A COMPARATIVE STUDY OF FEATURE SELECTION METHODS FOR WIND SPEED

Nguyen Thi Hoai Thu*, Pham Nang Van, Nguyen Vu Nhat Nam, Pham Hai Minh, Phan Quoc Bao
Hanoi University of Science and Technology

ARTICLE INFO

Received: 21/01/2022

Revised: 19/4/2022

Published: 21/4/2022

KEYWORDS

Feature selection
Wind speed
Pearson's Correlation
Random Forest
Boruta

ABSTRACT

Forecasting wind speed or capacity of wind power is playing an important role to serve the problem of resource mobilization of the power system. However, forecasting is still a difficult problem because there are many factors affecting wind speed. In this paper, our goals are using some feature selection methods to find the best approach as well as to select the meteorological parameters that have great influence on wind speed, thereby helping to improve our predictive model. Pearson's Correlation, Random Forest, Boruta were 3 feature selection methods to be used on 2 weather datasets in 2 different locations. Firstly, each data was analyzed with separate autocorrelation and partial autocorrelation analysis. From this, the hysteresis characteristics of the data were obtained then added to the methods. Following that, we carried out and compared the performance of the feature selection methods based on the evaluation criteria of each method. The results show that the wind speed depends heavily on the lags closest to it and in different geographical locations gives different results.

MỘT NGHIÊN CỨU SO SÁNH VỀ CÁC PHƯƠNG PHÁP LỰA CHỌN ĐẶC TRƯNG CHO VẬN TỐC GIÓ

Nguyễn Thị Hoài Thu*, Phạm Năng Văn, Nguyễn Vũ Nhật Nam, Phạm Hải Minh, Phan Quốc Bảo
Trường Đại học Bách khoa Hà Nội

THÔNG TIN BÀI BÁO

Ngày nhận bài: 21/01/2022

Ngày hoàn thiện: 19/4/2022

Ngày đăng: 21/4/2022

TỪ KHÓA

Lựa chọn đặc trưng
Vận tốc gió
Phương pháp hệ số Pearson
Thuật toán Random forest
Phương pháp Boruta

TÓM TẮT

Dự báo vận tốc gió hay công suất của điện gió đang là vấn đề được quan tâm lớn hiện nay để phục vụ bài toán huy động nguồn của hệ thống điện. Tuy nhiên, việc dự báo vẫn còn gặp khó khăn do có rất nhiều yếu tố ảnh hưởng tới vận tốc gió. Mục đích của bài báo này là sử dụng các phương pháp lựa chọn đặc tính khác nhau để xem xét ảnh hưởng của các yếu tố thời tiết tới vận tốc gió, từ đó cải thiện độ chính xác cho việc dự báo vận tốc hay công suất gió. Các phương pháp được đề cập đến là Pearson's Correlation, Random Forest và Boruta và được sử dụng trên hai tập dữ liệu thời tiết tại hai thành phố khác nhau. Đầu tiên, chúng tôi sử dụng hàm tự tương quan và hàm tương quan riêng để có thể xem xét tổng quan mối quan hệ giữa vận tốc gió và các trễ của chính nó trong quá khứ. Tiếp theo, các trễ có ảnh hưởng lớn và trễ bậc nhất của các yếu tố thời tiết khác được sử dụng làm đầu vào cho ba phương pháp lựa chọn đặc tính. Cuối cùng, chúng tôi so sánh kết quả giữa các phương pháp với nhau. Kết quả thu được cho thấy tốc độ gió phụ thuộc rất lớn vào chính nó ở các trễ gần nhất.

DOI: https://doi.org/10.34238/tnu-jst.5487

* Corresponding author. *Email: thu.nguyenthihoai@hust.edu.vn*

1. Introduction

Due to the energy crisis and environmental issues, renewable energy (RE) has attracted global attention and becomes an alternative to fossil-based energy sources all over the world in the last decade [1]. Among various kinds of RE resources, wind energy is one of the important RE sources with the rapid development and significant penetration into power systems [2], [3]. However, the intermittency of wind power makes power system operation and control more challenging [3]. To effectively integrate wind energy into systems, the wind speed (WS) and wind power (WP) need to be accurately predicted.

