Original Article

DOI: 10.5582/bst.2018.01126

Forecasting hand, foot, and mouth disease in Shenzhen based on daily level clinical data and multiple environmental factors

Ren Zhong[1,§],Yongsheng Wu[2,§], Yunpeng Cai[1], Ruxin Wang[1], Jing Zheng[3], Denan Lin[3], Hongyan Wu[1,*] ,Ye Li[1,*]

[1] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China;*
[2] *Shenzhen Center for Disease Control and Prevention, Shenzhen, China;*
[3] *Shenzhen Health Information Center, Shenzhen, China.*

Summary **Hand, foot, and mouth disease (HFMD) is caused by a group of enteroviruses. It infects millions of children in the Southeast Asian area. An accurate forecasting of outbreaks of HFMD could facilitate public health officials to suggest public health actions earlier. Many researchers tried to develop an early warning system for HFMD to lower the damage caused by a HFMD outbreak. The research data based on daily level could help figure out the relationship between HFMD and environmental factors, but nevertheless is difficult to collect. In this study, we collected the daily clinical data from the Shenzhen Health Information Center and multiple environmental factors to analyze the outbreaks of HFMD. Considering the incubation period of HFMD, we fed the previous 60 days' HFMD rates, 7 days' temperature factors and 7 days' air-quality factors into the tree model, XGBoost. The following conclusions were drawn in this study: *i*) Compared with the model only using the previous HFMD rate and temperature factors, the addition of the air-quality factors could make the model better, improving MAE nearly 16.7%. *ii*) By analyzing the Pearson correlation, we found that the temperature showed a positive correlation and the air quality showed a negative correlation for the HFMD outbreaks. Improving the air quality, especially decreasing $PM_{2.5}$ and PM_{10} could decrease the risk of HFMD outbreaks.**

Keywords: Hand, foot, and mouth disease (HFMD), tree model, XGBoost, Correlation

1. Introduction

Hand, foot, and mouth disease (HFMD) is caused by a group of enteroviruses, of which the coxsackievirus type A and the enterovirus 71 (EV71) are the most frequently seen. This illness mainly occurs in children aged less than 5 years. Although a majority of patients have only mild symptoms, some patients rapidly develop neurological and cardiopulmonary symptoms that can be fatal, particularly when the cases are associated with EV71. Since 1997, numerous large outbreaks of HFMD have occurred in Eastern and Southeastern Asian countries, including Singapore, Malaysia, Japan, and China (*1-5*). In China, the Chinese Ministry of Health has listed HFMD as a class C communicable disease and all doctors are required to report HFMD cases to the National Disease Surveillance Reporting and Management System.

Many researchers used data collected from the surveillance system to develop an early warning system of HFMD to lower the damage caused by a HFMD outbreak. A study by Nanjing Medical University attempted to predict the HFMD epidemics in Nanjing city, the main epidemic area of eastern China, by developing a weather-based forecasting model (*6*). A research team from Beijing explored the data collected in Beijing to find seasonal and other potential effects of weather factors on HFMD (*7*). In the southern part of China, some researchers from Guangzhou Center for Disease Control and Prevention also tried to estimate the effects of diverse climate variables on the incidence of HFMD (*8*). Another team from

[§] These authors contributed equally to this work.

**Address correspondence to:*
Drs. Hongyan Wu and Ye Li, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences,1068 Xueyuan Avenue, Shenzhen University Town, Nanshan, Shenzhen 518055, China.
E-mail: hy.wu@siat.ac.cn (Wu Y), ye.li@siat.ac.cn (Li Y)

Chongqing adopted multiple meteorological parameters to quantify the association of daily weather variation with HFMD incidence (*9*). Besides the meteorological parameters, studies from the Central South University also explored the relationship between HFMD and fine particulate matter less than 10 (PM_{10}) (*10*). This research has achieved state-of-the-art models which deliver sufficiently accurate forecasts of HFMD. However, to the best of our knowledge, these studies on building HFMD monitoring and forecasting largely focus on weekly forecasts due to the lack of daily data.

