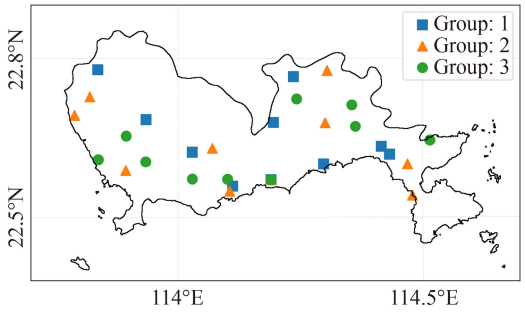

Fig. 5 The locations of the stations in the three randomly selected groups.

green disk indicate the stations belonging to groups 1, 2, and 3, respectively. The elevation and slope frequency distributions of each group are different.

Table 4 shows the interpolation evaluation of IDW, MIDW, and GIDW for each group. It can be seen that the accuracy of GIDW decreased significantly in the LOGO-CV test. This is because the regression coefficients used in GIDW are affected by the frequency distribution of stations' elevation. When the elevations of the training set and testing set do not have the same distribution, the performance of GIDW is not good. The evaluation results show that the IDW is a more robust interpolation method in the non-station area in Shenzhen.

The performance of GBRT correction is evaluated by LOGO-CV, and the results are summarized in Table 5. The comparison of Table 4 and Table 5 shows that GBRT correction is also effective for non-station areas, but the improvement effect is not as significant as that for the stratified sampling test. The difference in the elevation frequency distribution between the natural terrain and the weather stations may lead the accuracy of GIDW to decrease. Therefore, the IDW interpolation result corrected by GBRT is recommended to estimate the wind field in the non-station areas.

3.3 The importance of each input feature

Previous analysis shows that the performance of GBRT correction of the spatial wind speed interpolation depends on the historical wind observations and the topography features. The importance of these features included in Table 1 is evaluated by permutation importance and MDI. Figure 6 shows that the results obtained by the three IDW based interpolation methods, i.e., IDW, MIDW, GIDW, are the most important features. The terrain-related features (i.e., elevation, slope, and aspect) are the second important ones. The day of a year and the hour of a day slightly contribute to the model.

Table 4 Interpolation evaluation of IDW, MIDW, and GIDW

Group	MSE			R^2		
	IDW	MIDW	GIDW	IDW	MIDW	GIDW
1	1.223	1.297	1.471	0.321	0.280	0.183
2	1.392	1.598	1.631	0.326	0.227	0.211
3	1.193	1.313	1.497	0.364	0.300	0.202

Table 5 Performance evaluation of the GBRT correction

Group	MSE			R^2		
	GBR-IDW	GBRM-IDW	GBRG-IDW	GBR-IDW	GBRM-IDW	GBRG-IDW
1	1.210	1.181	1.288	0.328	0.344	0.285
2	1.356	1.384	1.453	0.344	0.330	0.297
3	1.190	1.214	1.306	0.366	0.353	0.304

The MDI and permutation importance of terrain-related features for GBRGIDW are lower than those for GBRIDW and GBRMIDW. Combined with the interpolation evaluation in Table 2, it can be speculated that GIDW has considered the inner relationship between elevation and wind speed. Therefore, it is less dependent on elevation when GBRT is used to correct GIDW. In contrast, when GBRT corrects interpolations by the IDW and MIDW methods, it is more necessary to adjust the interpolation according to the elevation of the actual terrain. In addition to accounting for the nonlinear relationship between the elevation and wind speed, GBRT correction considers the influence of slope and orientation of topography on wind speed which is ignored by the IDW based interpolation methods. From Fig. 6, it is concluded that the contribution from the time-related features (the day of a year and the hour of a day) on wind interpolation over the Shenzhen area is not significant. This is because when interpolating the wind speed over an unobserved area, the time related wind variation is already included in the surrounding known wind observation.

