Science of the Total Environment 647 (2019) 752–762

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

An ensemble forecast model of dengue in Guangzhou, China using climate and social media surveillance data

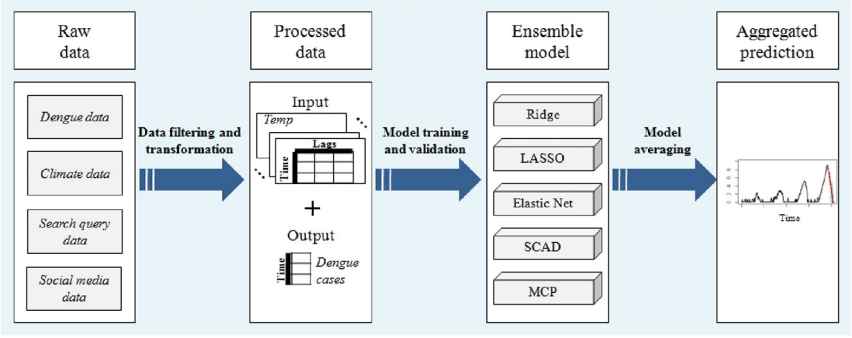
Pi Guo [a], Qin Zhang [b], Yuliang Chen [a], Jianpeng Xiao [c], Jianfeng He [d], Yonghui Zhang [d], Li Wang [e,*], Tao Liu [c,*], Wenjun Ma [c,**]

[a] Department of Preventive Medicine, Shantou University Medical College, No. 22 Xinling Road, Shantou 515041, China
[b] Good Clinical Practice Office, Cancer Hospital of Shantou University Medical College, Shantou 515041, China
[c] Guangdong Provincial Institute of Public Health, Guangdong Provincial Center for Disease Control and Prevention, Guangzhou 511430, China
[d] Guangdong Provincial Center for Disease Control and Prevention, Guangzhou 511430, China
[e] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Avenue, Shenzhen University Town, Shenzhen 518055, China

HIGHLIGHTS

- A novel ensemble model has been developed for timely tracking the timing and magnitude of dengue epidemics.
- The proposed algorithm had the best performance in comparison with regression models using single penalties.
- The ensemble forecast model provides skillful forecasts of the dynamics of dengue in China.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 15 February 2018
Received in revised form 8 July 2018
Accepted 3 August 2018
Available online 04 August 2018

Editor: SCOTT SHERIDAN

Keywords:
Dengue
Ensemble model
Forecast
Outbreak
Real time

ABSTRACT

Background: China experienced an unprecedented outbreak of dengue in 2014, and the number of dengue cases reached the highest level over the past 25 years. There is a significant delay in the release of official case count data, and our ability to timely track the timing and magnitude of local outbreaks of dengue remains limited.
Material and methods: We developed an ensemble penalized regression algorithm (EPRA) for initializing near-real time forecasts of the dengue epidemic trajectory by integrating different penalties (LASSO, Ridge, Elastic Net, SCAD and MCP) with the techniques of iteratively sampling and model averaging. Multiple streams of near-real time data including dengue-related Baidu searches, Sina Weibo posts, and climatic conditions with historical dengue incidence were used. We compared the predictive power of the EPRA with the alternates, penalized regression models using single penalties, to retrospectively forecast weekly dengue incidence and detect outbreak occurrence defined using different cutoffs, during the periods of 2011–2016 in Guangzhou, south China.
Results: The EPRA showed the best or at least comparable performance for 1-, 2-week ahead out-of-sample and leave-one-out cross validation forecasts. The findings indicate that skillful near-real time forecasts of dengue and confidence in those predictions can be made. For detecting dengue outbreaks, the EPRA predicted periods of high incidence of dengue more accurately than the alternates.
Conclusion: This study developed a statistically rigorous approach for near-real time forecast of dengue in China. The EPRA provides skillful forecasts and can be used as timely and complementary ways to assess dengue dynamics, which will help to design interventions to mitigate dengue transmission.

* Corresponding authors.
** Correspondence to: W. Ma, Guangdong Provincial Institute of Public Health, Guangdong Provincial Center for Disease Control and Prevention, Guangzhou 511430, China.
E-mail address: mawj@gdiph.org.cn (W. Ma).

https://doi.org/10.1016/j.scitotenv.2018.08.044
0048-9697/

1. Introduction

Dengue is a serious infectious disease and the number of infections has sharply increased over the past 50 years, resulting in a huge impact on population health worldwide (Diseases PFIT and WH Organization, 2009). A recent study about the global distribution and disease burden of dengue showed that the number of dengue infections is estimated to be 390 million per year, of which around 96 million are symptomatic (Bhatt et al., 2013). In China, dengue is of special concern since the areas affected by the disease have expanded and the incidence is steadily rising over the recent years (Lai et al., 2015; Chen and Liu, 2015). It is worthy to note that China experienced an unprecedented outbreak of dengue in 2014, and the number of dengue cases reached the highest level over the past 25 years (Li et al., 2017; Xiao et al., 2016).

Although the effects of climatic conditions such as rainfall and temperature on mosquito abundance and dengue transmission rate were suggested in temperate regions in China (Xu et al., 2017), but at present dengue is still characterized as an imported disease in the country due to localized transmission sparked by regular virus importations from returned travelers or visitors (Lai et al., 2015). Therefore, our ability to predict these dengue outbreaks by relying solely on the information of climatic conditions is insufficient. Indeed, in the absence of an effective vaccine against dengue in China, much public health benefit can be gleaned from early and skillful forecasts of local dengue outbreaks, which will enable the government to take effective control measures and minimize the menace.

