PLOS ONE

RESEARCH ARTICLE

A dengue fever predicting model based on Baidu search index data and climate data in South China

Dan Liu[1☯], Songjing Guo[2☯], Mingjun Zou[1☯], Cong Chen[1], Fei Deng[3], Zhong Xie[2,4], Sheng Hu[2], Liang Wu[2,4]*

1 School of Medicine, Wuhan University of Science and Technology, Wuhan, China, **2** School of Geography and Information Engineering, China University of Geosciences, Wuhan, China, **3** State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China, **4** National Engineering Research Center for GIS, Wuhan, China

☯ These authors contributed equally to this work.
* wuliang@cug.edu.cn

Citation: Liu D, Guo S, Zou M, Chen C, Deng F, Xie Z, et al. (2019) A dengue fever predicting model based on Baidu search index data and climate data in South China. PLoS ONE 14(12): e0226841. https://doi.org/10.1371/journal.pone.0226841

Editor: Abdallah M. Samy, Faculty of Science, Ain Shams University (ASU), EGYPT

Received: July 16, 2019

Accepted: December 4, 2019

Published: December 30, 2019

Data Availability Statement: The data underlying the results presented in the study are third party data available from http://index.baidu.com/ (The Baidu index data). Dengue fever case data used in this article are owned by the Chinese Center for Disease Control and Prevention and are available through an application on the website (http://www.phsciencedata.cn/Share/ky_sjml.jsp). The users need to fill in a large amount of personal information, and corresponding staff will reply after a period of time. the authors did not have any special access that other researchers would not

Abstract

With the acceleration of global urbanization and climate change, dengue fever is spreading worldwide. Different levels of dengue fever have also occurred in China, especially in southern China, causing enormous economic losses. Unfortunately, there is no effective treatment for dengue, and the most popular dengue vaccine does not exhibit good curative effects. Therefore, we developed a Generalized Additive Mixed Model (GAMM) that gathered climate factors (mean temperature, relative humidity and precipitation) and Baidu search data during 2011–2015 in Guangzhou city to improve the accuracy of dengue fever prediction. Firstly, the time series dengue fever data were decomposed into seasonal, trend and remainder components by the seasonal-trend decomposition procedure based on loess (STL). Secondly, the time lag of variables was determined in cross-correlation analysis and the order of autocorrelation was estimated using autocorrelation (ACF) and partial autocorrelation functions (PACF). Finally, the GAMM was built and evaluated by comparing it with Generalized Additive Mode (GAM). Experimental results indicated that the GAMM (R^2: 0.95 and RMSE: 34.1) has a superior prediction capability than GAM (R^2: 0.86 and RMSE: 121.9). The study could help the government agencies and hospitals respond early to dengue fever outbreak.

Introduction

Dengue fever (DF), an acute vector-borne disease caused by dengue virus, belongs to the Flaviviridae family, and is transmitted by mosquito vectors [1–3]. People who become infected can develop clinical symptoms with different levels, such as mild fever, headache, muscle and joint pain. In severe cases, bleeding, shock and even death can occur [2–4]. DF spreads widely in tropical and subtropical regions, such as Africa, Americas, Southeast Asia, and the western Pacific [5]. Evidence has shown that nearly half of the world's population face the threat of DF, and 390 million individuals were infected with dengue per year, of which nearly 96 million

have. Guangzhou Meteorological Data are available within the Supporting Information files.

Funding: This project was supported by the National Key Research and Development Program (grant number2017YFB0503601) and was supported by the National Science Foundation of China (grant number 41871311, 71774127). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

have clinical symptoms [6]. In the past 50 years, the incidence of DF has increased 30-fold [7–9]. In the 21st century, DF has transmitted rapidly and has become a serious public health problem.

There were no cases of dengue between 1949 and 1977 in mainland China until the first dengue outbreak occurred in Guangdong province in 1978. Since then, different scales of DF have appeared in Fujian, Guangxi, Zhejiang, and other provinces in China [10, 11]. In recent years, DF has spread further and has also been observed in Henan province of central China. Up to now, there are still no definitive treatment for DF, and popular dengue vaccines have not achieved satisfactory results [12–16]. Therefore, establishing an accurately and early prediction system is an important means for adequate preparedness and response to the outbreak of DF.

Many early warning models were developed in term of different data sources and different methods. The traditional data mainly included dengue cases data, meteorological data, media data (density of mosquito vector) and social factors data. Among them, meteorological data is an important factor for warning models to forecast dengue fever outbreaks [17–19]. Besides, with the development of the Internet in recent years, Internet search data, such as Google, Wikipedia, Yahoo, has been demonstrated as a good complement to traditional monitoring data and can reflect the outbreak of some diseases in some extent. [20–26]. The warning models can be divided into two categories according to analyze the change features of various variables in time latitude: the qualitative warning methods and the quantitative warning methods. The qualitative warning methods estimated disease of development trends and intensity based on the occurrence and development of historical cases. But it relied on the baseline that compared the current number of cases with the historical data in the same period which has a lot of room for improvement in predictive performance [27]. The quantitative warning models predicted future development trend of the disease based on the established mathematical model, such as regression analysis [28–30], time series methods [31–33]. Previous studies have found relationships between DF and meteorological data was non-linear[34]. Nowadays, Generalized Additive Model (GAM) represented by natural cubic splines are the main models used in environmental epidemiology [35–37]. However, GAM is essentially a probabilistic model, explaining how dependent variables depend on independent variables randomly, not showing how dependent variables depend on each other. Besides, GAM requires that each observation is independent, autocorrelation of the cases may cause the problematic estimates of the model.

The aim of this study is to find a better prediction model for estimate and predict the time and scale of the DF. Therefore, we proposed the Generalized Additive Mixed Model (GAMM), which combined the GAM and an autocorrelation term (AR). In this model, we introduced the Dengue Baidu Search Index (DBSI) as variable besides climatic factors (mean temperature, relative humidity, and precipitation) and added the autocorrelation item of DF cases. Firstly, the time series DF data were decomposed into seasonal, trend and remainder components by the seasonal-trend decomposition procedure based on loess (STL). Secondly, the time lag of variables were determined in cross-correlation analysis and the order of autocorrelation was estimated using autocorrelation (ACF) and partial autocorrelation functions (PACF). Finally, the GAMM was built and evaluated by comparing with the GAM. This study can help hospital management to allocate medical resources, and also help us to monitor the abnormal incidence of diseases.

Materials and methods

Study setting

Different levels of dengue fever have occurred in China, especially in Guangzhou, Guangdong province, China. Guangzhou, located between 112˚57'E and 114˚03' E longitude and 22˚26'N