Research data based on daily levels could help figure out the relationship between HFMD and environmental factors, but nevertheless is difficult to collect. In this study, we collected the daily clinical data from the Shenzhen Health Information Center and multiple environmental factors to analyze the outbreaks of HFMD. We consider three kinds of predictors: temperature, historcal HFMD rate, and air quality. Considering the incubation period of HFMD, we fed the previous 60 days' HFMD rates, 7 days' temperature factors and 7 days' air-quality factors into the tree model, XGBoost. XGBoost is a highly sophisticated algorithm, powerful enough to deal with kinds of data irregularities. In addition, its feature importance functionality could help us understand the relationship between HFMD rate, temperature, and air quality.

2. Materials and Methods

2.1. *Data sources*

The weather and air-quality data were obtained from the Weather Underground app (*https://www.wunderground.com/*). The HFMD rate data were obtained from the Shenzhen Health Information Center, which collected the clinic visit information from January 1, 2010, to September 12, 2017, from 60 state hospitals, 6 mother and child care centers, and 619 community rehabilitation centers. Figure 1 illustrates the data, in which the Y-axis represents the daily HFMD rate and the X-axis represents the outbreak time.

2.2. *Methodology*

XGBoost, A scalable machine learning system for tree boosting, was first popularized by Tianqi Chen for improving the operating efficiency and reducing the memory space usage of the current tree boosting system. XGBoost is a kind of assembly algorithm, named gradient-boosted decision trees (GBDT), which creates and combines a high number of individually weak but complementary classifiers, to produce a robust estimator. In XGBoost, a new weak classifier is constructed to be maximally correlated with the negative gradient of the loss function associated with the whole assembly for each iteration. XGBoost belongs to the group of widely used tree learning algorithms. A decision tree makes predictions on an output variable based on a series of rules arranged in a tree-like structure. They consist of a series of split points, the nodes, in terms of the value of an input feature and gives us the specific value of the output variable at the leaf node. As an efficient implementation of GBDT, XGBoost remarkably improves the efficiency of the training model but slightly decreases the accuracy of the model (*11*). In this study, three XGBoost models were built by gradually adding predictor factors into the predictor space to obtain the best model performance with minimum regression error.

Predictor space. In this study, three kinds of components were chosen as the predictor space: historical daily HFMD rate, temperature conditions, and air-quality conditions. Assuming the current predicted point was X0, the first component was the sequence X_1, X_2, X_3, …, X_{t-1}, X_t, where t was 60 and was filled with the values of the previous 60 HFMD rate observations before X0. The second components were maximum and minimum temperatures, which were the optimal-related factors with HFMD in previous studies (*12*). The third component was composed of the air-quality index (AQI), fine particulate matter less than 2.5 ($PM_{2.5}$), PM10, sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), and ozone (O_3).

Metrics. The mean absolute error (MAE) was used to measure the prediction accuracy. MAE was defined using the following formula:

$$\mathrm{MAE}\ (A_t, F_t) = \frac{1}{N}\sum_{i=0}^{N-1} |A_t - F_t|$$

Where A_t is the actual value and F_t is the forecast value.

Variable importance. Variable importance in XGBoost or GBDT is often called relative importance of predictor variables, which is useful to learn the relative importance or contribution of each input variable in predicting the response (*13*). In this study, XGBoost was tree-based, and therefore the variable importance was computed as follows:

(1) For a single decision tree *T*, the following formula was used:

$$I_l^2(T) = \sum_{t=1}^{J-1} \hat{\imath}_t^2\, i(t = l)$$

Calculating the relevance for each predictor variable F_t: The sum was over the $J - 1$ internal nodes of the tree. At each such node *t*, one of the input variables F_t was used to partition the region associated with that node into two subregions; within each, a separate constant was fit to the response values. The particular variable chosen was the one that gave the maximal estimated improvement $\hat{\imath}_t^2$ in squared error risk over that for a