4 Conclusions

In this study, the GBRT model is used to revise the wind interpolation by three IDW based interpolation methods, i.e., IDW, MIDW, and GIDW. The input features for the GBRT model are the interpolation results by IDW, MIDW, and GIDW, the terrain related features, including the elevation, slope, and aspect, and the time related features, including the day of a year and the hour of a day. By using the hourly observation data of 100 weather stations in Shenzhen from January 2008 to December 2018, the wind interpolation by the three interpolation methods and their corresponding corrected results by GBRT are evaluated in terms of R^2 and MSE, respectively. The results of stratified sampling show that GBRT significantly improves the accuracy of interpolation results. This means that the GBRT model can be used to revise the interpolation results by IDW, MIDW, and GIDW methods when the interpolation results are used to fill the occasionally missing values of the observation stations.

The LOGO-CV result reflects the GBRT correction of the wind interpolation over the non-station area in Shenzhen. Even if there is no historical wind observation over these areas as a reference, GBRT is still robust to correct the IDW based wind interpolation in the non-station area. Since the difference of elevation frequency distribution between the natural terrain and the weather stations may lead the accuracy of GIDW to decrease, the IDW interpolation result corrected by GBRT is more suitable to be used in this situation.

By analyzing the importance of input features for the

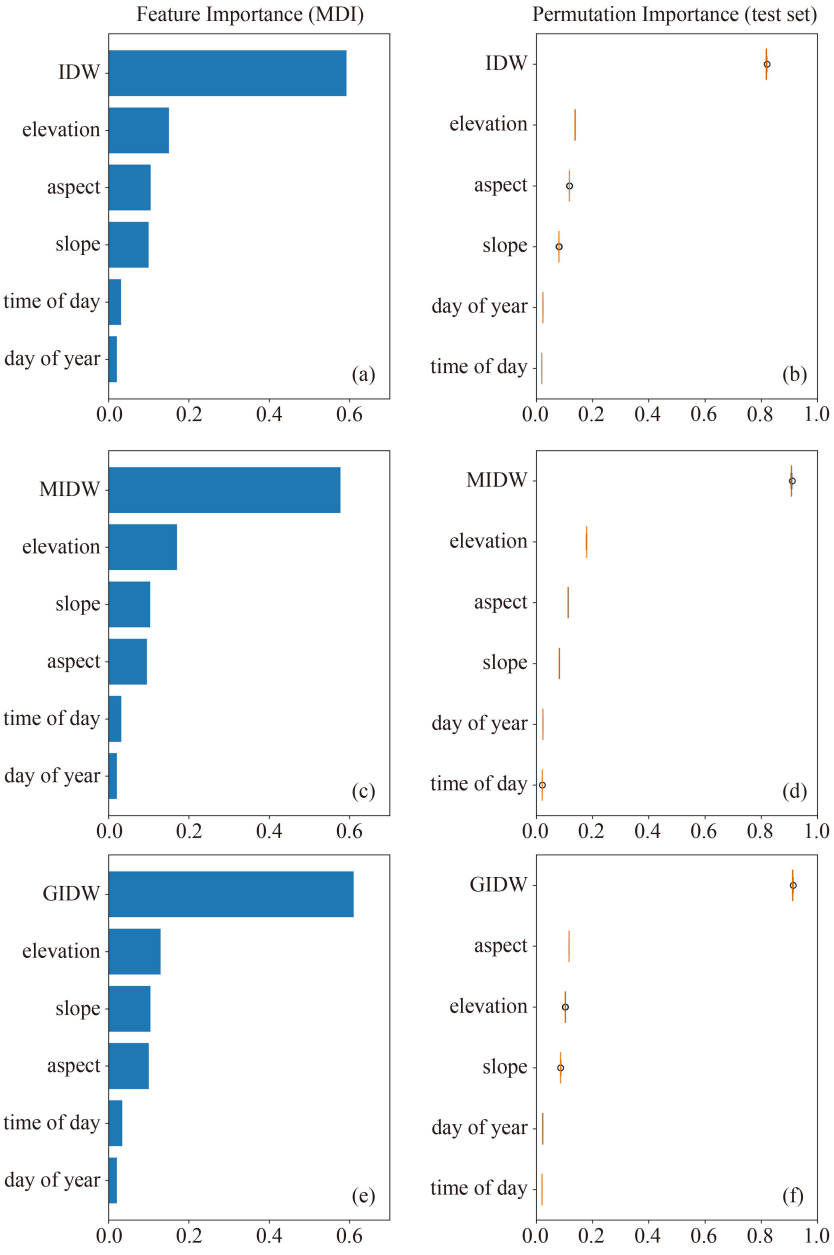

Fig. 6 The importance of each input feature. (a) and (b) are for GBRIDW; (c) and (d) are for GBRMIDW; (e) and (f) are for GBRGIDW.