In recent years, AI-based models are well developed and widely used for forecasting WS/WP. However, when developing and applying these models, several following challenges have been raising: (1) the complexity of models and (2) the volume of data [4]. WS/WP is usually predicted based on its historical data and the meteorological data which are commonly represented by extensive time series with multi-dimensions, data noises, and redundancy or lack of attributes. Moreover, using a very complex AI-based forecasting model with a huge volume of input data to train the model is not always feasible because of the limited calculation time and memory [5], [6]. Therefore, approaches for reducing the number of features in the model to eliminate extraneous and redundant data, avoid the loss of relevant information, and reduce computation time are necessary. Feature selection (FS) is usually considered as an efficient solution for this problem, which plays an important role in the data analysis process.

The basic feature selection method can be analyzed into 3 categories based on interaction with the model machine: Filter, Wrapper, Embedded method [5]-[7]. In WS/WP forecasting, various studies were conducted using a certain FS method to find out the features affecting the WS/WP to improve the forecasting accuracy as well as decrease the computational cost. In [2], Jiang et al. proposed a feature selection method using kernel density estimation (KDE)-based Kullback-Leibler divergence (KLD) and energy measure to reduce the influence of the illusive components. The original features include the subset of components after decompositions and their lags. The results of the FS process provided the subseries with lags of 3 to 6 to conduct a one-step prediction. Similarly, the optimal feature subset was selected from the combinations of the original features which were constructed from all the intrinsic mode functions and the residual after the decomposition of the wind speed series [3]. In other studies, the effects of meteorological variables such as air temperature, wind direction, relative humidity, incoming and reflected shortwave to the WS were considered using ReliefF method [8].

However, studies on the influence of weather parameters on the WP/WS have been still limited. Generally, the FS for WS/WP was usually based on time series or decomposed series of wind speed and their lags. Moreover, almost all studies have used only one method for FS without comparison. Therefore, in this paper, we aimed at studying different FS methods for WS to access the attributes of other meteorological parameters into the WS, find out the elements which have a significant impact on WS. Additionally, a comparison analysis was carried out using 2 weather datasets in Osaka (Japan) and Basel (Switzerland).

The paper is organized as follows. First, we briefly introduced different methods for FS in section 2. We use 3 methods, namely Pearson Correlation (PC), Random Forest, and Boruta which represented for 3 categories of Filter, Wrapper, Embedded methods, respectively. The simulation results and discussions were presented in section 3. Finally, conclusions are drawn in Section 4.

2. Feature selection methods

As mentioned above, Feature Selection (FS) is an important step in data analysis to improve the accuracy of the model while reducing the computational burden and complexity of the model. The basic feature selection methods can be classified into 3 types: Filter method, Wrapper

method, and Embedded method. Filter method is based on the different characteristics of the data to estimate and select a subset of the features, using evaluation measures extracted from the data set, such as distance, information, dependency, consistency [9], [10]. The wrapper method needs a machine learning algorithm and uses the performance of the algorithm as a benchmark for evaluation. This method finds the features that are best to machine learning algorithms for the purpose of improving data opening performance. This method uses predictive accuracy to classify features. Some of the commonly used methods here are Forward Feature Selection, Backward Feature Elimination, and Recursive Feature Elimination (RFE) [11]. Embedded feature selection techniques are quite similar to Wrapper, based on different classifiers, predictors, or clustering procedures. Some of the more commonly known methods include: L1(LASSO) [12], Random Forest(RF) [13], [14] or Decision Tree algorithm [15]. Figure 1 shown the block diagram of three methods of feature selection.

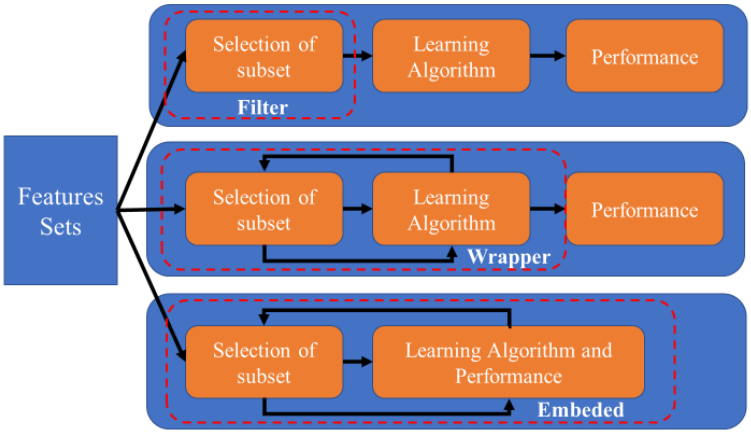


Figure 1. *Block diagram of feature selection methods*

2.1. Pearson's Correlation (PC)

The Pearson's correlation coefficient is a measure of the linear relationship between two interval- or ratio-level variables. Its value ranges from -1 to +1. The closer the value is to the boundary, the higher the correlation. A positive value of the correlation coefficient indicates the covariance of one variable with another. In contrast, a negative value of the correlation coefficient gives the inverse among the analyzed variables. Values of the correlation coefficient close to 0 mean that the variables have almost no correlation, and values close to -1 or +1 indicate a strong linear association between the two variables [16].

