Forecasting incidence of infectious diarrhea using Random forest in Jiangsu province, China

Xinyu Fang

Nanjing Medical University

Wendong Liu

Jiangsu Province Center for Disease Control and Prevention

Jing Ai

Jiangsu Province Center for Disease Control and Prevention

Ying Wu

Jiangsu Province Center for Disease Control and Prevention

Yingying Shi

Jiangsu Province Center for Disease Control and Prevention

Wenqi Shen

Jiangsu Province Center for Disease Control and Prevention

Changjun Bao (✉ bao2000_cn@163.com)

Jiangsu Province Center for Disease Control and Prevention <https://orcid.org/0000-0002-0546-1338>

Research article

Keywords: Infectious diarrhea; Forecasting; Random forest

Posted Date: October 9th, 2019

DOI: <https://doi.org/10.21203/rs.2.15862/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at BMC Infectious Diseases on March 14th, 2020. See the published version at <https://doi.org/10.1186/s12879-020-4930-2>.

Abstract

Background: Infectious diarrhea caused considerable disease burden around the world. An accurate forecast of infectious diarrhea epidemic based on predictive models is crucial for public health authorities. We aimed to develop an optimal random forest (RF) model to predict infectious diarrhea incidence using meteorological factors in Jiangsu province, China. **Methods:** A RF model was constructed to predict infectious diarrhea incidence in Jiangsu province based on the morbidity during 2012-2017 and synchronous meteorological data, then compared with classical ARIMA and ARIMAX model to evaluate the model prediction effect. **Results:** The RF model was established including atmosphere pressure, precipitation, relative humidity and their lagged terms, one to four weeks lagged morbidity and time variable as predictors. Meanwhile, a univariate model ARIMA(1,0,1)(1,0,0) 52 (AIC=-575.92, BIC=-558.14) and a multivariable model ARIMAX(1,0,1)(1,0,0) 52 with precipitation and its one-week lagged term as external variables (AIC=-578.58, BIC=-578.13) were developed as benchmark models. Predictive values of RF model matched the real morbidity very well, with a mean absolute percentage (MAPE) about 20%. The performance of the multivariate ARIMAX model was similar to that of the univariate ARIMA model. Both of them performed much worse than the RF model, with a MAPE as high as near to 30%. **Conclusions:** In brief, RF model well integrated autocorrelation and seasonality of morbidity. It also comprehensively combined meteorological factors and their hysteresis effects. This model well fitted the dynamic of infectious diarrhea epidemic and achieved ideal predication accuracy. It could be used to predict the epidemic level, and has a good prospect of practical application.

Background

Infectious diarrhea is one of the crucial causes of morbidity and mortality in infants and young children, which is a global major public health issue, especially in developing countries [1]. In 2015, diarrheal disease caused an estimated 688 million illnesses and 499,000 deaths worldwide among children younger than 5 years [2]. In the past decade, the morbidity has also shown an increasing trend in various regions in China [3-4]. An accurate forecast of infectious diarrhea activity based on predictive models is crucial for public health authorities to better understand the epidemic characteristics, track its seasonal changes in advance and facilitate key response actions, such as disease surveillance, deployment of emergency supplies, and hospital resource management [5-7].

Recently, some studies have been focused on predicting infectious diarrhea, using autoregressive integrated moving average (ARIMA) model. For example, Yang E et al. used ARIMA model in the early warning systems for diarrhea [5]. However, it does not consider the influence of the external condition on the morbidity. Many studies have investigated the association between climate and infectious diarrhea. Chou WC, et al [8] indicated that the maximum temperature and extreme rainfall days were strongly related to diarrhea morbidity in Taiwan. Significant positive associations were found between increased diarrhea incidence and high temperature, high relative humidity, and high cumulative rainfall in a Mekong Delta area [9]. Thus, climatic factors could be pivotal predictors of infectious diarrhea. Multivariable ARIMA (ARIMAX) model was established to predict bacillary dysentery morbidity in Changsha city with

temperature, relative humidity and air pressure [6]. Long Y, et al [7] also developed an ARIMAX model with temperature and rainfall. Unfortunately, ARIMAX is essentially a linear method. Meteorological factors were proved to be non-linearly associated with the epidemic of infectious diarrhea [8-9]. Therefore, ARIMAX model is inappropriate to fit the relationships between climatic factors and the incidence of infectious diarrhea.