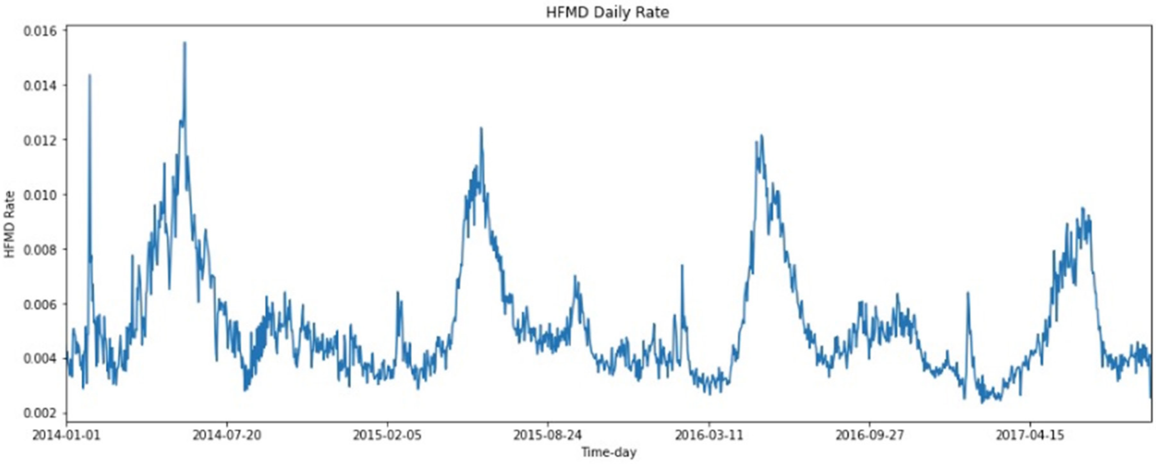

Figure 1. Data from the Shenzhen Health Information Center. The Y-axis represents the daily HFMD rate and the X-axis represents time.

constant fit over the entire region. The squared relative importance of variable F_l was the sum of such squared improvements over all internal nodes for which it was chosen as the splitting variable. i was the characteristic function whose value was 1 if $t = l$ and 0 if $t \neq l$.

(2) As XGBoost used trees to predict, the variable importance was easily generalized to additive tree expansions. It was simply averaged over the trees as shown in the following formula:

$$I_l^2(T) = \frac{1}{M}\sum_{m=1}^{M} I_l^2(T_m)$$

3. Results

The data from January 1, 2014, to December 31, 2016, were used as the training data, and the half-year data from January 1, 2017, to September 1, 2017, were used as the test data. The experiments were performed by the training regression model using the training data and calculating the regression error between forecasted value and test data. The experiments were carried out three times by gradually combining more predictors into the predictor space to investigate the influence of different predictors on the prediction accuracy.

3.1. *Improved prediction accuracy*

In the first process, the 14 recent temperature conditions (the maximum and minimum temperatures of the previous week) were chosen. In the second, 60 recent observation variables X_1, X_2, X_3, …, X_{59}, X_{60} were added. In the third, the air-quality factors and the foregoing predictor space were combined together as the last predictor space.

Figure 2A illustrates the result of the first experiment, which shows the model's ability to forecast the future HFMD trend, but the prediction accuracy was not good. Figure 2B illustrates the second experiment of the daily HFMD rate with the predictor space of temperature and history HFMD rate, showing that the prediction accuracy was improved. Figure 2C shows that the prediction got better by adding air quality. Table 1 shows the improved forecast results by adding more predictor factors into the predictor space.

3.2. *Comparison of variable importance*

As the final model provided the best result and the predictor space included all three kinds of factors, only the variables in the final model were checked. This study used the get*_fscore* function in the XGBoost package to calculate the variable's importance. The larger the value returned by the function, the greater the influence of a feature in the modeling process. According to their variable importance, the top 10 variables were obtained and are shown in Table 2. The following observations were made: (1) The historical HFMD rate, especially the nearest days' rate, seemed pretty important. (2) The air-quality factors showed more importance than the temperature factor. Two days ago, NO_2 and AQI occupied the fourth and ninth places, respectively.