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \quad (1)$$

Where *Cov(X, Y)* indicates the covariance of *X* and *Y*, *Var(X)* and *Var(Y)* denote the variance of *X* and *Y*, respectively. Correlation methods such as the ones above will usually favor any quantitative measurement performed on two or more variables simultaneously, the relationship between the two variables is linear, and both must be analyzed. standard distribution. However, it would be unwise to try to compute the correlation to a nonlinear relationship. In this paper, the relationship between weather characteristics and wind speed is nonlinear.

2.2. Random Forest (RF)

Random forest is a supervised learning algorithm that can solve both regression and classification problems. Random Forest algorithm builds many decision trees using the Decision Tree algorithm, but each decision tree will be different (with random factor). The prediction results are then aggregated from the decision trees. The weaknesses of the decision tree, whose

results fluctuate considerably depending on the training data, are compensated for in continuous learning and have a feature that focuses more on wrong answers from prior learning [15], [17].

In the Random forests algorithm used in this paper, we have 2 factors to consider calculating model accuracy: Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDI). The MDI assumes that the amount of impurity reduction when the individual variable is decided as the partition node is the contribution in the random forests. Classification tree using Gini coefficient index or information collection tree and regression tree using mean of variables to remove impurities. The equation (2) calculates the importance of variable x_j. To calculate variable importance for the MDI method, it adds up the decrease of Gini index of each of the variables from 1 to n_{tree}, which means the number of trees, and gets the average of all. MDA is a method of calculating the significance of a variable by permutation and it uses Out-of-bag (OOB) to split its sample data. With $t \in \{1,2,3,\dots,n_{tree}\}$, the importance of the variable x_j in the tree t is the mean of the difference between the predicted class before the permuting x_j, which is $y_i = f(x_i)$, and after that, which is $y_i = f(x_i^j)$, for a given observation *I* (Eq(3)).

$$MDI(x_j) = \frac{1}{n_{tree}} [1 - \sum_{k=1}^{n_{tree}} Gini(j)^k] \quad (2)$$

$$MDA(x_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \frac{\sum_{i \in OOB} I(y_i = f(x_i)) - \sum_{i \in OOB} I(y_i = f(x_i^j))}{|OOB|} \quad (3)$$

2.3. Boruta

Boruta is the algorithm to be designed as a wrapper method that combines with the Random Forest algorithm. The original data set is expanded by adding so-called drop shadows valuable features are randomly swapped between training instances to eliminate correlation with a decision variable. Estimating the significance of a calculated feature because of the loss of classification accuracy caused by random permutations of feature values of cases. First and foremost, the loss of classifier accuracy is calculated individually for all decisions trees in the forest use a certain trait to classify cases, and then an average and the standard deviation of the classification accuracy loss is calculated. An important measure is the Z-score calculated by dividing the mean loss by its standard deviation. The importance measure is used to determine the ranking of features [18]. The criterion for assessing the importance of this method is the Z score can be written as Eq. (4).

$$Z = \frac{x - \mu}{\sigma} \quad (4)$$

Where x is the raw score, μ is the population mean, and σ is the population standard deviation. After calculating Z score for each feature, maximum Z-score (MZS) between the defined shadow features and a specified hit for every feature that scores better than the MZS. The two-sided equality test with MZS is applied. Features are important significantly lower than the MZS are considered irrelevant (rejected). Features with significantly higher importance than MZS are considered relevant (confirmed) features. The remaining features are treated as tentative ones.