Random forest (RF) model can effectively extract nonlinear relationships in data, which starts by creating classification and regression trees and each constituent tree trains on a potentially nonlinear regression space and is then combined with others. It has been widely used in infectious diseases prediction such as to predict West Nile virus infection and avian influenza incidence [10-11]. Meanwhile, Gustavo M, et al [12] trained a RF model of Bovine viral diarrhea which presenting a satisfactory predictive power. To date, however, there has been no literature report on using RF model to predict the epidemic of infectious diarrhea.

In this study, we aimed to develop an optimal RF model to predict infectious diarrhea epidemic using meteorological factors in Jiangsu province, China. An early warning system for infectious diarrhea might be based on this model to facilitate preventive strategies more effectively.

Methods

Data sources

In China, infectious diarrhea (excluding cholera, dysentery, typhoid and paratyphoid) is an intestinal infectious disease with diarrhea and/or vomiting as the main symptom, which has been listed as a legal Class C infectious disease [3-4].

The weekly number of infectious diarrhea cases in Jiangsu province during 2012-2017 were downloaded from the National Notifiable Disease Surveillance System[13] (<http://www.cdpc.chinacdc.cn>). The demographic data were collected from the Jiangsu provincial statistics department. And the weekly meteorological variables were calculated based on the daily data which were gained from Jiangsu Meteorological Service Center, including atmospheric pressure, mean temperature, maximum temperature, minimum temperature, precipitation, relative humidity and sunshine duration.

ARIMA and ARIMAX model

Autoregressive integrated moving average (ARIMA), namely Box-Jenkins model, is widely used to time series analysis [14]. Developed from the ARIMA model, seasonal ARIMA incorporating seasonal variation performs better in the presence of an obvious seasonal pattern [15-17]. It is denoted as $ARIMA(p,d,q)(P,D,Q)_s$, in which p, d, q indicate orders of general auto-regression (AR), differencing and moving average (MA) terms; P, D, Q are orders of seasonal AR, differencing and MA terms, respectively; s means seasonal periodicity ($s=52$ weeks in this study)[18].

Fitting ARIMA model consists of three essential steps as follows:

First, augmented Dickey-Fuller (ADF) test is implemented to detect whether the original time series is stable or not. If not, logarithmic transformation or difference is adopted to achieve stability.

Second, alternative ARIMA models are established for the stationary time series, the one with the minimum Akaike information criterion (AIC) and Bayesian information criterion (BIC) values is considered as the optimal model. Whereafter, model parameters were estimated using conditional least square method.

Third, Box-Ljung test is conducted to test if the model is statistically significant or not. Finally, prospective prediction is conducted using the optimal model.

Based on the optimal ARIMA model, the multivariate ARIMA including climate factors as external regressors [19] was further developed, which was denoted as ARIMAX . To date, both ARIMA and ARIMAX have been deemed as a classical time series method for forecasting the incidence of infectious diseases. In this study, they were used as a reference to evaluate the performance of random forest model. And cross correlations were used to identify the lagged associations (lags of up to four steps [20-21]) between meteorological factors and the incidence of infectious diarrhea.

RF Model

Random forest (RF) model is an ensemble machine learning method proposed by Breiman [22]. RF creates multiple classification and regression (CART) trees, each trained on a bootstrap sample of the original training data with a randomly selected subset of input variables, and takes the average of outputs produced by the trees as the final prediction. One of the most important features of RF is to calculate the variable importance which measures the association between a given variable and the predication accuracy.

Fitting RF model consists of four essential steps as follows [11]:

First, bootstrap sampling method is used to randomly select sample units from the original training data to create multiple CART trees.

Second, bootstrap sampling method is also used to select candidate variables for each CART tree. In this study, the related climate variables were chosen as the predictors. Meanwhile, one-week lagged morbidity and variable “time” also incorporated into the RF model in order to consider the effect of autocorrelation and seasonality of the dependent, respectively.

Third, the average of outputs produced by all the CART trees was calculated as the final predictive value.

Fourth, the importance of each variable was assessed by the decrease in accuracy.

Model evaluation

Three models were fitted in this study, RF model with meteorological variables, univariate ARIMA model and multivariate ARIMAX model. The data subset during 2012-2016 were used as training set to fit models, and the rest in 2017 were used as testing set to evaluate the forecasting accuracy of different models. Root mean square error (RMSE) and mean absolute percentage error (MAPE) were selected as the measures to evaluate the performance of each model, which were calculated as the following formulas:

$$RMSE = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}}$$

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|\hat{y}_t - y_t|}{y_t}$$

where n means number of real data or predicted values, means real data and means predicted value.

