TDDF: HFMD Outpatients Prediction Based on Time Series Decomposition and Heterogenous Data Fusion in Xiamen, China

Zhijin Wang[1(✉)], Yaohui Huang[2], Bingyan He[1], Ting Luo[2], Yongming Wang[3], and Yingxian Lin[1(✉)]

[1] Computer Engineering College, Jimei University, Yinjiang Road 185, Xiamen 361021, China
{zhijin,hebingyan,yxlin}@jmu.edu.cn
[2] Chengyi University College, Jimei University, Jimei Road 199, Xiamen 361021, China
{yhhuang,201641051007}@jmu.edu.cn
[3] China Electronics Technology Group Corporation, Shandong Middle Road 337, Shanghai 200001, China
ymwang819@gmail.com

Abstract. Hand, foot and mouth disease (HFMD) is a common infectious disease in global public health. In this paper, the time series decomposition and heterogeneous data fusion (TDDF) method is proposed to enhance features in the performance of HFMD outpatients prediction. The TDDF first represents meteorological features and Baidu search index features with the consideration of lags, then those features are fused into decomposed historical HFMD cases to predict coming outpatient cases. Experimental results and analyses on the real collected records show the efficiency and effectiveness of TDDF on regression methods.

Keywords: HFMD prediction · Meteorological factor · Baidu search index

1 Introduction

Hand, foot and mouth disease (HFMD) is a common global infectious disease [4]. This disease is easy to cause fever, oral ulcers, blisters and rashes on hands, feet and buttocks, some serious and potentially fatal complications will lead to serious sequelae and even death [11]. Millions of people, in particular for children less than 5 years old, suffer from HFMD-related disease over the past decade [7,8,16]. The control and prevention of HFMD is a public health issue that receives attentions by government agencies, medical institutions and the public [14]. It became the 38th legally notifiable disease in the China's National Notifiable Disease Reporting and Surveillance System [1,3] on May 2, 2008 [2].

J. Li et al. (Eds.): ADMA 2019, LNAI 11888, pp. 658–667, 2019.
https://doi.org/10.1007/978-3-030-35231-8_48

Auxiliary data are collected to alleviate the uncertainty of disease occurrences and improve the prediction performance. Typically, meteorological factors have been proved to be associated with the incidence of infectious diseases [9,15]. For example, temperature and relative humidity are presented as continuous variables, and connected with HFMD cases using by using linear models. But other weak correlated climate factors are usually ignored, such as dew point and atmospheric pressure. Recently, as an important entry of the Internet, search engines are adopted to track infectious disease epidemics in countries and provinces [6]. These methods commonly use "手足口病" as query keyword, and search engine returns the search indices.

This motivate us to leverage both meteorological factors and search indices to provide predictions for short-term HFMD outpatient visits with respect to feature construction and representation. The challenges of constructing features are: (1) *Weak correlated feature discovery.* Many factors affect infectious disease propagation, and weak correlated factors are usually ignored. (2) *Time concerned feature representation.* There exists a time interval between factors and disease outbreaks. The time interval consists of incubation period and diagnosing duration. Usually, the incubation period is 2–5 days, and it takes another 1–2 days to be diagnosed. In total, it spans 2–7 days from infection time to report time. Hence, the time interval should be taking into consideration for feature representation.

To address the above challenges, the time series decomposition and heterogeneous data fusion (TDDF) method is proposed to enhance the features for a better prediction performance. The TDDF first represent meteorological factor features and Baidu search index features with the consideration of lags, then those features are fused into decomposed historical HFMD cases to predict coming outpatient cases. In the first stage, the related data are collected, organized and count variables in weeks. The correlation between variables and outpatients was analyzed based on bivariate correlation analysis. We try time difference analysis on two observing variables to figure out the connections between current outpatients and the outpatients of several days before. In the second stage, we consolidate a feature matrix based on previous analyses and figures. Thus, outpatients predictions are provided via well trained regression models.