3. Result and discussion

3.1.Case study

In general, for feature selection, a lot of meteorological data need to be used. The time series should be at least 1 year to cover the possible seasonal characteristic. However, it is not necessary to use too much data due to the time consuming and the difficulty in collecting data in a long time. Moreover, the purpose of the study is to find out which parameters have the great impact on wind speed, not the precise relationship between them. Therefore, in this study, we used the

weather dataset of two regions: Basel and Osaka during 5 years from 2010 to 2014 [19], [20] to determine the statistics of the characteristics that affected the wind speed. The weather characteristics of Basel and Osaka are given in Table 1 and 2, respectively. While weather data for Osaka is collected by the Japan Meteorological Agency (JMA), weather data for Basel is collected by Meteoblue - a meteorological service created at the University of Basel, Switzerland, in partnership with the US National Oceanic and Atmosphere Administration and the National Center for Environmental Prediction located in Basel city. Both data are recorded hourly.

In this paper, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) functions are used to determine the lagging features of the wind speed itself in Basel and Osaka, shown in Figure 2 and 3.

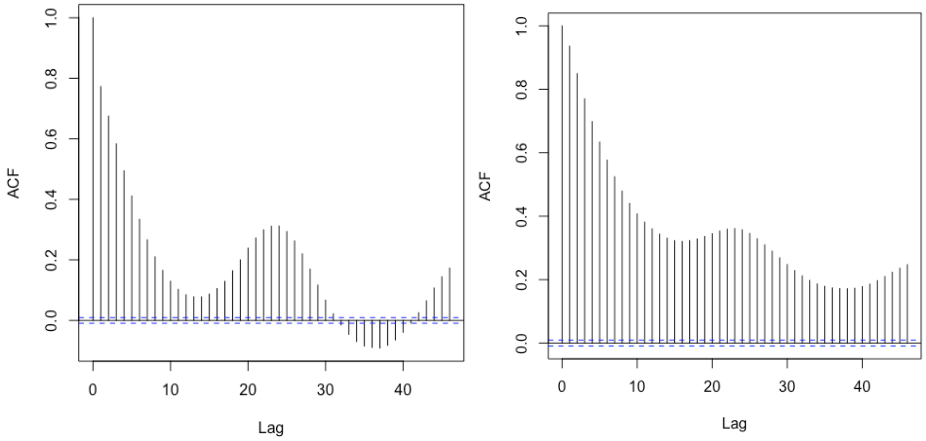
Figure 2 and 3 reveal information on the relationship of wind speed to itself at different lags. First of all, Osaka's autocorrelation function oscillated gradually according to the exponential law while with that of Basel, we couldn't conclude yet, and these PACF plots showed that the partial correlation function of Basel was the damped sine wave law while that of Osaka was the exponential damping oscillation. In these partial correlation function plots, the wind speed of Basel depended a lot on lag 1, lag 2, and lag 3 while that of Osaka depended almost on lag 1, lag 2. The reason why the ACF and PACF were to be considered is that these plots helped us to see which wind speed lags had a great influence on itself at the time t. After filtering out the important lags, they were combined with the first lag of other features to form a set of lagged features and used the above methods to find the good ones.

Table 1. *Descriptive statistics of raw historical weather data set of Osaka city from 2010 to 2014*

	Pressure (hPa)	Air Temp. (°c)	Dewpoint (°c)	Humidity (%)	Wind speed (m/s)	Hours sunlight (h)	Solar radiation (MJ/m^2)	Cloudiness
Mean	1005.17	16.94	9.35	62.71	2.49	0.45	0.59	6.72
St.dev	6.7	8.95	9.57	15.93	1.42	0.43	0.89	3.64
Min.	971	-2.7	-14.2	10	0	0	0	0
Max.	1024	38.1	26.2	99	11.7	1	4.11	10