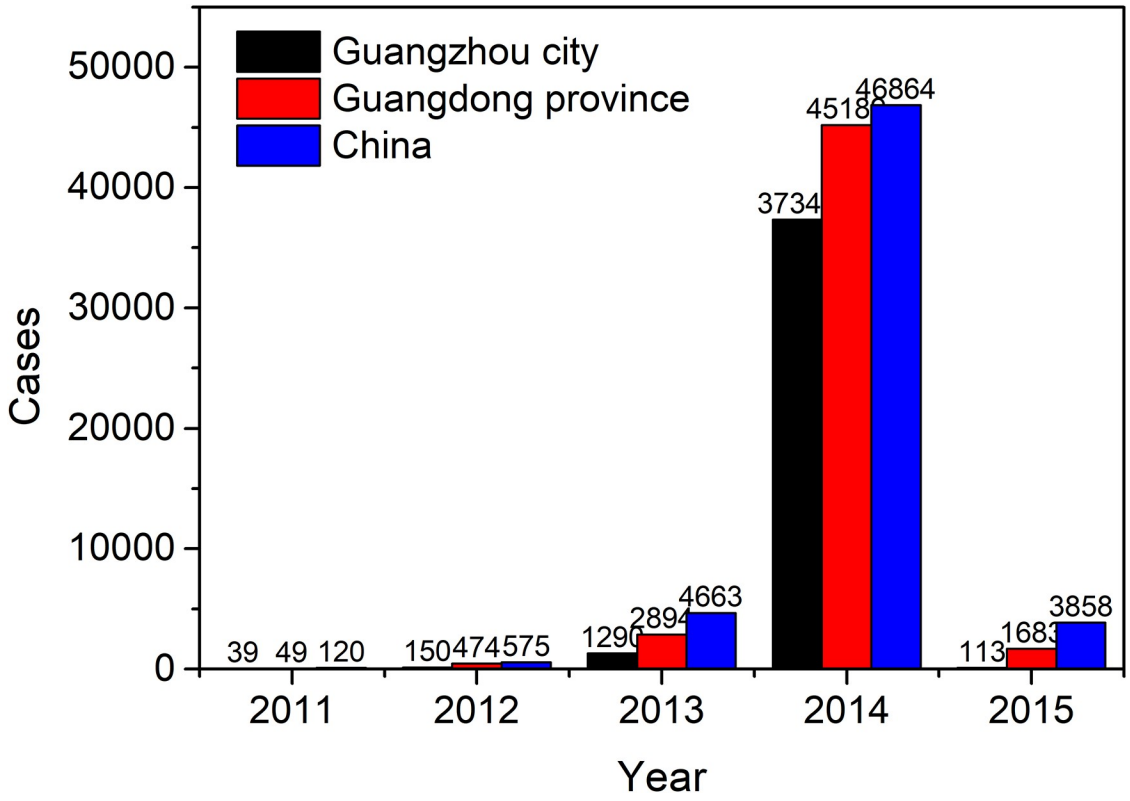

Fig 1. Annual dengue incidence in China, Guangdong, and Guangzhou from 2011 to 2015 respectively.

https://doi.org/10.1371/journal.pone.0226841.g001

and 23˚56' N latitude, has 12 municipal districts with an area of 7434 km^2 and a population of about 13 million [38]. It has a maritime subtropical monsoon climate with an annual average temperature of around 22˚C, and more rainfall, which is suitable for the growth of mosquito vectors. The above factors increase the risk of dengue transmission.

Guangzhou has always been being a high-risk area for DF in mainland China, and the number of reported dengue cases have ranked first in the country. During 2011–2015, the number of dengue cases in Guangzhou accounted for 77.4% (38932/50289) of the cases in Guangdong Province and 69.4% (38932/56080) of the cases in China during the same period (Fig 1). In particular, in 2014, there were 37,354 dengue cases in Guangzhou city, accounting for 70.57% of the total number of cases in China since 1978 [30]. Hence, Guangzhou is an ideal area to study dengue in China.

Data sources

Dengue cases data. The dengue cases data of Guangzhou from January 2011 to December 2015 in this study were obtained from the Public Health Science Data Center (http://www.phsciencedata.cn/Share/ky_sjml.jsp) whose disease data come from a statutory infectious diseases report database established by Chinese Center for Disease Control and Prevention. The database collected all statutory reported infectious diseases data since 2004. Information of dengue cases included location of the report, sex, age, occupation, the number of the infected in multi-dimension, morbidity, death toll, and mortality. Dengue cases were diagnosed according to the China National Diagnostic Criteria for dengue fever (WS216–2008) enacted by the Chinese Ministry of Health [39].

Meteorological data. Meteorological data from January 2011 to December 2015 in Guangzhou were obtained from Guangzhou Meteorological Data Website (http://data.tqyb.com.cn/weather/history_weather_fm.jsp) including the monthly average temperature (˚C), monthly mean relative humidity, and precipitation (mm) (S1 Dataset). The climate data used were collected by meteorological station widely distributed in China. The monthly average meteorological data of a city were calculated using the area-weighted average method.

The Baidu index data. Baidu is the most popular search engine in China, accounting for more than 80% market share [40]. Therefore, the Baidu search index website (http://ndex.baidu.com/) is used as a data source, which based on the search behavior of users in Baidu. We could obtain the daily counts at the national, provincial, and city level since January 2011. Search data were extracted based on a monthly basis and city level for the study period.

Statistical analysis

Keywords selection and filtering. The network search volume of different keywords could affect the precision of the prediction model. Therefore, the selection of keywords is critical. Unfortunately, there are no clear principles and standards for the selection of keywords [25, 41–43]. Previous studies have mainly used the names of the disease, clinical symptoms, and diagnosis as the main terms to search more related terms. In this study, we obtained the related keywords from a Chinese website (http://tool.chinaz.com/baidu/words.aspx) to minimize the omission of main terms [40]. Typing original terms, we obtained 40 related keywords (S1 Table) suggested by different websites: recommendations of Baidu, portals websites, blogs, and online reports using semantically related analytics. Some recommended keywords may not be closely related to DF occurrence, which could reduce the detective ability of models [28, 44]. Therefore, keywords were filtered as follows: 1) the keywords irrelevant to DF and those with a search volume of zero in the website of the Baidu index were eliminated; and 2) the Spearman's rank correlation coefficients(ρ_1) between monthly dengue cases and search volumes were calculated during the corresponding period. We excluded the terms with correlation coefficient less than 0.4 and those correlations without statistical significance ($P > 0.05$). 11 keywords were left finally (S2 Table).

The remaining terms were used to combine the DBSI. Weights of terms were defined by the values of the correlation coefficients (ρi). The DBSI was calculated as follow:

$$weight_i = \frac{\rho_i}{\sum_{i=1}^{n} \rho_i} \tag{1}$$

$$\mathrm{DBSI} = \sum_{i=1}^{n} weight_i \times keyword_i \tag{2}$$

Where n is the number of keywords, $keyword_i$ and $weight_i$ represent the *i*th keyword of the search volume in the Baidu index website and the weight of the *i*th keyword, respectively.

Seasonal-trend decomposition procedure based on loess (STL). Local dengue fever of outbreaks are affected by many factors including meteorological factors, medium, and human factors. Therefore, dengue cases data showed obvious volatility and it is difficult to see the seasonal trend from the original time series data. Seasonal-trend decomposition procedure based on loess (STL) can analyze local dengue cases and find whether there were long-term and seasonal trends, which laid the foundation for whether to add the long-term trend and seasonal trend of control in the prediction model. Here, the STL method was used to divide dengue

cases data into trend term, seasonal term, and residual term. The expression was as follows:

$$Y_t = Trend_t + Seasonal_t + \mathrm{R}emainder_t \quad (3)$$

Where Y_t represents the local dengue cases at time t, $Trend_t$ represents the trend term, $Seasonal_t$ represents the seasonal term, $Remainder_t$ represents the residual term, and t is the time in unit of month, t ranges from 1 to N. Long-term trends and seasonal trends of local dengue cases were observed.

