Original Article

DOI: 10.5582/bst.2017.01035

Time series analysis of weekly influenza-like illness rate using a one-year period of factors in random forest regression

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

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

Summary **Influenza, a disease caused by a respiratory virus, sickened over 5,043,127 citizens in Shenzhen, China, from January 2014 to April 2016. An accurate forecasting of outbreaks of influenza-like illness (ILI, here we refer to ILI as the upper respiratory infection) could facilitate public health officials to suggest public health actions earlier. In this study, a random forest regression constructed with a one-year period of factors was adopted to forecast the weekly ILI rate using the clinical data from Shenzhen Health Information Center. The following conclusions were drawn based on this method: *i*) Compared to the predication with 52 (one-year) history observations, the accuracy of the predication was improved by adding another 52 first-order difference variables: mean absolute percentage error (MAPE) decreased from 5.04% to 4.35% and mean squared error (MSE) decreased from 2.85E-04 to 1.97E-04. *ii*) The variables with the first-order difference seemed more significant than the original history observations during the predication. In addition, both the recent observations and the later observations seemed important in the predicating procedure. *iii*) Analysis using the Pearson correlation concluded that weather conditions, the influence of which could have been implied by history observations and seemed insignificant for the predication, showed correlation to the weekly average temperature and maximum temperature. The correlation coefficients were -0.3656 and -0.3583, respectively.**

Keywords: Time series analysis, random forest regression, influenza-like illness (ILI), mean absolute percentage error (MAPE), mean squared error (MSE), correlation

1. Introduction

Influenza is a disease caused by a respiratory virus, and can infect any age group. The illness ranges from mild to severe, and results in the death of thousands annually. An outbreak puts tremendous pressure on both clinicians and patients. The accurate forecasting of influenza outbreaks could facilitate public health officials in taking more timely public health actions, such as suggesting school closures and allocating or temporarily readjusting medical resources for hospitals and medical centers. Studies suggest that by accurately forecasting the outbreak of influenza and by taking preventative and control measures, such as school closures, the impact of influenza could be minimized (*1-3*).

Time-series forecasting methods, which play an important role in disease prediction, analyze the patterns of past outbreaks and formulate a forecasting model from underlying temporal relationships (*4*). The auto-regressive integrated moving average (ARIMA) method was first popularized by Box-Jenkins for analyzing time-series data (*5*). The study used this approach to investigate the influence of winter holiday break on

Released online in J-STAGE as advance publication May 8, 2017.

**Address correspondence to:*
Dr. Denan Lin, Shenzhen Health Information Center, 2210 North of Renmin Road, Luohu, Shenzhen, 518000, China.
E-mail: ldn308@163.com

Dr. Ye Li, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences,1068 Xueyuan Avenue, Shenzhen University Town, Nanshan, Shenzhen 518055, China.
E-mail: ye.li@siat.ac.cn

weekly influenza-like illness rates (*6*). Unfortunately, the ARIMA model suffers from two drawbacks. First, it assumes linear relationships between independent and dependent variables, and second, a constant standard deviation in errors develops in the model over time. Reference compared the performance of ARIMA and random forest time series to predict avian influenza H5N1 outbreaks (*7,8*), which revealed that random forest time series modeling provided enhanced results over existing time series models for the prediction of infectious disease outbreaks. Instead of utilizing clinical data, Google Flu Trends attempted to make accurate predictions by aggregating search queries. Although it achieved an impressive accuracy of 97% in its early stage, Google Flu Trends team no longer published current estimates because of its drop in accuracy in the interval of 2011-2013 (*9,10*). A study by the Institute of Cognitive Science Osnabrück also attempted to predicate flu trends by combining social media data (*e.g.* Twitter) with CDC data (*11*).

A key intuition in this study is that a flu season could be influenced by the conditions of the past year. Therefore the forecasting of weekly influenza-like illness (ILI) rate should consider not only recent observations as used in the traditional approaches (*7,8*) but also much later observations and their difference. Therefore, this paper adopted the weekly rate of the previous one-year observations and their first-order difference to the recent observation as the predicator space, and applied this novel predicator space to the random forest regression method to forecast the weekly ILI rate.

2. Materials and Methods

2.1. *Data sources*

Weather data were obtained from the Weather Channel (*https://weather.com/*), and clinical data were obtained from Shenzhen Health Information Center, which collected clinic visit information from January 1, 2014 to April 10, 2016, 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 weekly ILI rate and the X axis represents outbreak time.