2 Study Area and Auxiliary Data

2.1 Study Area

Xiamen is located in the southeast part of China, and is also an important special economic zone in China. It covers a land area of 1,699.39 $km^2$and a sea area of over 390 km^2 until 2017. Xiamen has a monsoonal humid subtropical climate, characterised by long, hot and humid summers (but moderate compared to much of the rest of the province) and short, mild and dry winters, the annual mean is 20.7 °C (69.3 °F) [5]. There are 4.01 million permanent residents as of 2018. As the sample of Xiamen's total population is relatively stable between 2012 and 2016, with an annual growth rate between 1.3% and 2.1%, the trend of

morbidity during this period can be stipulated by the trend of the number of disease cases. Therefore, the number of HFMD outpatient visits is helpful for monitoring disease status of a city within a period.

2.2 Data Source

HFMD Outpatient Visits. The temporal variation of weekly outpatient cases from January 1, 2012 to December 31, 2016 is collected. It contains 261 weeks data pairs with mean value 113, and ranges from 2 to 435 cases. The cases were all clinical or laboratory-confirmed cases of HFMD and reported by hospital diagnostic.

Meteorological Factor (MF). MF information is provide in days, such as daily average temperature (°C), and daily average dew point (°C). Therefore, we calculate 261 weekly MF pairs by the downloaded daily records, and their correlations with the variable of HFMD cases can be figured out based on these data pairs.

Baidu Search Index (BSI). The key concern is to find proper query words, to make sure that provided indices are connected to the number of HFMD cases. The commonly used key word "手足口病" is choosed in this work. BSI engine returns daily counts of a given keyword under a given region (e.g., city, province or countrywide) and a given platform (e.g., mobile, PC, or total). Therefore, weekly data pairs are calculated by the returned 6 groups of daily indices.

2.3 Time Difference Analysis

We focus on discovering the autocorrelation of C_{hfmd} and correlations between observation variables (i.e., MF and BSI) and C_{hfmd}. The Pearson Correlation Coefficient (PCC) is adopted to measure degrees of relevance among these continuous variables. All statistical analyses are two-sided and p-value < 0.05 is considered statistically correlated.

Results of statistical analysis are carried out and listed at Table 1. The first 17 rows (from 1st row to the 2nd last row) give time difference correlations between auxiliary variables and HFMD outpatients variable under d lagged weeks. The last row gives the autocorrelation of HFMD outpatients variable under d lagged weeks. These analyses reveal that the lagging period is short, and commonly less than 1 weeks. Technically, weather conditions change 1 week before the disease happen, and people use search engines when disease is occurring or close to it. Possible reasons are: HFMD is quickly onset and its incubation period is short usually 3–5 days. These analyses give suggests of time difference settings to data decomposition and data fusion as well.

Table 1. The results of time difference relevance between variables in auxiliary data and the variable of HFMD cases. The most significant correlated value of each row is in bold type. Symbol "d" denotes the number of weeks of time difference. *: $0.01 < p\text{-value} < 0.05$, correlation. **: $p\text{-value} < 0.01$, significant correlation.