GBRT model, it can be seen that elevation, aspect, and slope play important roles in the wind interpolation, following the original interpolation results by IDW, MIDW, and GIDW methods. It indicates that GBRT can reasonably integrate the historical records and topography features to correct the wind speed interpolation, which may be the reason why GBRT shows a good performance in correcting the wind speed interpolation by the IDW based methods.

Acknowledgments This study was supported by the Science and Technology Department of Guangdong Province (No. 2019B111101002) and the Innovation of Science and Technology Commission of Shenzhen Municipality Ministry (No. JCYJ 20210324101006016).

References

Aguilar F J, Agüera F, Aguilar M A, Carvajal F (2005). Effects of terrain morphology, sampling density, and interpolation methods on grid DEM accuracy. Photogramm Eng Remote Sensing, 71(7): 805–816

Altmann A, Toloşi L, Sander O, Lengauer T (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10): 1340–1347

Arain M A, Blair R, Finkelstein N, Brook J R, Sahsuvaroglu T, Beckerman B, Zhang L, Jerrett M (2007). The use of wind fields in a land use regression model to predict air pollution concentrations

for health exposure studies. Atmos Environ, 41(16): 3453–3464

Bañuelos-Ruedas F, Angeles-Camacho C, Rios-Marcuello S (2010). Analysis and validation of the methodology used in the extrapolation of wind speed data at different heights. Renew Sustain Energy Rev, 14(8): 2383–2391

Casalicchio G, Molnar C, Bischl B (2018). Visualizing the feature importance for black box models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer

Chang Q, Li S, Wang Y, Wu J, Xie M (2013). Spatial process of green infrastructure changes associated with rapid urbanization in Shenzhen, China. Chin Geogr Sci, 23(1): 113–128

Coccolo S, Mauree D, Naboni E, Kaempf J, Scartezzini J L (2017). On the impact of the wind speed on the outdoor human comfort: a sensitivity analysis. Energy Procedia, 122: 481–486

Curry C L, van der Kamp D, Monahan A H (2012). Statistical downscaling of historical monthly mean winds over a coastal region of complex terrain. I. Predicting wind speed. Clim Dyn, 38(7−8): 1281–1299

Daly C, Halbleib M, Smith J I, Gibson W P, Doggett M K, Taylor G H, Curtis J, Pasteris P P (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Intern J Climat, 28(15): 2031–2064

Elith J, Leathwick J R, Hastie T (2008). A working guide to boosted regression trees. J Anim Ecol, 77(4): 802–813

Friedman J H (2001). Greedy function approximation: a gradient boosting machine. Ann Stat, 29(5): 1189–1232

Friedman J H (2002). Stochastic gradient boosting. Comput Stat Data Anal, 38(4): 367–378

Ghosh S P (1958). A note on stratified random sampling with multiple characters. Calcutta Statistical Association Bulletin, 8(2−3): 81–90

Jarvis A, Reuter H I, Nelson A, Guevara E (2008). SRTM 90m Digital Elevation Database v4. 1

Johnson H K (1999). Simple expressions for correcting wind speed data for elevation. Coast Eng, 36(3): 263–269

Kikumoto H, Ooka R, Sugawara H, Lim J (2017). Observational study of power-law approximation of wind profiles within an urban boundary layer for various wind conditions. J Wind Eng Ind Aerodyn, 164: 13–21

Kleerekoper L, Van Esch M, Salcedo T B (2012). How to make a city climate-proof, addressing the urban heat island effect. Resour Conserv Recycling, 64: 30–38

Li J, Tian Y, Zhu Y, Zhou T, Li J, Ding K, Li J (2020). A multicenter random forest model for effective prognosis prediction in collaborative clinical research network. Artif Intell Med, 103: 101814

Li L (2019). Geographically weighted machine learning and downscaling for high-resolution spatiotemporal estimations of wind speed. Remote Sens, 11(11): 1378