Establishment of the GAMM. Firstly, the autocorrelation term was determined through the ACF and PACF of dengue cases. Secondly, spearman cross-correlation analysis was carried out to identify the correlation between dengue cases and the average temperature, precipitation, relative humidity, and the DBSI for the lag of 0–6 months. The lag term with the largest Spearman correlation coefficient for each variable was selected. Finally, GAMM was established. Considering the impact of past dengue cases on current dengue cases, the autocorrelation term was also added to improve the accuracy of the predicting model. Because dengue was a small probability event related to the entire population and satisfied the Poisson distribution, the study used the quasi-Poisson model to allow for excessive dispersion of dengue cases data. The smooth natural cubic spline functions on these risk factors were used to reflect the non-linear association between dengue cases and each dependent variable. In order to evaluate the quality of the proposed model, we made comparisons between the GAMM and the GAM without adding the autocorrelation term. The basic form of the models to be constructed were as follows:

$$\ln[E(\mu_t)] = \beta_0 + s(T_{t-a}, df) + s(H_{t-b}, df) + s(P_{t-c}, df) + s(DBSI_{t-d}, df) + year + month \quad (4)$$

$$\ln[E(\mu_t)] = \beta_0 + s(T_{t-a}, df) + s(H_{t-b}, df) + s(P_{t-c}, df) + s(DBSI_{t-d}, df) + year + month + \alpha_0 \quad (5)$$

$$\alpha_0 = \sum_{j=1}^{p} C_j \left((\ln(\mu^*_{t-j}) - \sum_{i=1}^{m} f(X_{t-ji}) \right) \quad (6)$$

Formula (4) and (5) are expressions of the GAM and the GAMM respectively. μ_t is the number of dengue cases in month t, $E(\mu_t)$ is the expected value of the t-month of dengue cases, log(*) is the link function of model, and β_0 is a constant term. Additionally, $s(T_{t-a}, df)$ is the natural cubic spline function of the average temperature in the previous a month with the corresponding *df*, $s(H_{t-b}, df)$ is the natural cubic spline function of the relative humidity in the previous b month with the corresponding *df*, $s(P_{t-c}, df)$ is the natural cubic spline function of the precipitation, in the previous c month with corresponding *df*, $s(DBSI_{t-d}, df)$ is the natural cubic spline function of the DBSI in the previous d month with corresponding *df*. Year controls long-term trends, and month controls seasonal trends. α_0 is an autocorrelation. $y^* = \max(y, \tau)$, $\tau = 0.5$, τ is to prevent *y* from being 0 and negative.

In the models, the *df* directly or indirectly affect the fitting effect of the model, and the choice of *df* values were essential. In this study, the *df* of independent variables were determined by the local minimum principle of Akaike Information Criterion(AIC)[45]. Autocorrelation term (*P* value) was determined by ACF and PACF for dengue cases. In addition, the model had to further control the volatility of long-term trends and seasonal trends so that the average temperature, relative humidity, 8precipitation, and the DBSI could better estimate the number of dengue cases.

In this study, we divided the disease data into two parts: The model was developed on the data from January 2011 to June 2015 and validated using the remaining data. For the GAM

and GAMM, R^2 was applied to evaluate the goodness of model fitting. Larger R^2 values were associated with stronger explanatory ability of the model. Furthermore, in order to fully evaluate the model, the ACF and PACF of the residual were used to test whether the residual of the model was independent. Moreover, we also analyzed the relationships between the Pearson residual and the time to test independence hypothesis of the model. If the Pearson residual of the model did not change with time, the model conformed to the independence hypothesis.

Finally, the root mean square error (RMSE) was devoted to evaluate the quality of the model, which tested the consistency between local dengue cases and predicted cases of the two models. The RMSE reflects the error of circumstances between the actual and predicted values. Smaller RMSE was associated with better prediction ability.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (\bar{y}_t - y_t)^2}{n}} \tag{7}$$

Where $\bar{y}_t$ is the actual of local dengue case data at time t, y_t is the predicted value of dengue case data in the model at time t, and n is the size of samples for prediction.

All the analyses were performed in R version 3.4.0. Results with *P* values of less than 0.05 were considered statistically significant in all statistical tests. The model was built and analyzed using the "mgcv" package. All relevant data are within the paper and Supporting Information files. The code needed to reproduce the results presented here are available on github at https://github.com/guoandwu08180914/hellodata123.git.

Results

Factors influencing dengue cases

During the period from 2011 to 2015, dengue fever cases of Guangzhou have indicted different degrees every year, mainly occurred from August to November, late summer and autumn, with a large outbreak in 2014 (Fig 2). The monthly average temperature, relative humidity, and precipitation all shows seasonal fluctuations and reaches the peak at April to September, March to July, May to October, respectively. Moreover, relative humidity and precipitation increased year by year, whereas average temperature do not change substantially. The time series of the monthly DBSI is significantly similar to the time series of monthly dengue cases, which also shows two peaks in 2013 and 2014. The peak in 2014 is particularly prominent, indicating that the DBSI could reflects changes in the dengue epidemic (S1 Fig).

The decomposition result showed that DF in Guangzhou city had an increased treed from 2011 to 2014, and then mildly decrease in 2015. Besides, DF also had a seasonal distribution, which were prone to have more case from August to November every year. Therefore, long-term and seasonal effects on local dengue cases should be considered when building the prediction model (Fig 3).

The auto-correlation analysis showed that the autocorrelation coefficient with a lag of 1 and the partial auto-correlation coefficient cutoff at lag 2, indicating that the number of dengue cases at lags 1 months affects the number of dengue cases in the current month (Fig 4). Therefore, the autocorrelation item of dengue cases must be considered when building the model. According to the results of the autocorrelation analysis, the *P* value of the autocorrelation item was set to 1.

Spearman cross-correlation analysis of dengue fever cases with monthly mean temperature, monthly relative humidity, monthly cumulative rainfall, and DBSI with a lag time of 0–6 months showed that the DF had the strongest cross-correlation coefficients with lag of 2

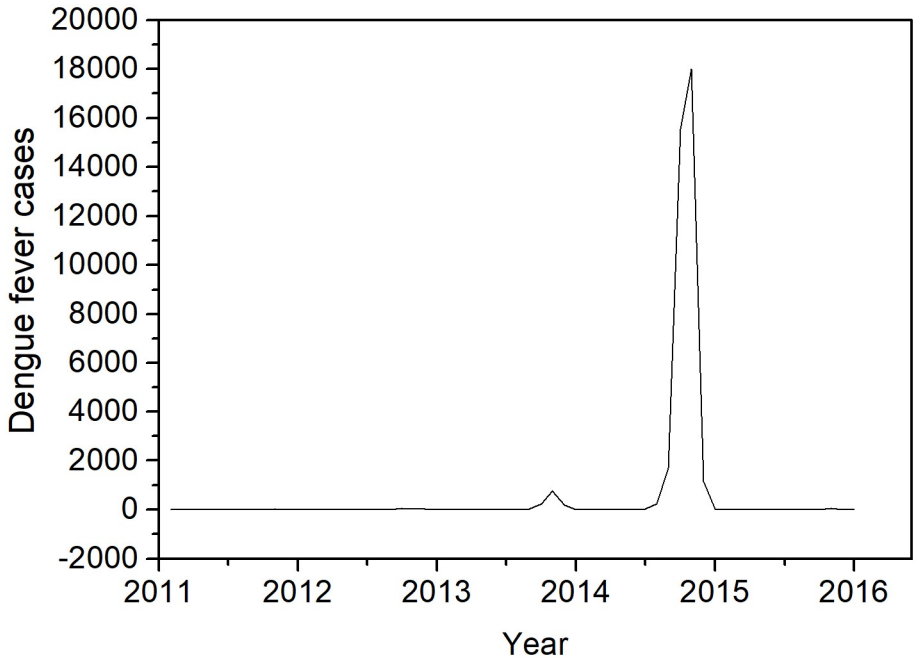

Fig 2. Temporal distribution of dengue fever cases in Guangzhou, 2011–2015.