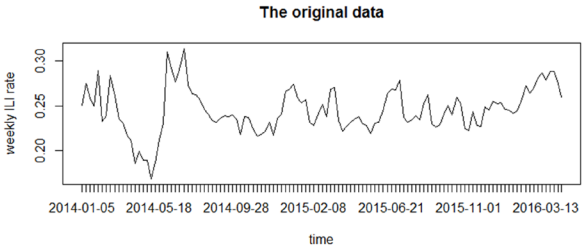

Figure 1. Data from Shenzhen Health Information Center. Y axis represents the weekly ILI rate; X axis represents outbreak time.

2.2. *Methodology*

Random forest regression is a tree-based method that involves stratifying or segmenting the predicator space into a number of simple regions. To make a prediction for a given observation, the mean of the response values of the training observations in the same region is typically applied. There are two steps to build a regression tree as follows:

i) Divide the predicator space $X_1,X_2, \ldots,X_p$ into *j* distinct and non-overlapping regions, $R_1,R_2, \ldots, R_j$;

ii) For every observation that falls into region R_j, the same predication is made, which is simply the mean of the response values for the training observations in R_j.

By bootstrapping the entire training data set multiple times, bagging reduces the high variance to overcome the coherent overfitting problem in decision trees. For K bootstrapped training sets, the final predication for the point x is as follows:

$$f(X) = \frac{1}{K}\sum_{k=1}^{K} T_k(x)$$

All bagged trees look similar to each other if a very strong predicator is always selected in the top split. Random forest de-correlates decision trees by allowing a randomly sampled subset (*m* features) from the full predicator space (*p* features). Random forest utilizes a group of "weak learners" to form a "strong learner" thereby improving the classification or regression performance. Two parameters are important in the random forest algorithm – the number of trees in the forest (*ntree*), and the number of predicators in each tree (*mtry*). In this study, package "randomForest" in R was used. The default value for *ntree* was adopted, and the function *tuneRF* was used to choose the optimal value of *mtry*.

Predicator space In this study, three kinds of components were chosen as the predicator space: history observations, first-order difference values and weather conditions. Assuming the current predicated point was X_0; the first component was the sequence $X_1,X_2,X_3,\ldots,X_{t-1},X_t$, where t was 52, and was filled with the values of the previous 52 observations before X_0. The second component was the sequence $D_1,D_2,D_3,\ldots,D_t$, where *t* was 52, and *Dt* meant the first order difference between X_1 and the previous *t*th observation. The third component was composed of weather conditions *Temperature*, *Humidity*, *Wind_speed*, and *Maximum_temperature*, which denoted the weekly average of temperature, humidity, wind speed, and maximum temperature, respectively.

Metrics Mean absolute percentage error (MAPE) and mean squared error (MSE) were used to measure the predication accuracy (*12*). MAPE and MSE are

defined as the following formula, respectively:

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(F_t - A_t)^2.$$

where A_t is the actual value and F_t is the forecasted value.

Variable importance is a predicator ranking based on the contribution that predicators make to construct a tree. In this study, variable importance was computed using the percent increase in MSE, based upon the mean decrease of accuracy in predications on the out of bag samples when a given variable was excluded from the model.

3. Results

3.1. *Improved predication accuracy*

The data from Jan. 1, 2014, to Oct. 12, 2015 (93 week) were used as training data, and the half-year data from Oct. 19, 2015, to Apr. 10, 2016 (26 weeks) were used as the test data. The experiments were performed by iteratively adding a new week of data, training a new model based on the updated data, and predicating the value for the following week. To investigate the influence of different predicators on the predication accuracy, the experiments were carried out four times by gradually combining more predicators into the predicator space. In the first process, 52 recent observation variables X1,X2,X3,…,X51,X52 were chosen. In the second, 52 difference variables D1,…,D52 were combined into the predicator space. In the third, the weather conditions were added into the predicator space. Finally, the weather conditions were changed into the first-order difference values of each weather condition.

Figure 2A illustrates the first experiment of the weekly ILI rate with the predicator space of 52 recent observations. Figure 2B shows the results of 104 predicators (52 difference predicators were added), which shows that the predication accuracy was improved. Compared to the forecast without using difference predicators, Table 1 shows that by adding the difference predictors MAPE decreased from 5.04% to 4.35% and MSE decreased from 2.85E-04 to 1.97E-04. Here, the detailed results of the last two experiments were not provided anymore because the weather conditions had almost made no influence on the predication accuracy.