Li L, Qian J, Ou C Q, Zhou Y X, Guo C, Guo Y (2014). Spatial and temporal analysis of air pollution index and its timescale-dependent relationship with meteorological factors in Guangzhou, China, 2001–2011. Environ Pollut, 190: 75–81

Liu C, Li Q, Zhao W, Wang Y, Ali R, Huang D, Lu X, Zheng H, Wei X (2020). Spatiotemporal characteristics of near-surface wind in Shenzhen. Sustainability, 12(2): 739

Louppe G, Wehenkel L, Sutera A, Geurts P (2013). Understanding variable importances in forests of randomized trees. In: Proceedings of the 26th International Conference on Neural Information Processing Systems: 431–439

Masseran N, Razali A M, Ibrahim K (2012). An analysis of wind power density derived from several wind speed density functions: the regional assessment on wind power in Malaysia. Renew Sustain Energy Rev, 16(8): 6476–6487

McCutchan M H, Fox D G (1986). Effect of elevation and aspect on wind, temperature and humidity. J Clim Appl Meteorol, 25(12): 1996–2013

Mitas L, Mitasova H (1999). Spatial interpolation. In: Longley P A, Goodchild M F, Maguire D J, Rhind D W, eds. Geographical Information Systems: Principles, Techniques, Management and Applications. New York: Wiley

Morris H, Wardle M, Mainelli M (2020). The global financial centres index 28. Long Finan Financ Centr Future. 1−69

Nalder I A, Wein R W (1998). Spatial interpolation of climatic normals: test of a new method in the Canadian boreal forest. Agric Meteorol, 92(4): 211–225

Ozelkan E, Chen G, Ustundag B B (2016). Spatial estimation of wind speed: a new integrative model using inverse distance weighting and power law. Int J Digit Earth, 9(8): 733–747

Palutikof J P, Kelly P M, Davies T D, Halliday J A (1987). Impacts of spatial and temporal windspeed variability on wind energy output. J Clim Appl Meteorol, 26(9): 1124–1133

Peng S (2017). Optimized study on spatial interpolation methods for meteorological element. Geospat Inform, 15(7): 86−89 (in Chinese)

Tobler W R (1970). A computer movie simulating urban growth in the Detroit region. Econom Geogra, 46(sup1): 234–240

Wang J, Hu J, Ma K (2016a). Wind speed probability distribution estimation and wind energy assessment. Renew Sustain Energy Rev, 60: 881–899

Wang P, Dou Y, Xin Y (2016b). The analysis and design of the job recommendation model based on GBRT and time factors. In: 2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA). IEEE: 29–35

Wang Y, Tang Y (2019). A recommendation algorithm based on item genres preference and GBRT. In: Journal of Physics Conference Series. IOP Publishing, 2019, 1229(1): 012053

Wanninkhof R (2014). Relationship between wind speed and gas exchange over the ocean revisited. Limnol Oceanogr Methods, 12(6): 351–362

Winstral A, Jonas T, Helbig N (2017). Statistical downscaling of gridded wind speed data using local topography. J Hydrometeorol, 18(2): 335–348

Xu Y, Wan X, Fu C, Liu C (2012). Wind speed interpolation under complex terrain conditions: a case study of Jilin Province. Yunnan Geogr Environ Res, 24 (4), 78−81 (in Chinese)

Yang J, Shi B, Shi Y, Marvin S, Zheng Y, Xia G (2020). Air pollution dispersal in high density urban areas: research on the triadic relation of wind, air pollution, and urban form. Sustain Cities Soc, 54: 101941

Yuan W, Xu B, Chen Z, Xia J, Xu W, Chen Y, Wu X, Fu Y (2015).

Validation of China-wide interpolated daily climate variables from 1960 to 2011. Theor Appl Climatol, 119(3−4): 689–700

Zhang Z, Yang W, Wushour S (2020). Traffic accident prediction based on LSTM-GBRT model. J Contr Sci Eng, 2020: 4206919

Zhi S, Qian G, Luo J (2001). The variation of coastal wind with height in Guangdong Province. Trop Geogr, 21(2): 131–134