In China, existed outbreak detection system using the China Infectious Disease Automated-alert and Response System (CIDARS) is overly dependent on the number of the routinely reported dengue cases, and the delay from receipt to dissemination of information of the notified dengue cases to the public is 1- to 2-week on average (Zhang et al., 2014; Yuan et al., 2013). To counter the delay of traditional surveillance, several studies proposed to use online informal sources such as news reports (Freifeld et al., 2008), social media data (Chew and Eysenbach, 2010; Polgreen, 2009; Ye et al., 2016), and search query data (Eysenbach, 2006) to achieve a near-real time surveillance of the spread of diseases. But for a formal statistical approach to the near-real time estimation of the epidemic trajectory of infectious diseases, Ginsberg, J. et al. first used large numbers of Google search query data to monitor influenza-like illness in the United States in 2009 (Ginsberg et al., 2009). However, many details remain unclear about the implementation of this model. In our previous study, we also found dengue-related Baidu search queries can improve the prediction of local dengue epidemics (Li et al., 2017). Fortunately, we can obtain near-real time online information including dengue-related internet search queries and social media data in China at present (Index, 2017; Weibo, 2017). Internet search queries from the Baidu Index database (http://index.baidu.com/) and social media data from the Sina Weibo platform (http://www.weibo.com/) can be available on a daily basis and at a city level since January 1, 2011 in China. Currently in China, the Baidu is the most popular search engine, and Sina Weibo is a microblogging website with the largest user group which is regarded as an equivalent version of USA Twitter, respectively. These population-level digital data make the application of near-real time forecast practicable.

In fact, our previous study has developed a method of variable selection using an iterative sampling technique to identify significant predictors and estimate the parameters in the model with low variability (Guo et al., 2015). Then we have further developed an internet search terms-based model of influenza using Baidu search queries within an ensemble prediction framework (Pi et al., 2017). On this basis, to track dengue dynamics with near-real time forecasts, this present study aimed to integrated multi-sources of data related to dengue dynamics, and established a novel ensemble forecast framework to fill in the gap in China.

This present study integrated different penalty functions with the techniques of iterative sampling and model averaging called Bagging (Breiman, 1996) to establish a novel ensemble penalized regression algorithm (EPRA) to timely estimate the timing and magnitude of local outbreaks of dengue. To get a forecast, a 100-member ensemble of penalized regression models is numerically integrated and calibrated for predictions of dengue epidemic progression in Guangzhou, the most densely-populated city in south China. Each ensemble prediction is initialized with a different randomly drawn sample of original data to train a penalized regression model with a kind of penalty, and incrementally adds more different kinds of penalized models to improve the ability of the EPRA to forecast the future trajectory of dengue epidemic by adjusting the model parameters. The suite of predictors and parameters in each penalized regression model was tuned using a cross-validation (CV) technique. In addition, we believe that a forecast model should be optimized under several performance measures simultaneously in order to provide more robust predictions of the future unfolding trajectory of a local dengue outbreak. To achieve this goal, a multi-objective optimization algorithm (Pihur et al., 2007) was used to optimize the performance of the EPRA. This study reported an attempt to establish such a prediction model of dengue with favorable performance.

2. Material and methods

2.1. Data sources

2.1.1. Dengue case data

Guangzhou, is the capital city of Guangdong province and the most densely-populated city in south China. Guangzhou city was selected as the study site since its infection rate of dengue was historically high and emerging as a continuous threat to the local people, especially it had a large outbreak of dengue in 2014 (Xiao et al., 2016). Guangzhou is the center of transportation, finance, industry and trade in south China and has frequent economic and cultural communication with the nations of Southeast Asia where dengue poses a great burden of disease. It has 12 districts with an area of 7473 km^2 and a typical subtropical humid monsoon climate, with an annual temperature of 22 °C on average (Li et al., 2017). Weekly dengue case data of Guangzhou from 1 January 2011 to 31 December 2016 were available from the Guangdong Provincial CDC. Both of the imported and indigenous dengue cases were notified according to the Guangdong Provincial CDC, and were included in this work. Related information of each case including basic demographic characteristics (gender, age, nationality and residential address) and times of disease-related incidents (date of illness onset, diagnosis and death) were recorded. Clinical diagnosed and laboratory confirmed cases were aggregated and included in our analysis. All dengue cases were diagnosed according to the diagnostic criteria for dengue fever (WS216–2008) enacted by the Chinese Ministry of Health (Jing et al., 2012).

2.1.2. Climate data

Climate data including weekly mean temperature (°C), maximum temperature (°C), minimum temperature (°C), average relative humidity and rainfall (millimeters) over the study period from the China Meteorological Data Sharing Service System (http://cdc.nmic.cn/home.do) were used for the analysis.

2.1.3. Search query surveillance data

Baidu is now the most popular search engine in China, making it the most representative data source for tracking online behavior of people seeking information. Baidu releases near-real time search query surveillance of keywords typed by users in the Baidu Index database (https://index.baidu.com/). The database contains logs of online search query volume for numerous keywords. Some previous studies proposed to choose the names or clinical symptoms of the studied diseases as the primary terms to search for more related keywords, which were usually obtained from a Chinese website (http://tool.chinaz.com/baidu/words.aspx) (Weibo, 2017; Min et al., 2013). Upon typing in 10 primary search terms, we obtained a group of related keywords (Table S1). We designed a crawler software using Python to automatically collect weekly data of search query surveillance in Guangzhou over the study period.