3.3. *Analysis of weather conditions*

The three experiments revealed that the addition of air-quality conditions into the predictor space could efficiently improve the prediction accuracy. The analysis of variable importance also showed that the air-quality factors had an influence on the prediction. In this section, the Pearson correlation analysis between the HFMD rate and both temperature and air-quality factors was performed. Table 3 shows the correlation coefficients between the HFMD rate and these factors. The coefficient was a value that ranged from 1 to –1. The closer to 1 the coefficient was, the more linearly

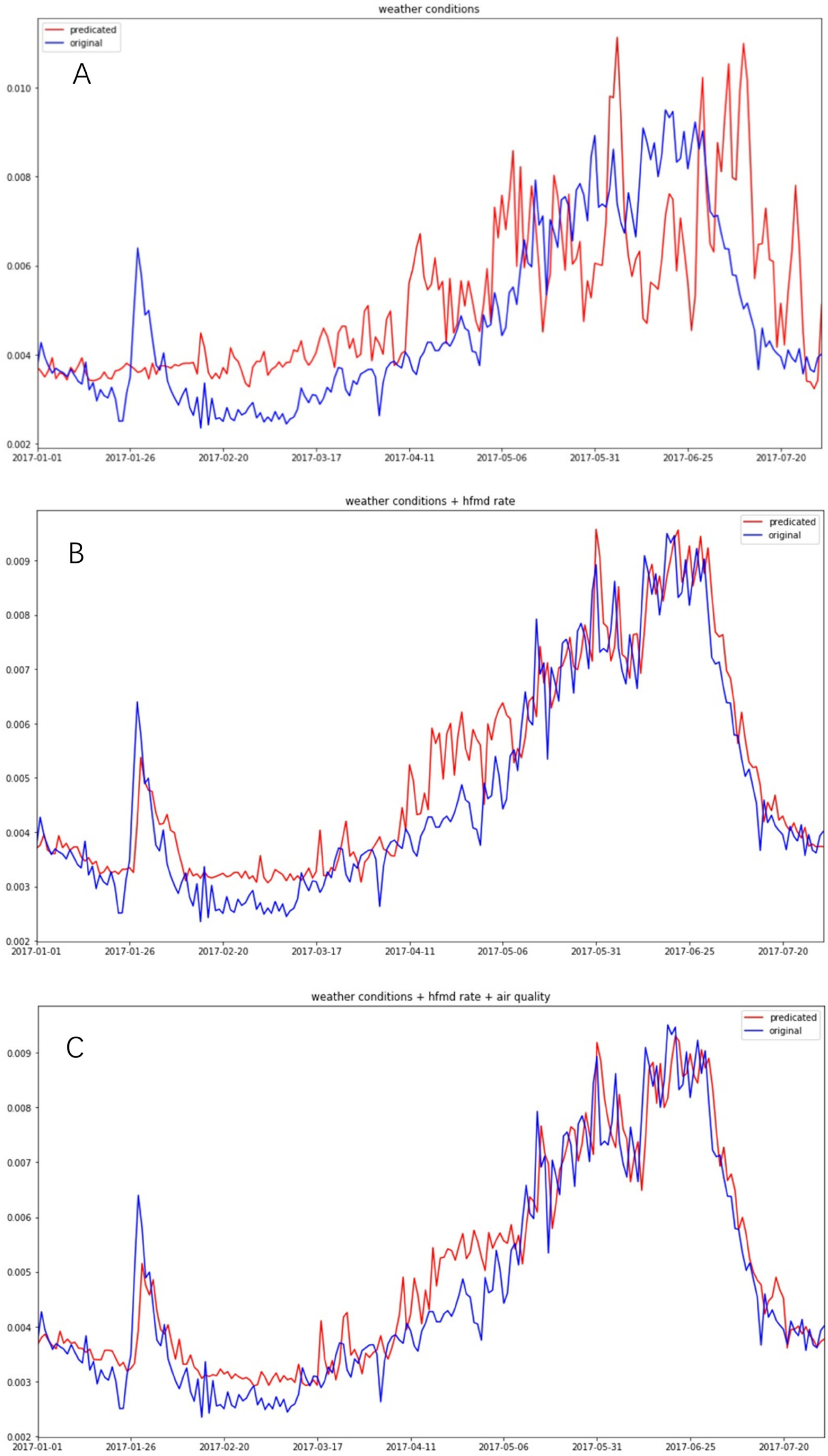

Figure 2. Predication of Weekly ILI rate with different predicator spaces. (A) Prediction of the daily HFMD rate using temperature factors. The blue line illustrates the original data, and the red line shows the corresponding predicted values. **(B)** Prediction of the daily HFMD rate using temperature factors and historical HFMD rate. The blue line illustrates the original data, and the red line shows the corresponding predicted values. **(C)** Prediction of the daily HFMD rate using temperature factors, historical HFMD rate, and air-quality factors. The blue line illustrates the original data, and the red line shows the corresponding predicted values.

