Acta Tropica 185 (2018) 391–399

Contents lists available at ScienceDirect

Acta Tropica

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

Mapping the transmission risk of Zika virus using machine learning models

Dong Jiang[a,b], Mengmeng Hao[a,b], Fangyu Ding[a,b,*], Jingying Fu[a,b], Meng Li[a,b]

[a] *State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China*

[b] *College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, 100049, China*

ARTICLE INFO

Keywords:
Zika virus
Transmission risk
Machine learning
Significant differences
Prediction uncertainty

ABSTRACT

Zika virus, which has been linked to severe congenital abnormalities, is exacerbating global public health problems with its rapid transnational expansion fueled by increased global travel and trade. Suitability mapping of the transmission risk of Zika virus is essential for drafting public health plans and disease control strategies, which are especially important in areas where medical resources are relatively scarce. Predicting the risk of Zika virus outbreak has been studied in recent years, but the published literature rarely includes multiple model comparisons or predictive uncertainty analysis. Here, three relatively popular machine learning models including backward propagation neural network (BPNN), gradient boosting machine (GBM) and random forest (RF) were adopted to map the probability of Zika epidemic outbreak at the global level, pairing high-dimensional multidisciplinary covariate layers with comprehensive location data on recorded Zika virus infection in humans. The results show that the predicted high-risk areas for Zika transmission are concentrated in four regions: Southeastern North America, Eastern South America, Central Africa and Eastern Asia. To evaluate the performance of machine learning models, the 50 modeling processes were conducted based on a training dataset. The BPNN model obtained the highest predictive accuracy with a 10-fold cross-validation area under the curve (AUC) of 0.966 [95% confidence interval (CI) 0.965-0.967], followed by the GBM model (10-fold cross-validation AUC = 0.964[0.963-0.965]) and the RF model (10-fold cross-validation AUC = 0.963[0.962-0.964]). Based on training samples, compared with the BPNN-based model, we find that significant differences (p = 0.0258* and p = 0.0001***, respectively) are observed for prediction accuracies achieved by the GBM and RF models. Importantly, the prediction uncertainty introduced by the selection of absence data was quantified and could provide more accurate fundamental and scientific information for further study on disease transmission prediction and risk assessment.

1. Introduction

Zika virus (ZIKV), a mosquito-borne virus, is a member of the family Flaviviridae, genus Flavivirus (Wikan and Smith, 2016), and has aroused wide attention in the international community in recent years (World Health Organization, 2016; Messina et al., 2016a). ZIKV can spread by sexual transmission (Mansuy et al., 2016a, b), but the main route of transmission is similar to that of Dengue virus (DENV), which is mosquito borne (Gardner et al., 2016; Medlock et al., 2017; Liu et al., 2017). In most cases, people infected with ZIKV tend to have mild fever, rashes, arthralgia and conjunctivitis symptoms (Pacheco et al., 2016; Campos et al., 2015); in severe cases, ZIKV virus can lead to neurological complications, such as Guillain–Barré syndrome (GBS) (Parra et al., 2016). Additionally, ZIKV infection during pregnancy has been associated with microcephaly of newborn babies, and several researchers have illustrated the relationship that ZIKV disrupts neural progenitor development (Li et al., 2016; Cordeiro et al., 2016).

ZIKV was first identified from the serum of a rhesus monkey in the Zika Forest of Uganda in 1947, and the second isolation was made from wild-caught *Aedes* africanus in 1948 in the same forest (Dick et al., 1952; Dick, 1952). The first reported case of human infection by ZIKV occurred in 1952 in Eastern Nigeria (Macnamara, 1954), but several scientists showed that the virus was initially misidentified as ZIKV and subsequently identified as close to Spondweni virus (Boorman and Draper, 1968; Simpson, 1964). Before 2007, only 13 natural ZIKV infections in humans were documented, sporadically, in Africa and Asia (Faye et al., 2014; Wikan and Smith, 2017). From 2007 to 2016, ZIKV caused several outbreaks of human infection (Wikan and Smith, 2016). The first reported ZIKV outbreak occurred in 2007 in Yap State, which belongs to the Federated States of Micronesia in the Western Pacific

* Corresponding author at: State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China.

E-mail addresses: jiangd@igsnrr.ac.cn (D. Jiang), haomm.16b@igsnrr.ac.cn (M. Hao), dingfy.17b@igsnrr.ac.cn (F. Ding), fujy@igsnrr.ac.cn (J. Fu), lim.17b@igsnrr.ac.cn (M. Li).

https://doi.org/10.1016/j.actatropica.2018.06.021
Received 22 March 2018; Received in revised form 11 June 2018; Accepted 18 June 2018
Available online 19 June 2018