https://doi.org/10.1371/journal.pone.0226841.g002

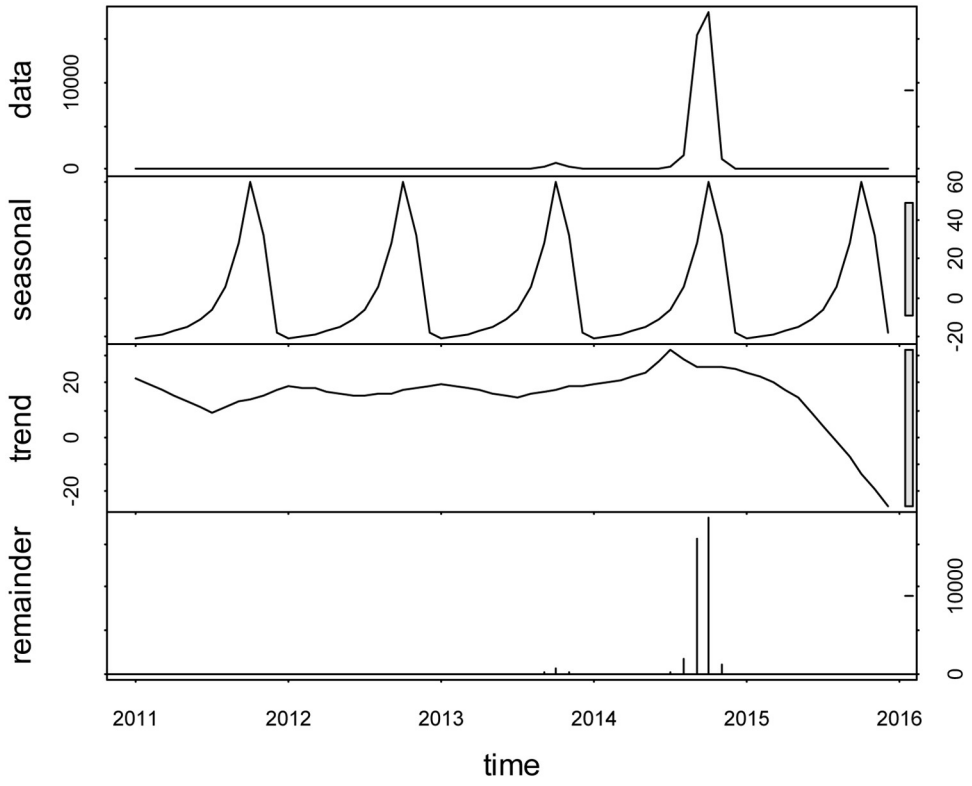

Fig 3. The decomposition plot of local dengue cases in the study areas from January 2011 to December 2015.

https://doi.org/10.1371/journal.pone.0226841.g003

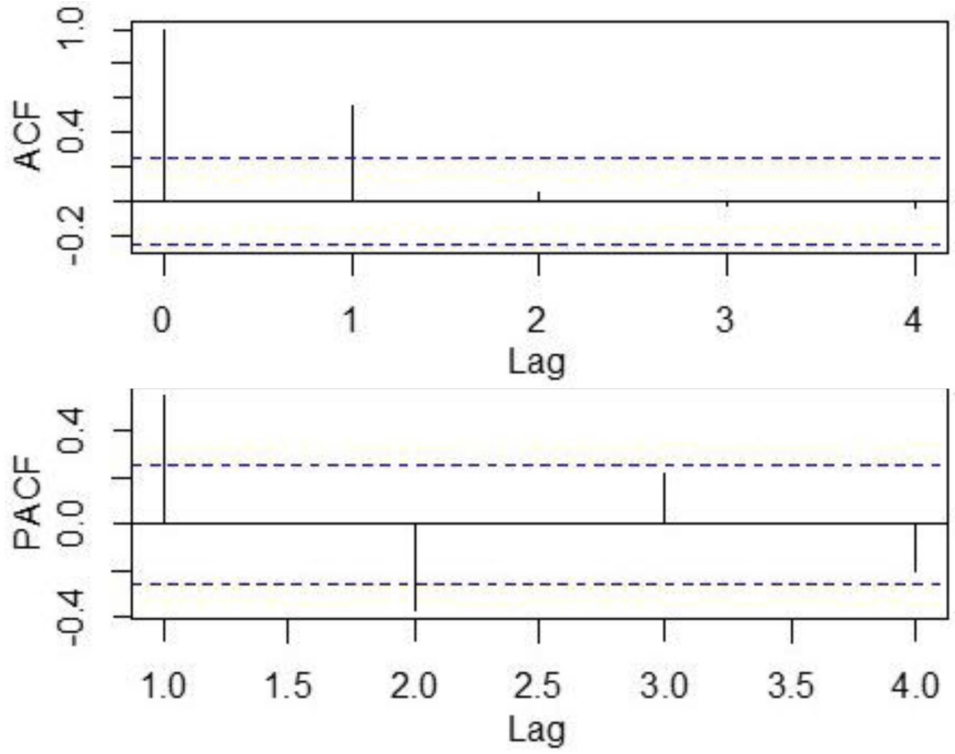

Fig 4. Auto-correlation and partial auto-correlation plots of dengue cases, 2011–2015.

https://doi.org/10.1371/journal.pone.0226841.g004

months for the average temperature, lag of 3 months for relative humidity, lag of 3 months for precipitation, and lag of 0 month for DBSI, their Spearman cross-correlation coefficients are 0.718, 0.605, 0.692, and 0.734, respectively (Table 1).

Measures of the ability of model fitting and forecasting

The R^2 of GAM was 0.86 and the R^2 of GAMM was 0.95. The fitting results exhibited that GAMM has a better fitting performance compared with GAM (Fig 5). The residual test (Fig 6) by ACF and PACF showed residuals of GAM have obvious auto-correlation. However, residual in the GAMM were not correlated, which indicated that the information of these variables was extracted sufficiently.

Table 1. Cross-correlation analysis between dengue cases and average temperature, relative humidity, precipitation, DBSI.

	Temperature	Humidity	Precipitation	DBSI
Lag0	**0.404***	-0.073	0.055	**0.734***
Lag1	**0.618***	0.143	0.238	**0.672***
Lag2	**0.718***	**0.456***	**0.497***	**0.536***
Lag3	**0.619***	**0.605***	**0.692***	**0.416***
Lag4	**0.356***	**0.600***	**0.594***	0.270
Lag5	0.022	**0.444***	0.318	0.135
Lag6	-0.361	0.225	-0.022	0.024

Notes: Each positive answer equals 1 point,

*$p<0.05$.

https://doi.org/10.1371/journal.pone.0226841.t001

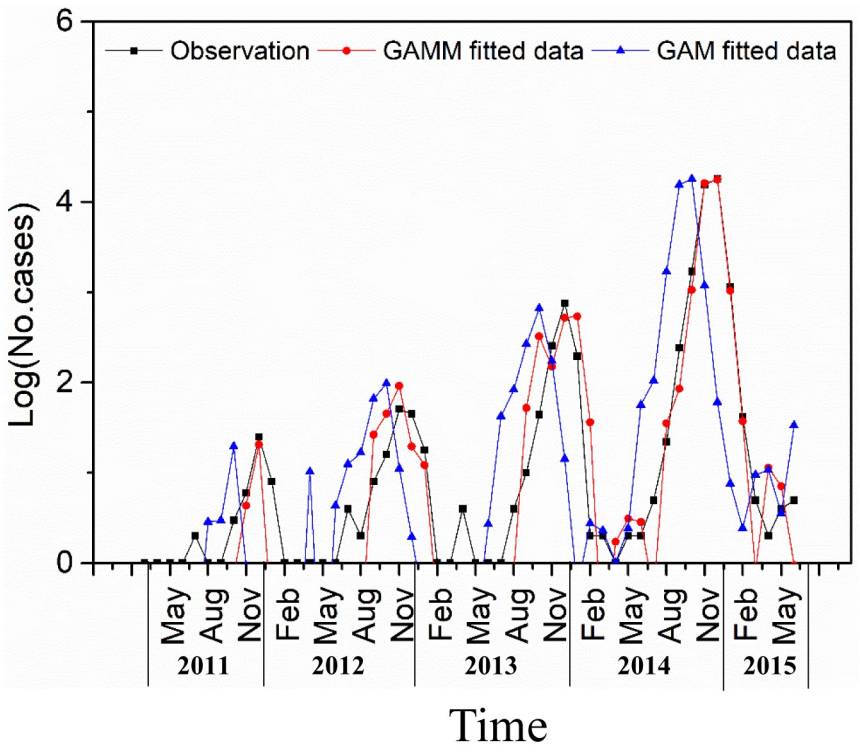

Fig 5. Monthly observed DF cases and fitted local DF cases using two different models from January 2011 to June 2015.