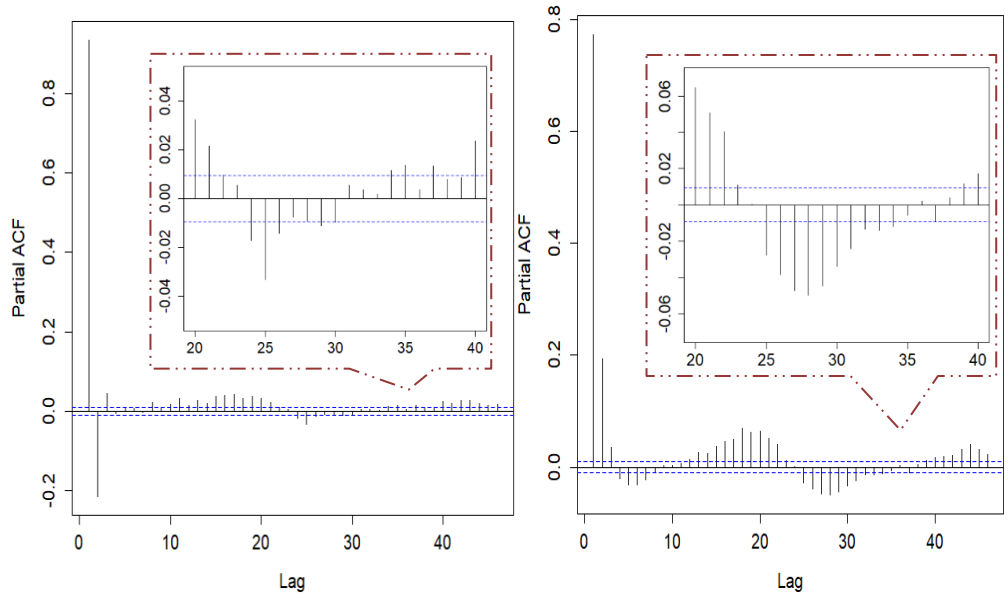
Table 2. *Descriptive statistics of raw historical weather data set of Basel city from 2010 to 2014*

	Temp. (°c) [2 m]	Humidity (%) [2 m]	Wind Speed (km/h)	Wind Direction (°) [10 m]	Cloud Cover Total	Solar Radiation (W/m^2)	Direct Radiation (W/m^2)	Diffuse Radiation (W/m^2)	Mean Sea Level Pressure (hPa)
Mean	11.78	72.36	10.62	201.41	53	163.07	96.55	66.52	1016.22
St.dev.	7.68	14.95	7.87	96.09	45.75	230.13	150.89	86.6	7.99
Min.	-12.52	17	0	0.64	0	0	-9.07	0	976.1
Max.	36.22	100	74.34	360	100	890.89	605.12	302.04	1040.5



(a) *b)*

Figure 2. *ACF plots of the wind speed: (a) Basel and (b) Osaka*



(a) *(b)*

Figure 3. *PACF plots of the wind speed in: (a) Basel and (b) Osaka*

3.2. Comparision and discussion

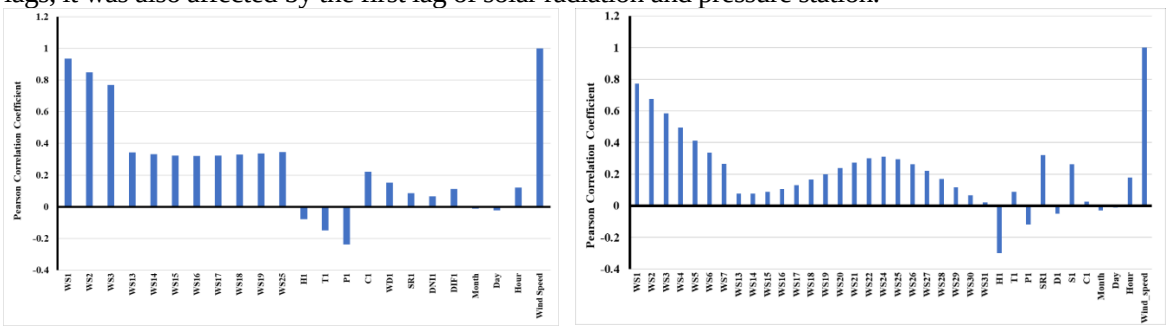
To achieve our goals, we developed programs for the above methods using RandomForest package in R which is an open-source software environment [21]. First, the correlation of wind speed was calculated by running the program according to Pearson's correlation method and got the result as shown in Figure 4. Except for the time features which are month, hour, and day, the rest of the features were all considered lags because our goal was finding the relationship between the wind speed at present and all the features in the past to be able to make accurate forecasts. And feature month, hour, day to stay at the current time to be able to see if time has a lot of influence on wind speed or not. Each of the past lags we denote by numbers, for example solar radiation at the first lag and wind speed at the third lag are SR1 and WS3, respectively. It is interesting that the wind speed in Osaka depended greatly on its first 7 lags, besides it was also affected by humidity at first lag and solar radiation at first lag while that of Basel depended almost entirely on its first 3 lags. But the correlation of wind speed lags from 13 to 25 was higher than the rest of the other features, which points that wind speed at this city to be affected almost by itself lags. However, if we compared the role of wind speed at first lag between the two cities, the highest importance index in Osaka would have significantly lower than that of Basel. Part of the reason is that in the data processing, Basel had no missing data while Osaka has quite a lot of Na values.