Statistics analysis

All analyses were performed in R (version 3.5.1). Seasonal decomposition was carried out to elucidate the temporal pattern of infectious diarrhea. RF models were fitted using the “randomForest” package. ARIMA and ARIMAX models were fitted using the “Forecast” package. Statistical significance was defined as $P < 0.05$.

Results

General description

Totally 102,020 cases were detected during 2012-2017 in Jiangsu province, China, reaching an average annual incidence rate of 21.40 per 100,000. As shown in figure 1, the incidence presented an increasing long-term trend in the 6 years. Moreover, there was a distinct seasonality, two incidence peaks were observed in each year, a higher winter peak from December to next February and a lower summer peak from July to September. The descriptive statistics for meteorological variables were summarized in table 1.

Correlation analysis

According to the cross-correlation analysis shown in table 2, atmospheric pressure and precipitation were significantly associated with the incidence of infectious diarrhea at lagged 0-2 and 0-3, respectively. Meanwhile, relative humidity was related with the incidence but no hysteresis effect was observed. Notably, the temperature variables as well as sunshine duration were not correlated with the incidence.

Model fitting

ARIMAX model

The original time series of infectious diarrhea morbidity was stationary (Dickey-Fuller=-4.26, $p<0.01$). Alternative univariate ARIMA models were developed. The best-fitting ARIMA model was determined to be ARIMA(1,0,1)(1,0,0)₅₂, with a minimum AIC=-575.92 and a minimum BIC=-558.14. Ljung-Box test ($\chi^2=0.01$, $P=0.93$) suggests that the residual series of the model was white noise sequence.

Accordingly, several multivariate ARIMAX models were fitted based on the optimal ARIMA model. Finally, ARIMAX(1,0,1)(1,1,0)₅₂ with precipitation and its one week lagged term as external predictors was identified as the optimal ARIMAX model, with a minimum AIC=-578.58 and a minimum BIC=-578.13 (Ljung-Box test: $\chi^2=0.00548$, $P=0.10$).

RF model

A RF model was constructed, including atmosphere pressure, precipitation and their lagged terms, relative humidity, one to four weeks lagged morbidity and time as predictors. As shown in figure 2, the lagged dependent terms were the most imperative in all predictors. Atmospheric pressure and its lagged term were proved to be the most vital climate factors, which was followed by lagged precipitation.

Note: lag1_rate, lag2_rate, lag3_rate indicate one-week lagged morbidity, two week lagged morbidity, three week lagged morbidity respectively. AP, lag1_AP, lag2_AP indicate atmospheric pressure and its lagged terms respectively. PPTN, lag1_PPTN precipitation, lag2_PPTN precipitation, lag3_PPTN precipitation, lag4_PPTN indicate precipitation and its lagged terms respectively. RH means relative humidity.

Prediction performance comparison

The comparison of the three models was summarized in table 3, and the predicting outputs were displayed in figure 3. RF model with climate variables outperformed ARIMA model and ARIMAX model in both model fitting stage and prospective forecasting stage. The predicted values by RF matched the real incidences very well, with a MAPE about 20%, which suggests that this RF model would be able to accurately estimate the prevalence and seasonal fluctuation of infectious diarrhea. The performance of ARIMAX model was similar with that of ARIMA model. Both of them performed much worse than the RF model, with a high MAPE near to 30%.

Note: the left side of the vertical line is model fitting stage and the right is forecasting stage.

Discussion

The incidence of infectious diarrhea in Jiangsu province presented a long-term trend of gradual growth. Mathematical models for predicting are urgently needed to reinforce an integrated management for monitoring, control and prevention of infectious diarrhea. Some researchers have introduced different prediction methods for this purpose [23-25]. In this study, a RF model with meteorological factors was constructed. This model presented a well accuracy in predicting the incidence of infectious diarrhea with

a MAPE about 20%, and accurately estimated the seasonal fluctuation of this disease. It may be used as an important tool for the public health authorities.