Symbols	$d=0$	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
T_{max}	**0.4563****	0.4506**	0.4252**	0.3819**	0.3414**	0.2969**	0.234**
T_{min}	**0.4566****	0.4389**	0.3879**	0.3308**	0.2796**	0.2204**	0.1551*
T_{avg}	**0.4694****	0.4559**	0.4060**	0.3512**	0.3051**	0.2508**	0.1853**
D_{max}	0.4937**	**0.5181****	0.4847**	0.4414**	0.4000**	0.3498**	0.2955**
D_{min}	0.4531**	**0.4564****	0.3963**	0.3390**	0.2927**	0.2247**	0.1895**
D_{avg}	0.5064**	**0.5222****	0.4708**	0.4144**	0.3688**	0.3067**	0.2555**
H_{min}	0.3384**	**0.3880****	0.3142**	0.2641**	0.2279**	0.1434*	0.1492*
H_{avg}	0.2968**	**0.3942****	0.3715**	0.3485**	0.3371**	0.3022**	0.3233**
A_{max}	−0.4078**	**−0.4126****	−0.3905**	−0.3494**	−0.3131**	−0.2670**	−0.2190**
A_{min}	−0.2927**	**−0.3201****	−0.2964**	−0.2724**	−0.2825**	−0.2719**	−0.2816**
A_{avg}	−0.3862**	**−0.3969****	−0.3740**	−0.3331**	−0.3097**	−0.2728**	−0.2386**
B_{pc}	**0.5015****	0.4654**	0.4116**	0.3370**	0.2620**	0.1756**	0.0836
B_{mo}	**0.7524****	0.7095**	0.6406**	0.5537**	0.4578**	0.3653**	0.2686**
B_{total}	**0.7551****	0.7105**	0.6396**	0.5488**	0.4501**	0.3519**	0.2489**
B^1_{pc}	**0.6326****	0.5818**	0.5055**	0.4287**	0.3380**	0.2306**	0.1218
B^1_{mo}	**0.7080****	0.6607**	0.5831**	0.5057**	0.4252**	0.3451**	0.2638**
B^1_{total}	**0.7935****	0.7401**	0.6531**	0.5633**	0.4674**	0.3712**	0.2730**
C_{hfmd}	–	**0.9120****	0.7807**	0.6516**	0.5242**	0.4006**	0.2785**

The calculations of these relevant values are formulate as follows. Let $\Omega_m = \{T_{max}, T_{min}, T_{avg}, D_{max}, D_{min}, D_{avg}, H_{min}, H_{avg}, A_{max}, A_{min}, A_{avg}\}$ denote variables of MF, $\Omega_b = \{B_{pc}, B_{mo}, B_{total}, B^1_{pc}, B^1_{mo}, B^1_{total}\}$ denote variables of BSI, and $\Omega = \{\Omega_m, \Omega_b\}$ denote all auxiliary variables. We use $\Omega(1:t)$ to present t data pairs in Ω, N to denote the number of all data pairs. The correlations are formulated as:

$$\begin{aligned} Corr(\Omega, C_{hfmd}, d) &= PCC(\Omega(1:N-d), C_{hfmd}(1+d:N)), s.t., d &>= 0; \\ Corr(C_{hfmd}, C_{hfmd}, d) &= PCC(C_{hfmd}(1:N-d), C_{hfmd}(1+d:N)), s.t., d &>= 1. \end{aligned} \tag{1}$$

3 TDDF and the Development for Prediction

3.1 Time Series Decomposition and Heterogenous Data Fusion (TDDF)

TDDF consists of time difference analysis, time series decomposition, and feature consolidation. For easy presentation, symbol $C_t = \{c_t\}_{t=1}^N \in \mathbb{R}^{N\times 1}$ is used to denote the time series of HFMD cases, symbol $\mathbf{M}_t = \{\mathbf{m}_t\}_{t=1}^N \in \mathbb{R}^{N\times 11}$ is adopted to describe 11 time series of MF, and symbol $\mathbf{B}_t = \{\mathbf{b}_t\}_{t=1}^N \in \mathbb{R}^{N\times 6}$ is employed to describe 6 time series of BSI.

Time Difference Analysis. The degree of time difference relevance is calculated according to Eq. 1, which measures the relevance between variable(s) with d weeks' lagging. The statistical analyses of variables are listed at Table 1.