https://doi.org/10.1371/journal.pone.0226841.g005

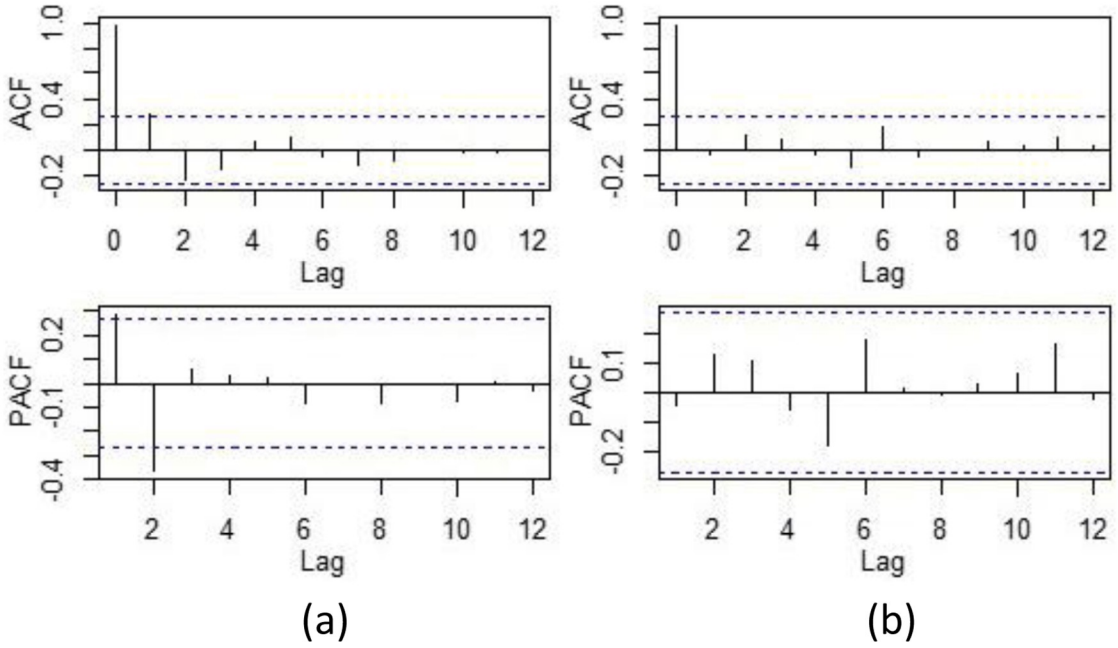

Fig 6. Auto-correlation and partial auto-correlation of residuals. (a) ACF/PACF plot of the Pearson residual of the GAM. (b) ACF/PACF plot of the Pearson residual of the GAMM.

https://doi.org/10.1371/journal.pone.0226841.g006

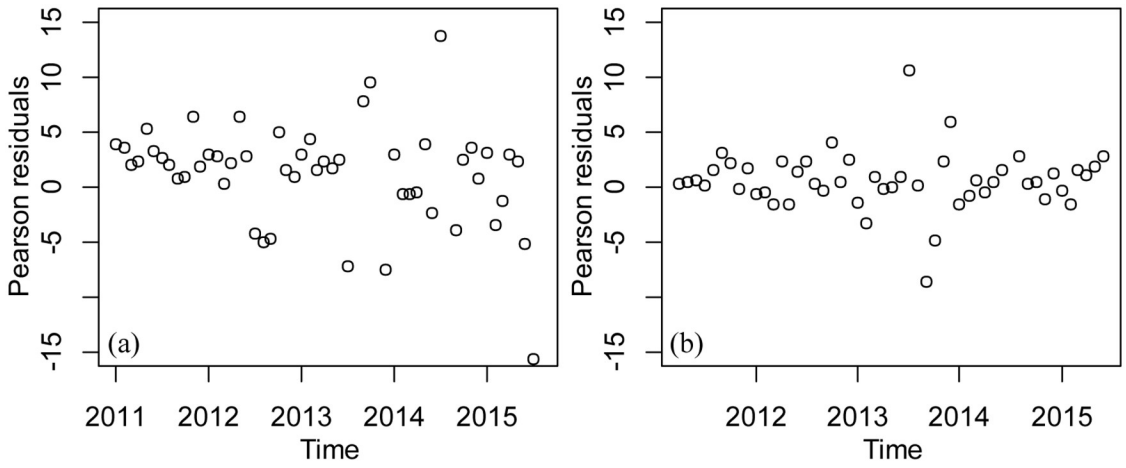

Fig 7. Scatter plot of residuals for predicted values using two different models and dates. (a) Scatter plot of residual for predicted values and dates in the GAM. (b) Scatter plot of residual for predicted values and dates in the GAMM.

https://doi.org/10.1371/journal.pone.0226841.g007

Besides, the Pearson residual of the GAM model has significant fluctuations over time, violating the model assumptions of GAM (Fig 7). The forecast result showed that the GAMM (RMSE: 121.9) gives a better prediction of DF cases than the GAM (RMSE: 34.1) (Fig 8).

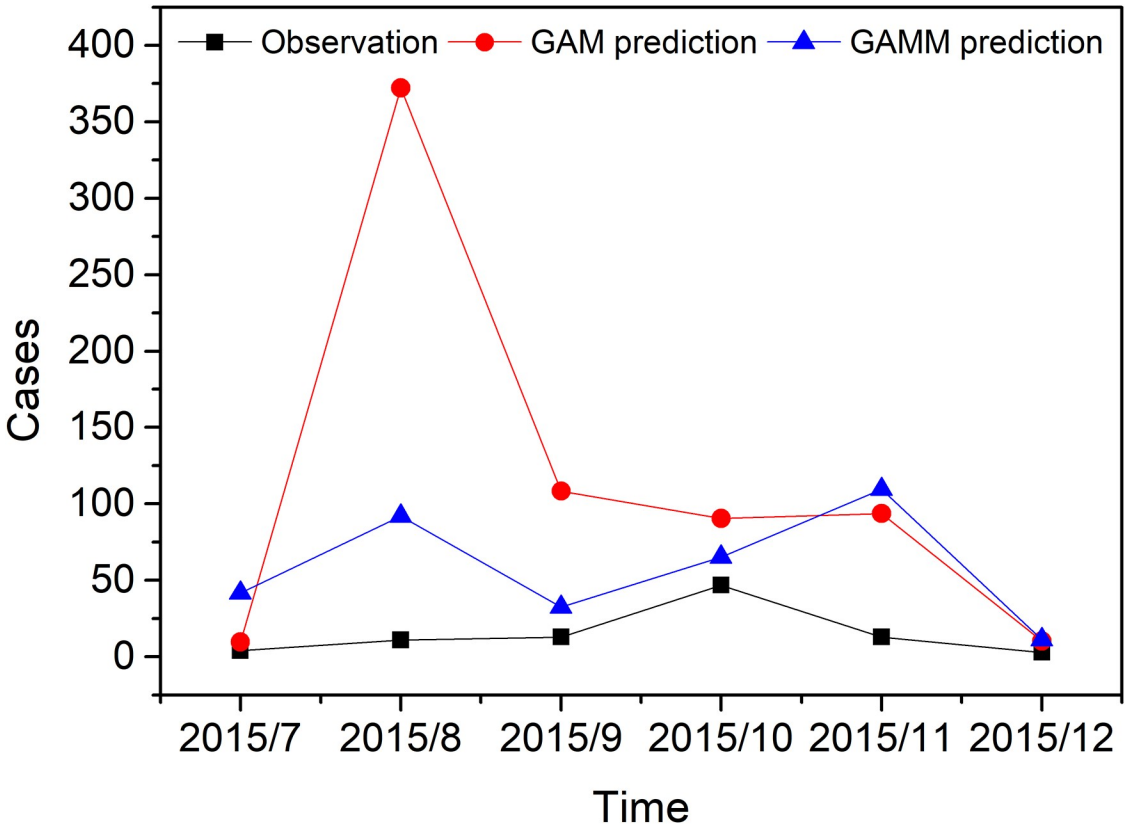

Fig 8. Observations and model predictions of DF case using two different models from July 2015 to December 2015.

https://doi.org/10.1371/journal.pone.0226841.g008

Discussion

Since the beginning of the 21st century, DF has become one of the most serious infectious diseases worldwide and has seriously affected public health in many countries. From January 2011 to December 2015, there were 38,277 cases of dengue fever in Guangzhou city, accounting for 76% of dengue cases in Guangdong Province during the same period. In 2014, Guangzhou city experienced the largest outbreak of dengue fever in history, during two months (September and October), the number of cases of dengue fever exceeded 30,000, which was higher than the number of cases in other countries and regions, leading to panic in Guangzhou residents. Given that there are currently no specific drugs for dengue treatment and no reliable vaccine for prevention, the World Health Organization (WHO) believes that the establishment of early warning system is a crucial means to deal with the outbreak of dengue fever.