To evaluate the performance of the RF model, we developed a univariate ARIMA model and a multivariate ARIMAX model which included precipitation and its one-week lagged term as external predictors. According to the results, the ARIMA and ARIMAX models reached a MAPE of 28.53% and 28.06%, respectively. The introduction of climate factors did not significantly optimize the prediction accuracy of the ARIMA model. And both of them performed much worse than the RF model. It may be due to two reasons. Firstly, ARIMA model is essentially one kind of linear methods [26]. However, the incidence of infectious diarrhea in the study area may not satisfy the assumption of linearity. Besides, meteorological factors were proved to be non-linearly associated with the epidemic of infectious diarrhea [27-28], so ARIMA model is inappropriate to fit the relationships between predictors and the disease. On the contrary, RF model can effectively extract nonlinear relationships contained in data [29]. Secondly, compared with ARIMAX model, RF is not subject to multicollinearity, many due to random selecting variables for each tree in RF [30]. In this study, we selected predictor variables through cross correlation analysis. The meteorological factors and their lagged terms were incorporated into the models so long as being identified to be significantly correlated with the incidence of infectious diarrhea. And all of them presented some degree of importance, which suggests that the RF model comprehensively combined the climatic variables and their hysteresis effects. To sum up, RF model is more suitable than ARIMA method to predict the infectious diarrhea epidemic in the study region.

In this study, we found that the lagged incidences in the previous weeks were more important than meteorological factors to improve the prediction accuracy of the RF model. It suggests that RF model took advantage of the autocorrelation of the dependent variable. Particularly, the infectious diarrhea epidemic in Jiangsu province presented an obvious seasonality which is a critical feature to fit predicting model. However, RF is unable to learn the seasonal pattern because of randomly selecting samples for each tree [11,22]. In this study, we introduced a variable “time” into the model to fit the seasonal variance of the disease. The importance analysis shows that it played a significant role to optimize the model. This strategy is worthy of reference when fitting the similar RF models.

It is worth mentioning that atmospheric pressure, precipitation and relative humidity were all correlated with the incidence of infectious diarrhea in Jiangsu province with relevant lag times from zero to three weeks. It is not completely consistent with other studies [20-21], which may due to the regional differences of pathogen composition and climate conditions. Furthermore, the meteorological factors significantly contributed to the forecasting power of RF model, with atmospheric pressure and its lagged values playing the leading role. Under high atmospheric pressure, the radiation cooling at night may form the ground inversion layer, pathogens of infectious diarrhea, like viruses and bacteria, are easy to stagnate near the ground, thus inducing disease. Atmospheric pressure also may influence survival of diarrhea causing microorganisms in the environmental though effects tend to be species specific with both negative and positive survival impacts [31]. Precipitation and its lagged values showed moderate importance in RF model, especially three-week lagged effect. It indicated that precipitation in the previous

three weeks may availablely influenced the following morbidity and can be used for prediction of it. Especially, relative humidity was identified as the least important factor. The relative humidity in Jiangsu Province had a narrow variation at the weekly level and failed to fit better with the incidence rate. These findings may provide clues for the further study of specific relationship between climate and infectious diarrhea.

There are some limitations in this study. First, some mild cases might use home therapies, and some cases with atypical symptoms may be misdiagnosed, so the data reported may underestimate the morbidity. Second, in this study, just meteorological variables were considered to improve the predication ability. Other factors associated with infectious diarrhea may also be used as good predictors, which deserves further studies. Third, our study based on provincial-wide data, does not take into account geographical disparities and pathogen heterogeneity within the province. Further studies accounting for geographical disparities, such as rural or urban, may be useful. Fourth, similar to many other machine learning methods such as artificial neural networks, RF model can not explain the specific association between relative factors and disease.

Conclusions

The RF model with climate factors presented satisfactory prediction accuracy. It could be used to predict the epidemic level, and has a good prospect of practical applications. The autocorrelation and seasonal variation contained in the data of the dependent variables are crucial for the predication model. Meanwhile, the effects of meteorological factors and cumulative effects over a period of time were combined to improve the model. Further researches are warranted to explore RF model with meteorological factors as well as other variables has the potential to be a useful tool for predicting other major infectious diseases.

Abbreviations

RF: random forest ARIMA: autoregressive integrated moving average AP: atmospheric pressure, Mean_T: mean temperature, MAX_T: maximum temperature, MIN_T: minimum temperature, RH: relative humidity, PPTN: precipitation, SD: sunshine duration

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study are available from National Notifiable Disease Surveillance System but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Jiangsu Provincial Major Science & Technology Demonstration Project (No. BE2017749); Jiangsu Provincial natural science foundation (No. BK20151595); Jiangsu Provincial Medical Youth Talent (No. QNRC2016542); and Key Medical Discipline of Epidemiology (No. ZDXK A2016008).