Time Series Decomposition. Historical infectious disease outbreaks will affect the current status of these diseases. Based on this assumption, the next weeks' cases is formulate as:

$$c_{t+1} \leftarrow (c_{t-d_1}, c_{t-d_1+1}, \cdots, c_t) \quad s.t.,\ d_1 \geq 1. \tag{2}$$

We use symbol $\mathbf{c}_{t-d_1} = \{c_i\}_{i=t-d_1}^{t} \in \mathbb{R}^{d_1 \times 1}$ to present historical HFMD cases of previous d_1 weeks. Thus, Eq. 2 is presented as:

$$c_{t+1} \leftarrow \mathbf{c}_{t-d_1} \quad s.t.,\ d_1 \geq 1. \tag{3}$$

There are many signal processing methods, which decompose a time-series to extract (or represent) features. Such as, empirical mode decomposition (EMD) [13], wavelets [10], spline methods [6], and ARIMA [12]. In this study, in order to investigate advantages of the auxiliary data in improving prediction performance, we use linear transformation to process $c_t, t \in [1, N]$ and use the mean value of C_t to fill missing values.

Feature Consolidation. Moreover, let $\mathbf{m}_{t-d_2} \in \mathbb{R}^{11 \times 1}$ stand for climate factors of the $t - d_2$ period, and let $\mathbf{b}_{t-d_3} \in \mathbb{R}^{6 \times 1}$ stand for search indices of the $t - d_3$ period. The auxiliary factors are fused and formulated as:

$$c_{t+1} \leftarrow (\mathbf{c}_{t-d_1}, \mathbf{m}_{t-d_2}) \quad s.t.,\ d_1 \geq 1\ and\ d_2 \geq 0, \tag{4}$$

$$c_{t+1} \leftarrow (\mathbf{c}_{t-d_1}, \mathbf{b}_{t-d_3}) \quad s.t.,\ d_1 \geq 1\ and\ d_3 \geq 0, \tag{5}$$

$$c_{t+1} \leftarrow (\mathbf{c}_{t-d_1}, \mathbf{m}_{t-d_2}, \mathbf{b}_{t-d_3}) \quad s.t.,\ d_1 \geq 1,\ and\ d_2,\ d_3 \geq 0. \tag{6}$$

Symbol $f_1, \cdots, f_4$ stand for the bridge between features and their target. Thus, we get:

$$c_{t+1} \leftarrow f_4(f_1(\mathbf{c}_{t-d_1}), f_2(\mathbf{m}_{t-d_2}), f_3(\mathbf{b}_{t-d_3})) \quad s.t.,\ d_1 \geq 1\ and\ d_2,\ d_3 \geq 0. \tag{7}$$

According to Eq. 7, abundant of models can be developed for training and prediction. The models include but not limited to neural networks, Adaboost. For better investigation of data-driven improvements and significant correlations of those variables (see Table 1), we adopt linear transformation to carry out $f_1, \cdots, f_4$.

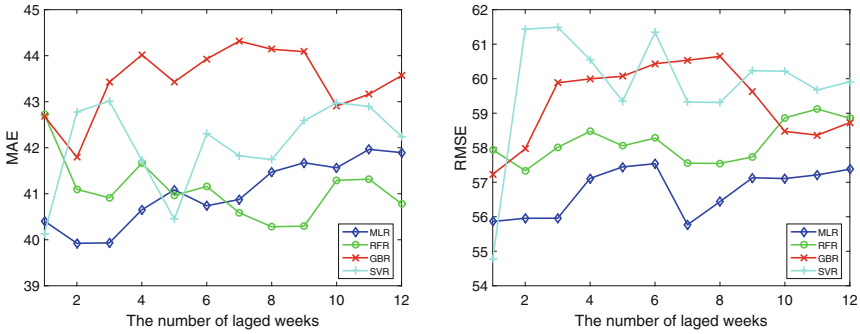
3.2 Performance Evaluation Criteria (PEC)

To date, a variety of performance evaluation criteria (PEC) have been proposed for evaluation and intercomparision of different models, but no single evaluation index is recognized as a universal standard. Therefore, we need to evaluate prediction performance based on multiple PEC and analyze the prediction accuracy performance of different prediction models under multiple metrics.

The disease dataset was divided into two subsets: the first part, from the 1st week of 2012 to the 52nd week of 2015, was used for model training and construction, and the subsequent part, from the 1st to the 52nd week of 2016, for external validity assessment.