Accurate and timely forecasts of dengue incidence is essential in China. Government and hospitals can timely grasp the time and scale of DF outbreak to take preventive and control measures that reduce the number of dengue cases. In this study, we introduced meteorological factors (mean temperature, relative humidity, precipitation) and DBSI into the GAMM model. The autocorrelation item of dengue cases was also considered in the prediction model. The results indicated that the proposed model could effectively monitor dengue fever cases in Guangzhou city for a short period.

DF is transmitted form of human-mosquito-human. At present, there is no effective vaccine against dengue in China. The control of mosquitoes is the most effective way to prevent DF in endemic countries. For example, *wolbachia* [46, 47], spray pesticides, and genetic modification [48] have achieved good effects. However, the establishment of an early DF warning system that enhance the predictability of dengue outbreaks is still a key step in dengue control. Various countries, including Nepal[49] Thailand [50], Singapore [51], Indonesia [52–54], Venezuela [55], and Trinidad [56], have conducted numerous studies on DF, which reported that meteorological factors play a crucial role in dengue transmission by direct or indirect effects on mosquito vectors[57, 58]. The impact of climate on dengue fever is not immediate but accumulates over time. Study found that the mosquito-borne mating rates were highest at 20˚C and 25˚C [59]. In this study, dengue cases of Guangzhou city are sharp increase in August to September, lag after two months (during June to July), the average annual temperature of Guangzhou city of 20˚C to 25˚C, which is consistent with previous studies. Notably, during this period, the growth of *Aedes* can be reduced by spraying with syrup or eliminating water accumulation. Our study also indicated that lag of 3 months for relative humidity had a large impact on the number of dengue. The possible reasons are that eggs and adult mosquitoes are affected by relative humidity, and the number of mosquito ovulations and adult mosquitoes are reduced in dry areas [60]. Besides, we also found that precipitation in the previous 3 months was positively correlated with dengue case in current month that agreement with previous studies [61], highlighting the importance of precipitation in the spread of dengue fever. Heavy precipitation can flush away the egg, larvae and pupae of *Aedes* mosquitoes in the short term, prevent the growth of mosquito vectors. In addition, precipitation shorten the time that people are exposed to mosquitoes and the rate of mosquito bites, leading to reduce the risk of DF infection in a certain extent. However, precipitation can also provide an excellent breeding condition for the growth of mosquito larvae, increasing the number of mosquitoes in the long run [62, 63].

When users search information by the Internet, the frequency of the specific phrases, time and place and other relevant information are preserved by web search engine, which reflect the behavior tendency of the Internet users. Internet search data has comprehensive, massive and real-time characteristics, which can better reflect the tendency and timeliness of people's

social behavior. With the rapid development of Internet technology, more and more people use Internet search for health information, which provides data basis for disease monitoring model. At present, many scholars have used Google, Yahoo, Wiki, Twitter, and other web searches to conduct research to predict the prevalence of seasonal diseases and have made certain achievements [22, 24, 26, 64]. According to the 39th Statistical Report on Internet Development, the total number of Internet users in China was 73.1 million by 2016, accounting for about 53.2% of the national population. Most of them use the Baidu engine[65]. Therefore, we added the DBSI to the model and found that there was a strong correlation (coefficient of association 0.73424) between current month DBSI and dengue cases in Guangzhou city, which suggests that dengue patients are usually aware of their condition and will conduct relevant Internet searches current month before see a doctor. Study proves that DBSI can provide a supplement f traditional disease surveillance.

Moreover, we validated our model by comparing the actual DF values and the predicted values of the GAM and GAMM in the last 6 months of 2015, we further discovered that the GAMM with autocorrelation item had better performance in the dynamic monitoring of dengue cases in Guangzhou city. The GAM only explained the effects of the climate factors and the DBSI on the change of DF, not included an unpredictable long-term trend of DF, whereas the GAMM took into account the maximum past information that affected the current dengue cases, which was more robust for adjusting the long-term trend. Therefore, the GAMM was a better choice for prediction models.

Although the study achieved a good result, there were some limitations to the model. Firstly, the amount of sample data was small. In this study, we only used data from January 2011 to December 2015. The monthly accuracy data for dengue cases in 1990–2015 could be obtained from the Public Health Science Data Center website, however, we could only obtain the Baidu search data after January 2011. Therefore, this study contains 5 years of experimental data at most, and this may have affected the robustness of the model. In future studies, weekly/day data can be considered to replace monthly data. Secondly, Baidu search data account for the largest market for Internet data in China and have become the largest data source for tracking Chinese search behavior. However, these data do not represent all Internet data. Moreover, the Baidu search index provides a dimensionless data, and the official website does not give a specific calculation method, indicating that these data only partly reflect the trends in dengue cases. Besides, dengue Baidu search behaviors are easily affected by the news media and government, resulting in uncertainties in such search data. There is no specific way to overcome this problem currently. In further studies, we will combine the commonly used Weibo data in China with the Baidu data to build models and evaluate media-driven interests or other events that change search behavior. Finally, other risk indicator data that may be associated with dengue fever does not examine, such as population data, socio-economic, urbanization status, mosquito control measures, and animal herd immunity.

Conclusions

In the 21st century, dengue fever has transmitted rapidly and has become a serious public health problem. Early warning system is great important to the treatment and prevention of DF. Therefore, in this study, we introduced the Dengue Baidu Search Index (DBSI) as variable besides climatic factors (mean temperature, relative humidity, and precipitation) and added the autocorrelation item of DF cases to establish the GAMM. Experimental results indicated that the GAMM has a superior prediction capability. The study could offer the assist with public health interventions to prevent and control the dengue outbreak.

Supporting information

S1 Dataset. Meteorological data.
(XLSX)

S1 Fig. Temporal distribution of climate variables and Baidu search variable from July 2015 to December 2015.
(TIF)

S1 Table. Search keywords from Baidu index website in this study.
(XLSX)

S2 Table. Dengue related Baidu search terms that were finally selected.
(XLSX)

Acknowledgments

We would like to thank China's National Engineering Research Center for Geographic Information Systems (NERCGIS) for providing hardware support.

Author Contributions

Data curation: Songjing Guo, Sheng Hu.

Formal analysis: Liang Wu.

Funding acquisition: Dan Liu.

Methodology: Dan Liu, Songjing Guo, Cong Chen, Liang Wu.

Resources: Cong Chen.

Supervision: Dan Liu, Mingjun Zou, Fei Deng, Zhong Xie, Liang Wu.

Validation: Songjing Guo.

Writing – original draft: Songjing Guo.

Writing – review & editing: Mingjun Zou, Cong Chen, Liang Wu.