After Pearson's correlation was used, the second method is the Random Forest model to be applied. The number of trees to grow in both data was 300. The reason why choosing this number of trees to be used was that if less then there might not be enough volume to give a suitable result while more could also lead to too large volume making it difficult during program running. Besides, the number of variables randomly sampled as candidates at each split in Basel and Osaka are 12 and 16, respectively. The effectiveness of the results were shown through two criteria in Figure 5. Overall, the dependence of wind speed of both two cities had many similar results of Pearson's correlation. Considering the city of Basel, in the MDA and MDI charts, the figure for the wind speed at first lag had the largest significant score far ahead of other features, while that of wind speed at lags 2 and 3 had the next highest scores. Some other features that

have high indexes in both two graphs were hour and temperature at first lag. And in Osaka city, the figure for the wind speed at first lag appeared to be by far the largest score, which was followed by that of solar radiation at first lag, wind speed at lag 2, and pressure station at first lag in the MDA chart. Another striking feature in Osaka's MDA chart is that the scores of Month and Day were respectively 8 and 45 times smaller than the score of the wind speed in the first lag and did not appear in the chart. In the MDI chart, wind speed at lags 1, 2, 3, 4 and solar radiation at first lag had a significantly higher score than the rest. In this city, the ordering of features in two different graphs with different scores made it difficult to choose the feature ranking. However, the gap between features in the MDI chart is much clearer than in the MDA, so we have preferred the results of the MDI chart over the MDA in Osaka city.

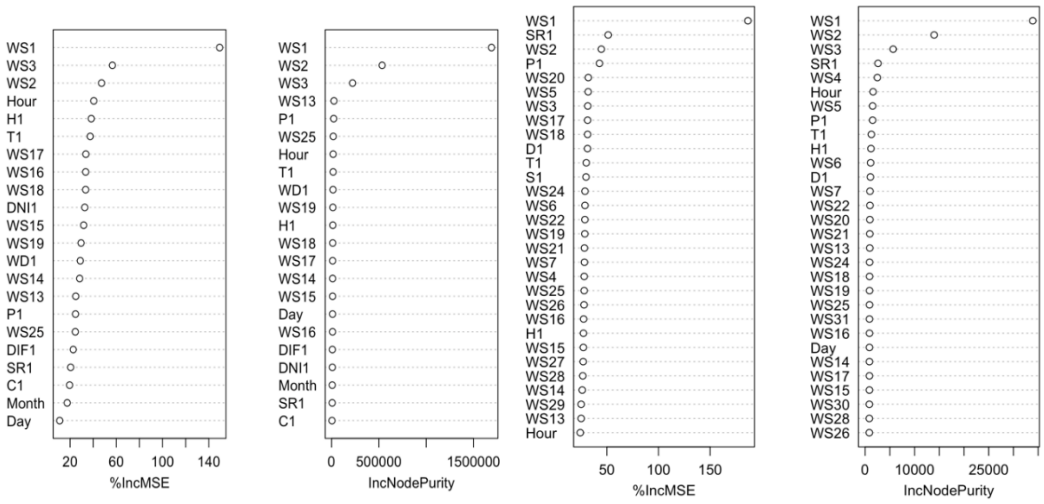
Finally, the Boruta method was put into operation to compare with the above 2 methods. Surprisingly, no features were rejected or in other words, all features had a higher Z-score than the maximum Z-score of its shadow features. In Figure 6a, the figure given for the results of this method was slightly similar to that of the random forests algorithm with the wind speed at first lag having the highest significant score and wind speed at lag 2 and 3 in the next highest scores. And in Figure 6b, the wind speed in Osaka was strongly influenced by itself at first lag. Following were other features such as solar radiation at first lag, wind speed at lag 2, hour, pressure station at first lag, and temperature at first lag.

Although the two cities had different features, it is shown that the wind speed depended greatly on itself at the first lag. In Basel almost the wind speed depended on itself in the first three lags and other features have little influence. In Osaka, in addition to the influence of wind speed in the first three lags, it was also affected by the first lag of solar radiation and pressure station.