4 Experimental Results

4.1 Predictions on the Basis of Historical Cases



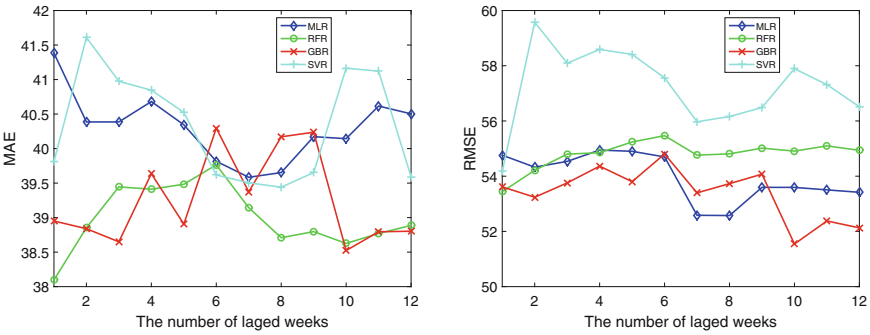
(a) MAE (Cases only). (b) RMSE (Cases only).

Fig. 1. The performance on historical cases in terms of MAE and RMSE.

Results of predictions on the basis of historical cases (see Eq. 3) in terms of MAE and RMSE are shown in Fig. 1. To observe time-series decomposition parameter d_1 in affecting prediction, we change d_1 from 1 to 12 to measure the performance of regression algorithms. As the results shows, the predicting performance of the 4 algorithms are very unstable, which suggests that predictions on the basis of historical cases are not robust, and have potentials to be improved. When $d_1 = 1$ and $d_1 = 2$, the optimal values of MAE and RMSE are found.

4.2 Predictions Based on Historical Cases and MF

Results of predictions based on historical cases and MF (see Eq. 4) in terms of MAE and RMSE are shown in Fig. 2. To observe time-series decomposition parameter d_1 in affecting prediction, we change d_1 from 1 to 12 while holding

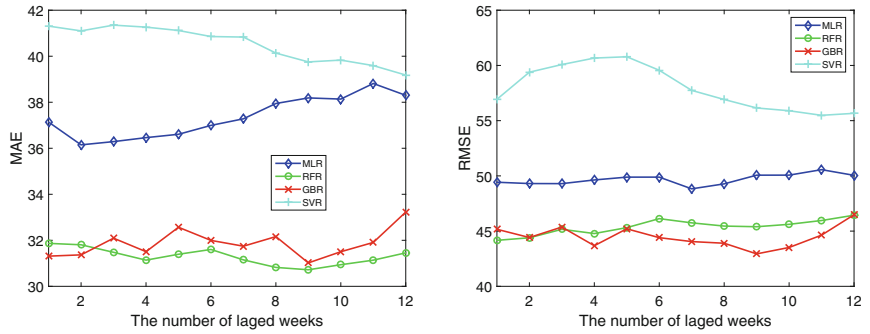


(a) MAE (Cases + MF). (b) RMSE (Cases + MF).

Fig. 2. The performance on historical cases and MF in terms of MAE and RMSE.

$d_2 = 1$, to measure the performance of regression algorithms. As the results shows, the predicting performance of the 4 algorithms gradually become stable with respect to RMSE in Fig. 2(b), which means the training is well convergent. But the MAE is very unstable in Fig. 2(a). When $d_1 = 2$, 3, and 10, the optimal values of MAE and RMSE are found.

4.3 Predictions Based on Historical Cases and BSI



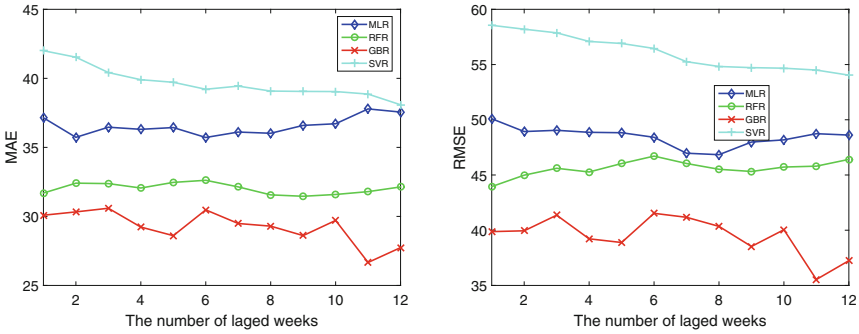
(a) MAE (Cases + BSI). (b) RMSE (Cases + BSI).