References

1. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature. 2013; 496:504. https://doi.org/10.1038/nature12060 PMID: 23563266
2. Tatem AJ, Hay SI, Rogers DJ. Global traffic and disease vector dispersal. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(16):6242–7. https://doi.org/10.1073/pnas.0508391103 PMID: 16606847
3. Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, Hoen AG, et al. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. PLoS neglected tropical diseases. 2012; 6(8):e1760. https://doi.org/10.1371/journal.pntd.0001760 PMID: 22880140.
4. Diseases PFIT, Organization WH. Dengue: guidelines for diagnosis, treatment, prevention and control. Geneva World Health Organization. 2009; 6(12):990.
5. Foster JE, Bennett SN, Vaughan H, Vorndam V, Mcmillan WO, Carrington CV. Molecular evolution and phylogeny of dengue type 4 virus in the Caribbean. Virology. 2003; 306(1):126–34. https://doi.org/10.1016/s0042-6822(02)00033-8 PMID: 12620805
6. Lanciotti RS, Lewis JG, Gubler DJ, Trent DW. Molecular evolution and epidemiology of dengue-3 viruses. J Gen Virol. 1994; 75(Pt 1):65–75. https://doi.org/10.1099/0022-1317-75-1-65 PMID: 8113741.
7. Lai S, Huang Z, Zhou H, Anders KL, Perkins TA, Yin W, et al. The changing epidemiology of dengue in China, 1990–2014: a descriptive analysis of 25 years of nationwide surveillance data. BMC Medicine. 2015; 13(1):100. https://doi.org/10.1186/s12916-015-0336-1 PMID: 25925417

8. Simmons CP, Farrar JJ, Vv N, Wills B. Dengue. New England Journal of Medicine. 2012; 366(15):399–401.
9. Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, et al. Developing a dengue forecast model using machine learning: A case study in China. PLoS neglected tropical diseases. 2017; 11(10):e0005973. https://doi.org/10.1371/journal.pntd.0005973 PMID: 29036169.
10. Wu JY, Lun ZR, James AA, Chen XG. Dengue Fever in Mainland China. American Journal of Tropical Medicine & Hygiene. 2010; 83(3):664.
11. Lo CL, Yip SP, Leung PH. Seroprevalence of dengue in the general population of Hong Kong. Tropical Medicine & International Health. 2013; 18(9):1097–102.
12. Capeding MR, Tran NH, Hadinegoro SRS, Ismail HIHJM, Chotpitayasunondh T, Chua MN, et al. Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial. The Lancet. 2014; 384(9951):1358–65. https://doi.org/10.1016/s0140-6736(14)61060-6
13. Achee NL, Gould F, Perkins TA, Reiner RC Jr., Morrison AC, Ritchie SA, et al. A critical assessment of vector control for dengue prevention. PLoS Negl Trop Dis. 2015; 9(5):e0003655. https://doi.org/10.1371/journal.pntd.0003655 PMID: 25951103.
14. Halstead SB. Dengue vaccine development: a 75% solution? The Lancet. 2012; 380(9853):1535–6. https://doi.org/10.1016/S0140-6736(12)61510-4
15. Villar L, Dayan GH, Arredondo-García JL, Rivera DM, Cunha R, Deseda C, et al. Efficacy of a tetravalent dengue vaccine in children in Latin America. N Engl J Med. 2015; 372(2):113–23. https://doi.org/10.1056/NEJMoa1411037 PMID: 25365753.
16. Ooi EE. The re-emergence of dengue in China. BMC Med. 2015; 13:99. https://doi.org/10.1186/s12916-015-0345-0 PMID: 25925732.
17. Sang S, Yin W, Bi P, Zhang H, Wang C, Liu X, et al. Predicting local dengue transmission in Guangzhou, China, through the influence of imported cases, mosquito density and climate variability. PloS one. 2014; 9(7):e102755. https://doi.org/10.1371/journal.pone.0102755 PMID: 25019967.
18. Banu S, Hu W, Guo Y, Hurst C, Tong S. Projecting the impact of climate change on dengue transmission in Dhaka, Bangladesh. Environment International. 2014; 63(3):137–42.
19. Acharya BK, Cao C, Xu M, Khanal L, Naeem S, Pandit S. Present and Future of Dengue Fever in Nepal: Mapping Climatic Suitability by Ecological Niche Model. International Journal of Environmental Research & Public Health. 2018; 15(2):187.
20. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Present and Future of Dengue Fever in Nepal: Mapping Climatic Suitability by Ecological Niche Model. Nature. 2008; 457:1012.
21. Althouse BM, Ng YY, Cummings DA. Prediction of dengue incidence using search query surveillance. Plos Neglected Tropical Diseases. 2011; 5(8):e1258. https://doi.org/10.1371/journal.pntd.0001258 PMID: 21829744
22. Chan EH, Sahai V, Conrad C, Brownstein JS. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. PLoS neglected tropical diseases. 2011; 5(5):e1206. https://doi.org/10.1371/journal.pntd.0001206 PMID: 21647308.
23. Huang J, Hui Z, Jie Z, editors. Detecting Flu Transmission by Social Sensor in China. IEEE International Conference on Green Computing & Communications & IEEE Internet of Things & IEEE Cyber; 2013.
24. Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for influenza surveillance in South China. PLoS One. 2013; 8(1):e55205. https://doi.org/10.1371/journal.pone.0055205 PMID: 23372837.
25. Qingyu Y, Nsoesie EO, Benfu L, Geng P, Rumi C, Brownstein JS. Monitoring influenza epidemics in china with search query from baidu. PLoS One. 2013; 8(5):e64323. https://doi.org/10.1371/journal.pone.0064323 PMID: 23750192
26. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol [Internet]. 2014/4//; 10(4):[e1003581 p.]. https://doi.org/10.1371/journal.pcbi.1003581 PMID: 24743682
27. Yang W, Li Z, Lan Y, Wang J, Ma J, Jin L, et al. A nationwide web-based automated system for outbreak early detection and rapid response in China. Western Pacific Surveillance & Response Journal Wpsar. 2011; 2(1):10.
28. Halide H, Ridd P. A predictive model for Dengue Hemorrhagic Fever epidemics. Int J Environ Health Res. 2008; 18(4):253–65. https://doi.org/10.1080/09603120801966043 PMID: 18668414.
29. Sang S, Yin W, Bi P, Zhang H, Wang C, Liu X, et al. Predicting local dengue transmission in Guangzhou, China, through the influence of imported cases, mosquito density and climate variability. PLoS One [Internet]. 2014; 9(7):[e102755 p.]. https://doi.org/10.1371/journal.pone.0102755 PMID: 25019967

30. Li Z, Liu T, Zhu G, Lin H, Zhang Y, He J, et al. Dengue Baidu Search Index data can improve the prediction of local dengue epidemic: A case study in Guangzhou, China. Plos Negl Trop Dis. 2017; 11(3): e0005354. https://doi.org/10.1371/journal.pntd.0005354 PMID: 28263988

31. Ho CC, Ting CY. Time Series Analysis and Forecasting of Dengue Using Open Data. 2015.

32. Shen JC, Luo L, Li LI, Jing QL, Chun Quan OU, Yang ZC, et al. The Impacts of Mosquito Density and Meteorological Factors on Dengue Fever Epidemics in Guangzhou, China, 2006–2014: a Time-series Analysis. Biomedical & Environmental Sciences. 2015; 28(5):321–9.

33. Siregar FA, Makmur T, Saprin S. Forecasting dengue hemorrhagic fever cases using ARIMA model: a case study in Asahan district. IOP Conference Series: Materials Science and Engineering. 2018; 300:012032. https://doi.org/10.1088/1757-899x/300/1/012032

34. Louis VR, Phalkey R, Horstick O, Ratanawong P, Wilder-Smith A, Tozan Y, et al. Modeling tools for dengue risk mapping—a systematic review. International Journal of Health Geographics. 2014; 13(1):50.