3.2. *Comparison of variable importance*

The top six variables in each model were checked according to their weights. Then, each variable was summed up in all the models, and the average weight was obtained and shown in Table 2. The following observations were made: *i*) Without using difference predicators, besides the recent variables (such as

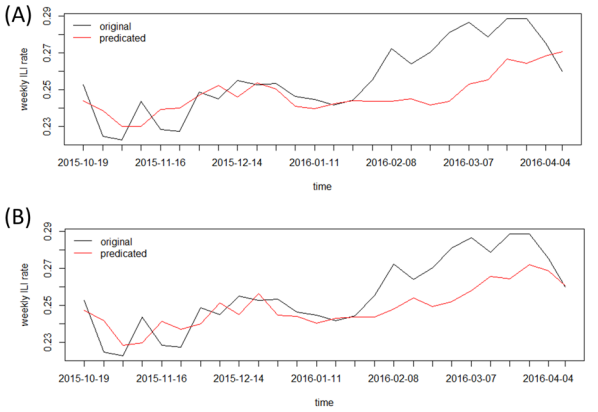

Figure 2. Predication of Weekly ILI rate with different predicator spaces. (A). The predicator space was composed of 52 recent observations. The black line illustrates the original data, and the red line shows the corresponding predicated values. **(B)**. Here, 52 difference predicators were combined into the predicator space. The black line illustrates the original data, and the red line shows the corresponding predicated values

Table 1. Comparison of forecasting with and without difference predicators

Predicator space	MAPE (%)	MSE
Without difference predicators	5.04	2.85E-04
With difference predicators	4.35	1.97E-04

Table 2. Comparisons of variable importance

Variables	V1	V2	V3	V4	V5	V6
Without difference predicators						
Name	X1	X5	X24	X25	X52	X23
Weight (%)	9.11	6.85	4.21	3.64	2.44	1.86
With difference predicators						
Name	D3	D4	D8	X5	D34	D43
Weight (%)	5.58	4.35	4.15	3.65	2.12	1.64

Table 3. Analysis of correlation between weather conditions and weekly ILI rate

Variables	T	H	WS	MaxT	T_d	H_d	WS_d	$MaxT_d$
Coefficient	-0.3656	-0.08224	0.03954	-0.3583	0.1229	-0.1015	-0.1575	0.334
P-Value	0.00235	0.5082	0.7507	0.002907	0.3219	0.4135	0.2029	0.1199

X1, X5, *etc*.), the variables of the middle-distance observations (such as X24, X25) and the one-year-away observation (such as X52) seemed to be important; *ii*) The variables with the first-order difference seemed more important than the original history observations since D3, D4 and D8 have heaver weights than X5.

3.3. *Analysis of weather conditions*

The third and fourth experiments revealed that the addition of weather conditions into the predicator space did not significantly change the predication accuracy. The analysis of variable importance also showed that the weather-condition-related variables have no significant influence on the predication. However, it is already known that weather conditions are somewhat related to ILI, the influence of which could be implied by the history observations during the predication. In this section we did Pearson correlation analysis between ILI and weather conditions. We investigated the Pearson correlation between weekly ILI rate and the weekly average of temperature (*T*), humidity (*H*), wind speed (*WS*), the maximum temperature (*MaxT*), and their first-order difference (the corresponding variables are notated with a subscript *d*), respectively. Table 3 shows the weekly ILI rate in Shenzhen correlated to the weekly average temperature and the maximum temperature. The correlation coefficients were calculated as -0.3656 and -0.3583, respectively.

4. Discussion

Although every flu season is different because of environmental conditions and changes in the flu virus itself, influenza outbreaks could be influenced and predicated by the conditions of past years. The random forest methods used in the current studies (*7,13,14*) utilize the window size for lags is no bigger than three, which means the influence of later observations are not considered. In this study, by evaluating the variable importance, we found that both the recent observations and the later observations were interesting and had significant influence on the predication. The top six variables of the Shenzhen data were X1, X5, X24, X25, X52, and X23 without the difference predicators, and D3, D4, D8, X5, D34, and D43 with the difference predicators. However, because of changes in the virus and environmental factors, it is difficult to explain how and why the later observations influence the current predictions of influenza outbreaks.