Fig. 3. The performance on historical cases and BSI in terms of MAE and RMSE.

Results of predictions based on historical cases and BSI (see Eq. 5) in terms of MAE and RMSE are shown in Fig. 3. To observe time-series decomposition parameter d_1 in affecting prediction, we change d_1 from 1 to 12, while holding $d_3 = 0$, to measure the performance of regression algorithms. As displayed in Fig. 3(a) and (b), predictions performance become stable, GBR and RFR run

better results than other two regressors. When $d_1 = 2$, 3 and 4, the optimal values of MAE and RMSE are found. It should be noted that, SVR's performance becomes worse when compared performance on previous 2 data groups. A possible reason is that few samples can not train a SVR well, either over-fitting or under-fitting.

4.4 Predictions Based on Historical Cases, BSI, and MF

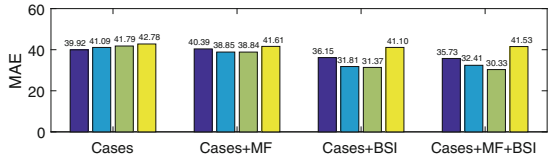


(a) MAE (Cases + MF + BSI). (b) RMSE (Cases + MF + BSI).

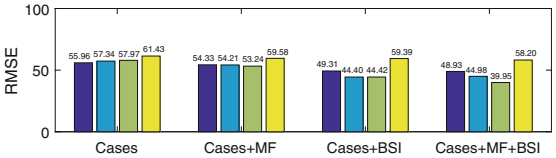
Fig. 4. The performance on historical cases, MF and BSI in terms of MAE and RMSE. (Color figure online)

Results of predictions based on historical cases, MF and BSI (see Eq. 6) in terms of MAE and RMSE are shown in Fig. 4. We change d_1 from 1 to 12, while holding $d_2 = 1$ and $d_3 = 0$, to measure the performance of regression algorithms. As illustrated in Fig. 4(a) and (b), the performance is very stable for all regressors. There are obviously different on regressors: GBR has best predictions (the red line with cross sign), while SVR (the aqua line with plus sign) is worst.

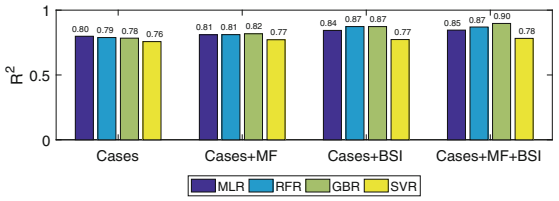
Given that the number of lagged weeks is set to 2, we compare the performance of regressors over the 4 data groups in Fig. 5. The R^2 value in Fig. 5(c) validate the confidence of experimental results. The TDDF (Cases + MF + BSI) outperforms predictions based on other data groups, which shows the effectiveness of our method in heterogenous data fusion for HFMD prediction. Compare Cases with Cases + MF or Cases + BSI, it can be found that MLR, RFR, and GBR benefit from auxiliary data (MF or BSI), but SVR is slightly enhanced. A possible reason is that SVR stacks in few sample training and multiple distribution data, while decision tree based methods perform well at these occasions.



(a) MAE comparison.



(b) RMSE comparison.



(c) R^2 comparison.

Fig. 5. Comparisons of 4 algorithms on 4 data groups in terms of MAE, RMSE, R^2.