35. Bouzid M, Colón-González FJ, Lung T, Lake IR, Hunter PR. Climate change and the emergence of vector-borne diseases in Europe: case study of dengue fever. BMC Public Health. 2014; 14(1):781. https://doi.org/10.1186/1471-2458-14-781 PMID: 25149418

36. Xu L, Stige LC, Chan K-S, Zhou J, Yang J, Sang S, et al. Climate variation drives dengue dynamics. Proceedings of the National Academy of Sciences. 2017; 114(1):113–8. https://doi.org/10.1073/pnas.1618558114 PMID: 27940911

37. Sang S, Gu S, Bi P, Yang W, Yang Z, Xu L, et al. Predicting unprecedented dengue outbreak using imported cases and climatic factors in Guangzhou, 2014. PLoS neglected tropical diseases. 2015; 9(5): e0003808. https://doi.org/10.1371/journal.pntd.0003808 PMID: 26020627

38. Shen JC, Luo L, Li L, Jing QL, Ou CQ, Yang ZC, et al. The Impacts of Mosquito Density and Meteorological Factors on Dengue Fever Epidemics in Guangzhou, China, 2006–2014: a Time-series Analysis. Biomedical and Environmental Sciences. 2015; 28(5):321–9. https://doi.org/10.3967/bes2015.046 PMID: 26055559

39. Sun J, Lin J, Yan J, Fan W, Lu L, Lv H, et al. Dengue virus serotype 3 subtype III, Zhejiang Province, China. Emerg Infect Dis. 2011; 17(2):321–3. https://doi.org/10.3201/eid1702.100396 PMID: 21291623.

40. Yuan Q, O Nsoesie E, Lv B, Peng G, Chunara R, Brownstein J. Detecting Flu Transmission by Social Sensor in China2013. e64323 p.

41. Gu Y, Chen F, Liu T, Lv X, Shao Z, Lin H, et al. Early detection of an epidemic erythromelalgia outbreak using Baidu search data. Sci Rep [Internet]. 2015; 5:[12649 p.]. https://doi.org/10.1038/srep12649 PMID: 26218589

42. Bao JX, Lv BF, Geng P, Na L, editors. Gonorrhea incidence forecasting research based on Baidu search data. International Conference on Management Science & Engineering; 2013.

43. Ying L, Lv B, Geng P, Yuan Q, editors. A preprocessing method of internet search data for prediction improvement: Application to Chinese stock market. Data Mining & Intelligent Knowledge Management Workshop; 2012.

44. Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. PLoS One. 2009; 4(2):e4378. https://doi.org/10.1371/journal.pone.0004378 PMID: 19197389.

45. Yang L, Qin G, Zhao N, Wang C, Song G. Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. Bmc Medical Research Methodology. 2012; 12(1):165. https://doi.org/10.1186/1471-2288-12-165 PMID: 23110601

46. Hoffmann AA, Montgomery BL, Popovici J, Iturbe-Ormaetxe I, Johnson PH, Muzzi F, et al. Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission. Nature [Internet]. 2011/8//; 476(7361):[454–7 pp.]. https://doi.org/10.1038/nature10356 PMID: 21866160

47. Walker T, Johnson PH, Moreira LA, Iturbe-Ormaetxe I, Frentiu FD, McMeniman CJ, et al. The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations. Nature [Internet]. 2011/8//; 476(7361):[450–3 pp.]. https://doi.org/10.1038/nature10355 PMID: 21866159

48. Harris AF, Nimmo D, McKemey AR, Kelly N, Scaife S, Donnelly CA, et al. Field performance of engineered male mosquitoes. Nat Biotechnol [Internet]. 2011/11//; 29(11):[1034–7 pp.]. https://doi.org/10.1038/nbt.2019 PMID: 22037376

49. Acharya BK, Cao C, Xu M, Chen W, Pandit S. Spatiotemporal Distribution and Geospatial Diffusion Patterns of 2013 Dengue Outbreak in Jhapa District, Nepal. Asia-Pacific journal of public health. 2018; 30(4):1010539518769809.

50. Focks, Dana A, Alexander, Neal, Villegas, Elci. Multicountry study of Aedes aegypti pupal productivity survey methodology: findings and recommendations. 2006.

51. Burattini MN, Chen M, Chow A, Coutinho FAB, Goh KT, Lopez LF, et al. Modelling the control strategies against dengue in Singapore. Epidemiol Infect. 2008; 136(3):309–19. https://doi.org/10.1017/S0950268807008667 PMID: 17540051.

52. Corwin AL, Larasati RP, Bangs MJ, Wuryadi S, Arjoso S, Sukri N, et al. Epidemic dengue transmission in southern Sumatra, Indonesia. Trans R Soc Trop Med Hyg. 2001; 95(3):257–65. https://doi.org/10.1016/s0035-9203(01)90229-9 PMID: 11490992.

53. Arcari P, Tapper N, Pfueller S. Regional variability in relationships between climate and dengue/DHF in Indonesia. Singapore Journal of Tropical Geography. 2010; 28(3):251–72.

54. Bangs MJ, Larasati RP, Corwin AL, Suharyono W. Climatic factors associated with epidemic dengue in Palembang, Indonesia: implications of short-term meteorological events on virus transmission. Southeast Asian J Trop Med Public Health. 2006; 37(6):1103–16. PMID: 17333762

55. Barrera R, Delgado N, Jiménez M, Villalobos I, Romero I. Estratificación de una ciudad hiperendémica en dengue hemorrágico. Revista Panamericana De Salud Pública. 2000; 8(4):225–33.

56. Chadee DD, Shivnauth B, Rawlins SC, Chen AA. Climate, mosquito indices and the epidemiology of dengue fever in Trinidad (2002–2004). Ann Trop Med Parasitol. 2007; 101(1):69–77. https://doi.org/10.1179/136485907X157059 PMID: 17244411.

57. Yang HM, Macoris MLG, Galvani KC, Andrighetti MTM, Wanderley DMV. Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue. Epidemiol Infect. 2009; 137(8):1188–202. https://doi.org/10.1017/S0950268809002040 PMID: 19192322.

58. Mariangela B, Giuliano G, Xioaguang C, James AA. The invasive mosquito species Aedes albopictus: current knowledge and future perspectives. Trends in Parasitology. 2013; 29(9):460–8. https://doi.org/10.1016/j.pt.2013.07.003 PMID: 23916878

59. Su T, Mulla MS. Effects of temperature on development, mortality, mating and blood feeding behavior of Culiseta incidens (Diptera: Culicidae). Journal of Vector Ecology. 2001; 26(1):83–92. PMID: 11469189

60. Githeko AK, Lindsay SW, Confalonieri UE, Patz JA. Climate change and vector-borne diseases: a regional analysis. Bull World Health Organ. 2000; 78(9):1136–47. PMID: 11019462

61. Sang S, Gu S, Bi P, Yang W, Yang Z, Xu L, et al. Predicting unprecedented dengue outbreak using imported cases and climatic factors in Guangzhou, 2014. PLoS neglected tropical diseases [Internet]. 2015/5//; 9(5):[e0003808 p.]. https://doi.org/10.1371/journal.pntd.0003808 PMID: 26020627

62. Chen S-C, Liao C-M, Chio C-P, Chou H-H, You S-H, Cheng Y-H. Lagged temperature effect with mosquito transmission potential explains dengue variability in southern Taiwan: insights from a statistical analysis. Sci Total Environ. 2010; 408(19):4069–75. https://doi.org/10.1016/j.scitotenv.2010.05.021 PMID: 20542536.

63. Naish S, Dale P, Mackenzie JS, McBride J, Mengersen K, Tong S. Climate change and dengue: a critical and systematic review of quantitative modelling approaches. BMC Infect Dis [Internet]. 2014; 14:[167 p.]. https://doi.org/10.1186/1471-2334-14-167 PMID: 24669859

64. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza forecasting with Google Flu Trends. PLoS One. 2013; 5(1):e56176.

65. Milinovich GJ, Williams GM, Clements AC, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infectious Diseases. 2014; 14(2):160–8. https://doi.org/10.1016/S1473-3099(13)70244-5 PMID: 24290841