Authors' contributions

Fang XY, Liu WD. Liu and Ai J designed the study. Wu Y, Shi YY, Shen WQ prepared and processed the initial data, Fang XY and Liu WD analyzed the data. X. Y. Fang and W. D. Liu drafted the manuscript. Bao CJ revised the manuscript. All authors read and approved the final manuscript

Acknowledgements

We gratefully acknowledge Jiangsu Meteorological Service Center for supplying the daily meteorological data.

References

- [1]GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015.Lancet. 2016; 388(10053): 1545-1602.
- [2]GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016; 388(10053):1459-1544.
- [3]Li J, Pan H, Xiao WJ, Gong XH, Zhuang Y, Kuang XZ, et al. Epidemiological and etiological surveillance study of infectious diarrhea in Shanghai in 2013-2015. Zhonghua Yu Fang Yi Xue Za Zhi. 2017; 51(12):1113-1117.
- [4]Zhang P, Zhang J. Surveillance on other infectious diarrheal diseases in China from 2014 to 2015. Zhong hua Liu Xing Bing Xue Za Zhi. 2017; 38(4):424-430.

- [5]Yang E, Park HW, Choi YH, Kim J, Munkhdalai L, Musa I, et al. A Simulation-Based Study on the Comparison of Statistical and Time Series Forecasting Methods for Early Detection of Infectious Disease Outbreaks. *Int J Environ Res Public Health*. 2018; 15(5): pii: E966. doi: 10.3390/ijerph15050966.
- [6]Gao L, Zhang Y, Ding G, Liu Q, Zhou M, Li X, et al. Meteorological variables and bacillary dysentery cases in Changsha City, China. *Am J Trop Med Hyg* 2014; 90(4): 697-704.
- [7]Yan L, Wang H, Zhang X, Li MY, He J. Impact of meteorological factors on the incidence of bacillary dysentery in Beijing, China: A time series analysis (1970-2012). *PLoS One*. 2017; 12(8): e0182937.
- [8]Chou WC, Wu JL, Wang YC, Huang H, Sung FC, Chuang CY. Modeling the impact of climate variability on diarrhea-associated diseases in Taiwan (1996-2007). *Sci Total Environ*. 2010; 409(1): 43-51.
- [9]Phung D, Huang C, Rutherford S, Chu C, Wang X, Nguyen M, et al. Association between climate factors and diarrhea in a Mekong Delta area. *Int J Biometeorol*. 2015; 59(9):1321-31.
- [10]Keyel AC, Elison Timm O, Backenson PB, Prussing C, Quinones S, McDonough K A et al. Seasonal temperatures and hydrological conditions improve the prediction of West Nile virus infection rates in *Culex* mosquitoes and human case counts in New York and Connecticut. *PLOS ONE*. 2019; 14(6): e0217854.
- [11]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.
- [12]Machado G, Mendoza MR, Corbellini LG. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. *Vet Res*. 2015; 24: 46-85.
- [13]Wang L, Wang Y, Jin S, Wu Z, Chin DP, Koplan JP et al. Emergence and control of infectious diseases in China. *Lancet*. 2008; 372(9649): 1598-605.
- [14]Lin Y, Chen M, Chen G, Wu X, Lin T. Application of an autoregressive integrated moving average model for predicting injury mortality in Xiamen, China. *BMJ Open*. 2015; 5(12): e008491.
- [15]Xu Q, Li R, Liu Y, Luo C, Xu A, Xue F, et al. Forecasting the Incidence of Mumps in Zibo City Based on a SARIMA Model. *Int J Environ Res Public Health* 2017; 14(18): 925.
- [16]Tian CW, Wang H, Luo XM. Time-series modelling and forecasting of hand, foot and mouth disease cases in China from 2008 to 2018. *Epidemiol Infect*. 2019; 147: e28.
- [17]Duan Y, Huang XL, Wang YJ, Zhang JQ, Zhang Q, Dang YW, et al. Impact of meteorological changes on the incidence of scarlet fever in Hefei City, China. *Int J Biometeorol*. 2016; 60(10):1543-1550.
- [18]Cryer JD, Chan KS. Time Series Analysis: with applications in R. *Journal of the Royal Statistical Society*. 2011; 174: 507-507.

[19]Soebiyanto RP, Adimi F, Kiang RK. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. PLoS One. 2010; 5(3): e9450.

[20]Tao Y, Liu ZM, Mi SQ, Song J, Qiang L. Effects of meteorological factors on other types of infectious diarrhea. Journal of Lanzhou University (Natural Sciences). 2015; 51: 646-651.

[21]Wang J, Xu MM, Mo YZ, Pan XC. Correlation between meteorological factors and infectious diarrhea in a district of Beijing. J Environ Health. 2013; 30: 991-995.