5 Conclusion

This paper contributes to the next week HFMD outpatient visits prediction in Xiamen, China. The time difference relevance analysis technique is leveraged to determine the number of lagged weeks for each related factors in meteorological data and Baidu search indices. The statistical model TDDF is proposed to fuse 3 data sources for training and predicting under general regression methods. Extensive experiments of 4 regression algorithms on 4 data groups show the effectiveness of 6 kinds of BSI data, 11 kinds of MF data in decreasing predictive errors, and TDDF in representing these auxiliary data.

The TDDF is a coarse-grained framework, and need to be further studied. One of our future work is windowed time series decomposition using signal processing methods for feature extraction, another is to develop model for windows features in order to better make prediction.

Acknowledgements. This work was supported by the Natural Science Foundation of Fujian Province of China (No. 2018J01539 and No. 2019J01713), and the Xiamen Center for Disease Control and Prevention. The authors would like to thank the editor and anonymous reviewers for their helpful comments in improving the quality of this paper.

References

1. Public Health Emergency Events Emergency Regulations. http://www.nhfpc.gov.cn/yjb/s3580/200804/b41369aac27847dba3e6aebccc72e2f8.shtml/chn (2005)
2. WHO Representative Office China. http://www.wpro.who.int/china/mediacentre/factsheets/hfmd/en/ (2008)
3. National Public Health Emergency Event Information Report and Management Regulations. http://www.nhfpc.gov.cn/mohbgt/pw10601/200804/27519.shtml/chn (2018). Accessed 1 Feb 2016
4. World Health Organization. http://www.who.int/infection-prevention/en/ (2018)
5. Xiamen from Wikipedia. https://en.wikipedia.org/wiki/Xiamen (2019)
6. Chen, S., et al.: The application of meteorological data and search index data in improving the prediction of HFMD: a study of two cities in Guangdong province, China. Sci. Total Environ. **652**, 1013–1021 (2019)
7. Ji, T., et al.: Surveillance, epidemiology, and pathogen spectrum of hand, foot, and mouth disease in mainland of china from 2008 to 2017. Biosaf. Health (2019)
8. Sun, B.J., Chen, H.J., Chen, Y., An, X.D., Zhou, B.S.: The risk factors of acquiring severe hand, foot, and mouth disease: a meta-analysis. Can. J. Infect. Dis. Med. Microbiol. **2018**, 1–12 (2018)
9. McMichael, A.J., Woodruff, R.E.: 14 - climate change and infectious diseases. In: Mayer, K.H., Pizer, H. (eds.) The Social Ecology of Infectious Diseases, pp. 378–407. Academic Press, San Diego (2008)
10. Nourani, V., Alami, M.T., Aminfar, M.H.: A combined neural-wavelet model for prediction of ligvanchai watershed precipitation. Eng. Appl. Artif. Intell. **22**(3), 466–472 (2009)
11. Ooi, M.H., et al.: Identification and validation of clinical predictors for the risk of neurological involvement in children with hand, foot, and mouth disease in sarawak. BMC Infect. Dis. **9**(1), 3 (2009)
12. Shao, Q., Yang, L.: Polynomial spline confidence bands for time series trend. J. Stat. Plann. Infer. **142**(7), 1678–1689 (2012)
13. Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P.: A complete ensemble empirical mode decomposition with adaptive noise. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4144–4147, May 2011
14. Wang, L., Jin, L., Xiong, W., Tu, W., Ye, C.: Infectious disease surveillance in china. In: Yang, W. (ed.) Early Warning for Infectious Disease Outbreak, pp. 23–33. Academic Press, San Diego (2017)
15. Wang, Y., Li, J., Gu, J., Zhou, Z., Wang, Z.: Artificial neural networks for infectious diarrhea prediction using meteorological factors in Shanghai (China). Appl. Soft Comput. **35**, 280–290 (2015)
16. Yang, S., et al.: Epidemiological features of and changes in incidence of infectious diseases in China in the first decade after the sars outbreak: an observational trend study. Lancet. Infect. Dis. **17**(7), 716–725 (2017)