Table 1. Comparison of forecasting with different predictors

Predictors space	MAE
Temperature	0.0013
Temperature + HFMD rate	0.0006
Temperature + HFMD rate + air quality	0.0005

HFMD, Hand, foot, and mouth disease; MAE, mean absolute error.

Table 2. Comparison of variable importance

Variables	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Name	$RATE_1$	$RATE_2$	$RATE_7$	$NO2_2$	$RATE_4$	$RATE_{59}$	$RATE_3$	$RATE_{48}$	AQI_2	$RATE_{57}$
VI	42	32	14	14	14	13	11	11	11	11

VI, Variable importance. Capital letters in name are the names of the features, and the suffix indicates how many days ago the value of the feature was obtained.

Table 3. Analysis of correlation between the HFMD rate and other factors

Variables	AQI	$PM_{2.5}$	PM_{10}	SO_2	CO	NO_2	O_3	MAXT	MINT
Coefficient	-0.2968	-0.3653	-0.3249	-0.1324	-0.1808	-0.2564	-0.2063	0.4414	0.4743
P value	5.31e-28	1.45e-42	1.57e-33	1.55e-06	4.53e-11	4.47e-21	4.91e-14	1.8e-63	2.39e-74

AQI, Air-quality index; CO, carbon monoxide; NO_2, nitrogen dioxide; O_3, ozone; $PM_{2.5}$, fine particulate matter less than 2.5; PM_{10}, fine particulate matter less than 10; SO_2, sulfur dioxide.

positive the feature was related to the HFMD rate; and the closer to –1 the coefficient was, the more linearly negative the feature was related to the HFMD rate. The closer the coefficient was to 0, the weaker was the linear correlation between the feature and the HFMD rate.

4. Discussion

In this study, the daily maximum and minimum temperatures in the first experiment were chosen to examine the relationship between the temperature and HFMD. Figure 2A showed that the temperature could reflect the future HFMD trend, although the prediction accuracy was not good. The Pearson correlation in Table 3 also showed that the daily maximum and minimum temperature was correlated with HFMD, with the coefficients of 0.44 and 0.47, separately. Our conclusion is consistent with the previous studies (*6-10*), which shows outbreaks of HFMD in different areas.

Although some researchers found no strong correlation between air quality and HFMD, we still chose air quality factors as a part of feature space, considering that the air quality could impact on virus diffusion, people's travel and respiratory diseases etc. We added the air-quality index (AQI), fine particulate matter less than 2.5 ($PM_{2.5}$), PM_{10}, sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), and ozone (O3) to the previous model, which improved MAE from 0.0006 to 0.0005, nearly 16.7%. The variable importance in this study shows the importance of air quality. The top 10 variables of the Shenzhen data were $RATE_1$, $RATE_2$, $RATE_7$, NO_2, $RATE_{59}$, $RATE_3$, $RATE_{48}$, AQI_2, and $RATE_{57}$. Furthermore, the Pearson correlation, some air-quality factors showed a negative correlation. The coefficient between $PM_{2.5}$ and HFMD is -0.37 while PM10 is -0.33. Our study is not consistent with the conclusion in the previous study (*10*), which showed no significant relationship between PM10 and HFMD (*10*).

The variable importance in our study showed that air quality seemed more important than temperature factors. We thought that the temperature factors could be collinear with historical rates, and therefore the importance of temperatures could be covered by the historical rates.