[22]Breiman L. Machine Learning: Kluwer Academic Publishers. California, USA, 2011.pp. 5-32.

[23]Zhao D, Wang L, Cheng J, Xu J, Xu Z, Xie M, et al. Impact of weather factors on hand, foot and mouth disease, and its role in short-term incidence trend forecast in Huainan City, Anhui Province. Int J Biometeorol. 2017; 61(3):453-461.

[24]Zhang X, Hou F, Qiao Z, Li X, Zhou L, Liu Y. Zhang, T. Temporal and long-term trend analysis of class C notifiable diseases in China from 2009 to 2014. BMJ Open. 2016; 6(10): e011038.

[25]Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White NJ, Kaewkungwal J. Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan. Malar J. 2010; 9: 251.

[26]Chadsuthi S, Modchang C, Lenbury Y, Iamsirithaworn S, Triampo W. Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time-series and ARIMAX analyses. Asian Pac J Trop Med. 2012; 5(7): 539-46

[27]Harper SL, Edge VL, Schuster-Wallace CJ, Berke O, McEwen SA. Weather, water quality and infectious gastrointestinal illness in two Inuit communities in Nunatsiavut, Canada: potential implications for climate change. Eco health. 2011; 8(1): 93-108.

[28]Alexander KA, Carzolio M, Goodin D, Vance E. Climate change is likely to worsen the public health threat of diarrhoeal disease in Botswana. Int J Environ Res Public Health. 2013; 10(4): 1202-30.

[29]Speybroeck N. Classification and regression trees. Int J Public Health. 2012; 57(1): 243-6.

[30]Biau G. Analysis of a random forests model. The Journal of Machine Learning Research 2012; 13: 1063–1095.

[31]Tang JW. The effect of environmental parameters on the survival of airborne infectious agents. J R Soc Interface. 2009; 15: 276.

Tables

Table 1. Summary of weekly meteorological factors in Jiangsu province, 2012-2017

Variable	Min	P25	P50	Median	P75	Max
AP (Pa) ¹	998.58	1007.02	1015.38	1015.15	1022.56	1032.09
Mean_T (°C) ²	-2.19	7.39	17.13	15.92	23.67	32.65
MAX_T (°C) ³	1.25	12.36	22.38	20.47	27.60	37.41
MIN_T (°C) ⁴	-4.77	3.59	13.08	12.22	20.63	28.24
RH (%) ⁵	45.93	68.06	74.69	73.96	80.40	91.88
PPTN (mm) ⁶	0.00	3.53	11.94	21.45	30.12	59.66
SD (hour) ⁷	2.25	27.71	37.50	38.52	48.72	82.01

Note: ¹atmospheric pressure, ²mean temperature, ³maximum temperature, ⁴minimum temperature, ⁵relative humidity, ⁶precipitation, ⁷ sunshine duration

Table2. Cross correlation coefficients between infectious diarrhea and meteorological factors in Jiangsu province, 2012-2017

Lag	AP (Pa) ¹	Mean_T (°C) ²	MAX_T (°C) ³	MIN_T (°C) ⁴	RH (%) ⁵	PPTN (mm) ⁶	SD (hour) ⁷
0	0.21*	-0.10	-0.09	-0.11	-0.13*	-0.23*	0.07
1	0.17*	-0.06	-0.06	-0.07	-0.08	-0.22*	0.05
2	0.12*	-0.02	-0.01	-0.02	-0.04	-0.14*	0.03
3	0.08	0.03	0.03	0.03	-0.02	-0.12*	0.04
4	0.04	0.08	0.08	0.08	0.04	-0.08	0.05

Note: ¹atmospheric pressure, ²mean temperature, ³maximum temperature, ⁴minimum temperature, ⁵relative humidity, ⁶precipitation, ⁷ sunshine duration, **P*<0.05

Table 3. Performance evaluation of RF and ARIMA model

model	RMSE		MAPE(%)	
	Training set	Testing set	Training set	Testing set
RF	0.04	0.31	6.88	20.89
ARIMAX (1,0,1)(1,0,0) ₅₂	0.08	0.46	13.64	28.06
ARIMA(1,0,1)(1,0,0) ₅₂	0.08	0.45	13.78	28.53

Figures

Figure 1

Temporal distribution of infectious diarrhea in Jiangsu province, 2012-2017

Figure 2

Variable importance in random forest regression model for infectious diarrhea

Figure 3

Plot of observed infectious diarrhea incidences and predicted values via different models