Shenzhen has a humid subtropical maritime climate. In the analysis of variable importance, we also checked weather conditions, which could influence influenza virus transmission (*15-17*). The weather conditions, including the weekly average temperature, humidity, wind speed, and maximum temperature, seemed insignificant for predication because their influence could be implied by the history observations. By analyzing the Pearson correlation, we found that the weekly average temperature and maximum temperature showed correlation to the predicated values with correlation coefficients of -0.3656 and -0.3583, respectively. Other factors, such as humidity, showed no apparent relationship. It was also noticed that by averaging the weekly value, the influence of weather conditions could be weakened. In the future, this conclusion should be verified with the investigation of more detailed daily data.

5. Conclusion

In this study, the random forest regression approach was adopted to forecast the weekly ILI rate. Compared to the predication with 52 one-year-previous observations, by adding an additional 52 first-order difference variables the accuracy was improved: the error decreased from 4.35% to 5.04% in MAPE and from 2.85E-04 to 1.97E-04 in MSE for the predication of the weekly ILI rate using the clinic data from the Shenzhen Health Information Center in China. The variables with the first-order difference seemed more important than the original history observations. However, both the recent observations and the later observations seemed to be important in the predicating procedure. By analyzing the Pearson correlation, the weather conditions, the influence of which could have been implied by the history observations and seemed insignificant for the predication, showed correlation coefficients of -0.3656 and -0.3583, respectively, to the weekly average temperature and the maximum temperature.

Acknowledgements

This work has been supported by National High-tech R&D Program (863 Program) of China (SS2015AA020109) and by National Natural Science Foundation of China (81601575).

References

1. Earn DJ, He D, Loeb MB, Fonseca K, Lee BE, Dushoff J. Effects of school closure on incidence of pandemic

influenza in Alberta, Canada. Ann Intern Med. 2012; 156:173-181.
2. Cauchemez S, Valleron AJ, Boëlle PY, Flahault A, Ferguson NM. Estimating the impact of school closure on influenza transmission from Sentinel data. Nature. 2008; 452:750-754.
3. Heymann AD, Hoch I, Valinsky L, Kokia E, Steinberg DM. School closure may be effective in reducing transmission of respiratory viruses in the community. Epidemiol Infect. 2009; 137:1369-1376.
4. Chatfield C. The analysis of time series: An introduction. CRC press, New York, USA, 2016; pp. 1-15.
5. Box GE, Jenkins GM, Reinsel GC. Time series analysis: Forecasting and control. John Wiley & Sons, Hoboken, New Jersey, USA, 2015; pp. 92-100.
6. Gao H, Wong KK, Zheteyeva Y, Shi J, Uzicanin A, Rainey JJ. Comparing observed with predicted weekly influenza-like illness rates during the winter holiday break, United States, 2004-2013. PLoS One. 2015; 10:e0143791.
7. Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics. 2014; 15:276.
8. Breiman L. Random forests. Machine learning. 2001; 45:5-32.
9. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: Traps in big data analysis. Science. 2014; 343:1203-1205.
10. Butler D. When Google got flu wrong. Nature. 2013; 494:155.
11. IBM, University of Osnabrück cooperate to study innovative use of Twitter to predict flu outbreaks. *http://www.research-in-germany.org/en/research-landscape/news/2016/09/2016-09-14-ibm--university-of-osnabr-ck-cooperate-to-study-innovative-use-of-twitter-to-predict-flu-outbreaks.html* (accessed December 16, 2016).
12. Lehmann EL, Casella G. Theory of point estimation. Springer Science & Business Media, New York, USA, 2006.
13. Akhoondzadeh M. Decision Tree, Bagging and Random Forest methods detect TEC seismo-ionospheric anomalies around the time of the Chile,(M_w = 8.8) earthquake of 27 February 2010. Adv Space Res. 2016; 57:2464-2469.
14. Gopakumar S, Tran T, Luo W, Phung D, Venkatesh S. Forecasting patient outflow from wards having no real-time clinical data. Healthcare Informatics (ICHI), 2016 IEEE International Conference on. 2016. IEEE. *http://ieeexplore.ieee.org/document/7776342/?reload=true* (accessed December 16, 2016).
15. Lowen AC, Steel J. Roles of humidity and temperature in shaping influenza seasonality. Virol J. 2014; 88:7692-7695.
16. Lowen AC, Mubareka S, Steel J, Palese P. Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathog. 2007; 3:1470-6.
17. Oong XY, Ng KT, Lam TT, Pang YK, Chan KG, Hanafi NS, Kamarulzaman A, Tee KK. Epidemiological and evolutionary dynamics of influenza B viruses in Malaysia, 2012-2014. PLoS One. 2015; 10:e0136254.

(Received February 22, 2017; Revised April 7, 2017; Accepted April 18, 2017)