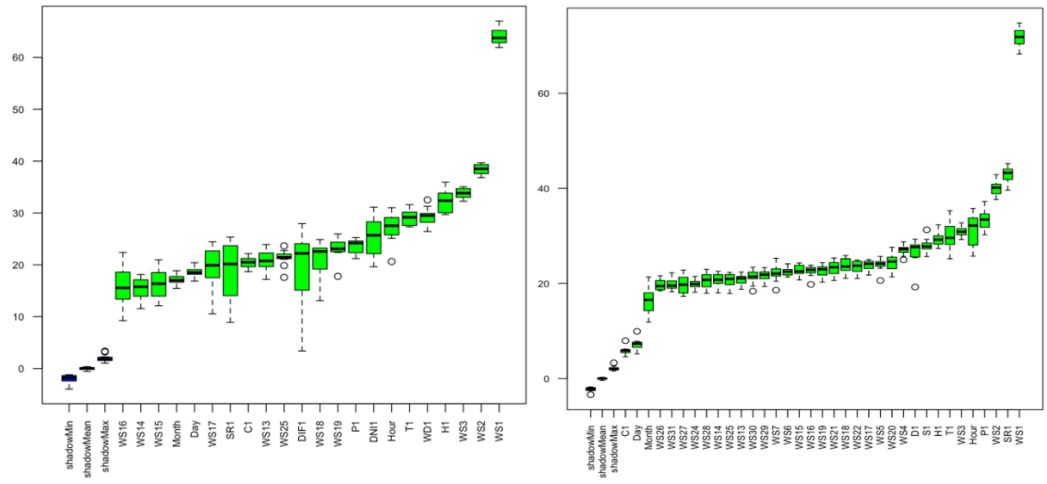
(a) *(b)*

Figure 4. *Correlation plots of the wind speed: (a) Basel and (b) Osaka*



(a) *(b)*

Figure 5. *MDA and MDI plots of the wind speed: (a) Basel and (b) Osaka*



(a) *(b)*

Figure 6. *Boruta plot of the wind speed: (a) Basel and (b) Osaka*

Table 3. *Summary the rank of features in two cities*

Feature ranking	Basel			Osaka		
	Pearson's correlation	***Random forest***	***Boruta***	***Pearson's correlation***	***Random forest***	***Boruta***
1	WS1	WS1	WS1	WS1	WS1	WS1
2	WS2	WS2	WS2	WS2	WS2	SR1
3	WS3	WS3	WS3	WS3	SR1	WS2
4	WS25	Hour	H1	WS4	WS3	P1
5	WS13	T1	WD1	WS5	P1	Hour

In terms of methods, the results of the Random Forest algorithm were nearly identical to the Boruta method in both cities. It is interesting to note that almost the important feature of Pearson's correlation method was the lags of the wind speed while the results of the remaining 2 methods had other features at first lag. The reason is that Pearson's correlation method was not suitable for large and nonlinear data sets while the Boruta method and Random Forest algorithm could overcome this drawback.

4. Conclusion

In this article, feature selection methods were used to find weather features affecting wind speed in 2 different cities gives the same result, which was highly dependent on the first 3 lags of wind speed. First of all, the wind speed of each data was input for ACF and PACF function to get an overall of lagging wind speed affected to itself. After that, 11 lags of wind speed with the first lag of others weather in Basel and 25 lags of wind speed with the first lag of the rest weather features to be added to each method. Finally, results of three methods and two cities were compared with each other to find out how effective they are. The results showed that Random Forest and Boruta methods seemed to be better than Pearson's correlation method in both two data. One of the reasons is that there were a number of data inputs, and they were non-linear, which was more suitable for Random Forest and Boruta method than Pearson's correlation. But Pearson's correlation had a much faster program running speed than the rest. In future study, we will use these feature selection results to decrease the input data of the wind speed forecasting model as well as improve the forecasting accuracy.

Acknowledgements

This research is funded by Hanoi University of Science and Technology (HUST) under grant number T2021-PC-004.

REFERENCES

[1] T. H. T. Nguyen, T. Nakayama, and M. Ishida, "Optimal capacity design of battery and hydrogen system for the DC grid with photovoltaic power generation based on the rapid estimation of grid dependency," *Int. J. Electr. Power Energy Syst.*, vol. 89, pp. 27-39, Jul. 2017, doi: 10.1016/j.ijepes.2016.12.012.

[2] Y. Jiang and G. Huang, "Short-term wind speed prediction: Hybrid of ensemble empirical mode decomposition, feature selection and error correction," *Energy Convers. Manag.*, vol. 144, pp. 340-350, Jul. 2017, doi: 10.1016/j.enconman.2017.04.064.