5. Conclusions

The clinic and environmental data based on daily levels could help figure out the relationship between HFMD and environmental factors. Considering the incubation period of HFMD, in this study we fed the previous 60 days' HFMD rates, 7 days' temperature factors and 7 days' air-quality factors into the tree model, XGBoost. Our study showed that the addition of air-quality factors to the historical HFMD rate and temperature data, could improve forecasting of HFMD. In addition, air quality showed a negative correlation to HFMD outbreaks. Improving air quality, especially decreasing $PM_{2.5}$ and PM_{10} , could decrease the risk of HFMD outbreaks.

Acknowledgements

This study was supported in part by the National High-tech R&D Program (863 Program) of China

(SS2015AA020109), Shenzhen Science and Technology Research Foundation (JCYJ20160229193120432), Major Special Project of Guangdong Province (2017B030308770), the National Natural Science Foundation of China (81601575), Layout of Shenzhen Basic Research Foundation, and the foundation from Shenzhen Healthy Big Data Analytics and Application Engineering Laboratory.

References

1. Ang LW, Koh BK, Chan KP, Chua LT, James L, James L, Goh KT. Epidemiology and control of hand, foot and mouth disease in Singapore, 2001–2007. Ann Acad Med Singapore. 2009; 38:106-112.
2. Chan LG, Parashar UD, Lye MS, Ong FG, Zaki SR, Alexander JP, Ho KK, Han LL, Pallansch MA, Suleiman AB, Jegathesan M, Anderson LJ. Deaths of children during an outbreak of hand, foot, and mouth disease in Sarawak, Malaysia: Clinical and pathological characteristics of the disease. For the Outbreak Study Group. Clin Infect Dis. 2000; 31:678-683.
3. Fujimoto T, Chikahira M, Yoshida S, Ebira H, Hasegawa A, Totsuka A, Nishio O. Outbreak of central nervous system disease associated with hand, foot, and mouth disease in Japan during the summer of 2000: Detection and molecular epidemiology of enterovirus71. Microbiol Immunol. 2002; 46:621-627.
4. Chen KT, Chang HL, Wang ST, Cheng YT, Yang JY. Epidemiologic features of hand-foot-mouth disease and herpangina caused by enterovirus71 in Taiwan, 1998–2005. Pediatrics. 2007; 120:e244-252.
5. Yang F, Ren L, Xiong Z, Li J, Xiao Y, Zhao R, He Y, Bu G, Zhou S, Wang J, Qi J. Enterovirus 71 outbreak in the People's Republic of China in 2008. J Clin Microbiol. 2009; 47:2351-2352.
6. Liu S, Chen J, Wang J, Wu Z, Wu W, Xu Z, Hu W, Xu F, Tong S, Shen H. Predicting the outbreak of hand, foot, and mouth disease in Nanjing, China: A time-series model based on weather variability. Int J Biometeorol. 2018; 62:565-574.
7. Dong W, Li X, Yang P, Liao H, Wang X, Wang Q. The effects of weather factors on hand, foot and mouth disease in Beijing. Sci Rep. 2016; 6:19247.
8. Li T, Yang Z, DI B, Wang M. Hand-foot-and-mouth disease and weather factors in Guangzhou, southern China. Epidemiol Infect. 2014; 142:1741-1750.
9. Wang P, Zhao H, You F, Zhou H, Goggins WB. Seasonal modeling of hand, foot, and mouth disease as a function of meteorological variations in Chongqing, China. Int J Biometeorol. 2017; 61:1411-1419.
10. Ruixue Huang, Guolin Bian, Tianfeng He, Lv Chen, Guozhang Xu. Effects of Meteorological Parameters and PM10 on the Incidence of Hand, Foot, and Mouth Disease in Children in China. Int J Environ Res Public Health. 2016; 13:481.
11. Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; 785-794.
12. Xiao X, Gasparrini A, Huang J, Liao Q, Liu F, Yin F, Yu H, Li X. The exposure-response relationship between temperature and childhood hand, foot and mouth disease: A multicity study from mainland China. Environ Int. 2017; 100:102-109.
13. Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning. Springer, Germany, 2009; pp. 367-369.

(Received June 8, 2018; Revised October 24, 2018; Accepted October 28, 2018)