[3] C. Zhang, H. Wei, J. Zhao, T. Liu, T. Zhu, and K. Zhang, "Short-term wind speed forecasting using empirical mode decomposition and feature selection," *Renew. Energy*, vol. 96, pp. 727-737, Oct. 2016, doi: 10.1016/j.renene.2016.05.023.

[4] I. M. Müller, "Feature selection for energy system modeling: Identification of relevant time series information," *Energy AI*, vol. 4, p. 100057, Jun. 2021, doi: 10.1016/j.egyai.2021.100057.

[5] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, "Embedded Methods," in *Feature Extraction: Foundations and Applications*, I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds. Berlin, Heidelberg: Springer, 2006, pp. 137-165, doi: 10.1007/978-3-540-35488-8_6.

[6] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Assessing feature selection method performance with class imbalance data," *Mach. Learn. Appl.*, p. 100170, Oct. 2021, doi: 10.1016/j.mlwa.2021.100170.

[7] U. Stańczyk, "Feature Evaluation by Filter, Wrapper, and Embedded Approaches," in *Feature Selection for Data and Pattern Recognition*, U. Stańczyk and L. C. Jain, Eds. Berlin, Heidelberg: Springer, 2015, pp. 29-44, doi: 10.1007/978-3-662-45620-0_3.

[8] K. P. Senthil and D. Lopez, "Feature Selection used for Wind Speed Forecasting with Data Driven Approaches," *J. Eng. Sci. Technol. Rev.*, vol. 8, no. 5, pp. 124-127, Oct. 2015, doi: 10.25103/jestr.085.17.

[9] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2015, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458.

[10] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, Oct. 2007, doi: 10.1093/bioinformatics/btm344.

[11] W. Liu and J. Wang, "Recursive elimination-election algorithms for wrapper feature selection," *Appl. Soft Comput.*, p. 107956, Oct. 2021, doi: 10.1016/j.asoc.2021.107956.

[12] J. Guenther and O. Sawodny, "Feature Selection for Thermal Comfort Modeling based on Constrained LASSO Regression," *IFAC-Pap.*, vol. 52, no. 15, pp. 400-405, 2019, doi: 10.1016/j.ifacol.2019.11.708.

[13] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123-140, Aug. 1996, doi: 10.1007/BF00058655.

[14] Jerome Friedman Trevor Hastie Robert Tibshirani, *The elements of statistical learning: Data Mining, Inference and Prediction*, 2nd ed. Springer 2009. [Ebook] Available: https://link.springer.com/book/10.1007/978-0-387-84858-7

[15] L. Breiman and J. H. Friedman, *Classification And Regression Trees*, Taylor and Francis group, 2017 [Ebook]. Available: https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman-jerome-friedman-richard-olshen-charles-stone. [Accessed Oct. 10, 2021].

[16] W. Kirch, Ed., "Pearson's Correlation Coefficient," in *Encyclopedia of Public Health*, Dordrecht: Springer Netherlands, 2008, pp. 1090-1091, doi: 10.1007/978-1-4020-5614-7_2569.

[17] M. Aria, C. Cuccurullo, and A. Gnasso, "A comparison among interpretative proposals for Random Forests," *Mach. Learn. Appl.*, vol. 6, p. 100094, Dec. 2021, doi: 10.1016/j.mlwa.2021.100094.

[18] H. Kaneko, "Examining variable selection methods for the predictive performance of regression models and the proportion of selected variables and selected random variables," *Heliyon*, vol. 7, no. 6, p. e07356, Jun. 2021, doi: 10.1016/j.heliyon.2021.e07356.

[19] Meteoblue, "Weather History Download Basel", 2021 [Online]. Available: https://www.meteoblue.com/en/weather/archive/export/basel_switzerland_2661604?daterange=2021-10-06%20-%202021-10-13&domain=NEMSAUTO&min=2021-10-06&max=2021-10-

13&utc_offset=2&timeResolution=hourly&temperatureunit=CELSIUS&velocityunit=KILOMETER_PER_HOUR&energyunit=watts&lengthunit=metric&degree_day_type=10%3B30&gddBase=10&gddLimit=30. [Accessed Dec. 27, 2021].

[20] Japan Meteorological Agency, "History weather data in Osaka," 2021 [Online]. Available: https://www.data.jma.go.jp/obd/stats/etrn/index.php?prec_no=62&block_no=47772&year=2014&month=01&day=01&view=p1 [Accessed Dec. 27, 2021].

[21] L. Breiman, "Random Forest", *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.