(Duffy et al., 2009). During this outbreak, 49 confirmed and 59 suspected cases of ZIKV infection were identified based on genetic and serological analysis, which represented expansion of ZIKV transmission outside Africa and Asia (Lanciotti et al., 2008). The following larger ZIKV epidemic occurred in French Polynesia in the South Pacific in 2013-14 (Hancock et al., 2014). Based on serological and surveillance data, 8750 probable infections with ZIKV were recorded, of which 383 were confirmed from October 2013 to April 2014 (Cauchemez et al., 2016). Subsequently, a small-scale ZIKV epidemic outbreak in the Cook Islands, Easter Island and the Solomon Islands was concentrated in the Pacific region (Musso et al., 2014; Roth et al., 2014; Musso and Gubler, 2016). ZIKV emerged in Brazil for the first time in 2015, with an estimated 440000–1300000 cases of infection in humans (Yakob and Walker, 2016) and spread rapidly to more than 20 countries or territories by the end of January 2016 (Enfissi et al., 2016). Sequence analyses showed that ZIKV strains isolated in Brazil showed a high degree of homology to sequences isolated in French Polynesia (Musso and Gubler, 2016; Zanluca et al., 2015), suggesting that the ZIKV found in the Pacific might have spread across Oceania and contributed to the epidemic in Latin America. With continuous geographic expansion, ZIKV is increasingly threatening global public health, especially that of pregnant women (Song et al., 2017).

Suitability mapping the transmission risk of ZIKV is essential, so that proper detection (i.e. RT-PCR) and prevention interventions (i.e. mosquito control) can be employed in the locations that likely have the greatest risk (Yakob and Walker, 2016; Waddell and Greig, 2016; Chan et al., 2017; Benelli and Mehlhorn, 2016). Based on surveillance cases of ZIKV data, Rodriguez-Morales et al. used geographical information systems (GIS) to estimate cumulative incidence rates at several departments in Colombia, guiding decision making for ZIKV prevention (Rodriguezmorales et al., 2017; Rodriguez-Morales et al., 2017, 2016). From a global viewpoint, Attaway et al. combined the predictive analysis method with five environmentally associated layers to create a ZIKV risk map, which shows that ZIKV probably occurs in many tropical regions all year round (Attaway et al., 2017). Combining known records of ZIKV occurrence with high spatial-resolution environmental datasets, Messina et al. used an ensemble boosted regression tress (BRT) model to produce the global risk map of ZIKV with an area under the curve (AUC) value of 0.829, illustrating that over 2.17 billion people inhabit high-risk regions (Messina et al., 2016a). These studies were primarily based on case-monitoring data, environmental variables and the demographic distribution of *Aedes* mosquitoes. However, the impacts of socioeconomic factors reflecting human movements and urbanization on the transmission of ZIKV were neglected. Additionally, the uncertainty of ZIKV prediction risk was also overlooked.

Given the availability of data, we collected comprehensive ZIKV occurrence records and multidisciplinary datasets (Table 1), including simulated areas of *Aedes* mosquitoes and meteorological, environmental and social-economic layers. With the high-dimensional datasets, three relatively popular machine learning models are adopted to map the global transmission risk of ZIKV at high spatial resolution. Further, we analyze the significant differences between prediction accuracies of these models for the first time. It is worth noting that the prediction uncertainty of the models is quantified, aiming at selecting more accurate models for identifying the areas susceptible to ZIKV outbreak.

2. Materials and methods

Several previous studies have shown some complex non-linear relationships between transmission risk of ZIKV and multiple variables (Messina et al., 2016a; Santos and Meneses, 2017). To handle these relationships efficiently, three robust machine learning models were adopted in the present research, including backward propagation neural network (BPNN), gradient boosting machine (GBM) and random forest (RF). To simulate the probability of ZIKV infection, the following datasets were required by these models: (a) a set of high-resolution global variables described to influence the transmission of ZIKV; (b) a set of georeferenced datasets for known ZIKV cases in humans; and (c) a set of absence points that represent inappropriate conditions for ZIKV transmission. All datasets used in this study were converted into a unified coordinate system, i.e., WGS-84. Open source GIS software (i.e., QGIS http://www.qgis.org/) and Python extension packages (i.e., GDAL http://www.gdal.org/) were employed to quantify the areas susceptible to ZIKV outbreak at the global scale with 5 × 5 km spatial resolution. The technical flow chart of the study is shown in Fig. 1.

Table 1
Multidisciplinary datasets used to map the global transmission risk of ZIKV.

Factors	Scale	Format	Data sources
Min Temperature	1 × 1 km	Grid	WorldClim database, version 2.0
Max Temperature	1 × 1 km	Grid	
Precipitation	1 × 1 km	Grid	
NDVI	8 × 8 km	Grid	Global Inventory Modeling and Mapping Studies Group
Relative humidity	—	Shapefile	Surface Meteorology and Solar Energy, NASA
Urbanicity	—	Shapefile	Socioeconomic Data and Applications Center, NASA
Nighttime Lights	1 × 1 km	Grid	The Earth Observation Group, NOAA
Urban Accessibility	1 × 1 km	Grid	European Commission Joint Research Center Global Environment Monitoring Unit
Predicted distribution of *Ae. aegypti* and *Ae. albopictus*	5 × 5 km	Grid	(Ding et al., 2018)

2.1. Data acquisition

2.1.1. Climatic factors

Previous literature has shown a close relationship between temperature and ZIKV transmission (Santos and Meneses, 2017). ZIKV transmission depends on the distribution of *Aedes* mosquito vectors, while temperature affects several key physiological processes in these vectors, such as adult female survivorship and length of the first gonotrophic cycle (Brady et al., 2014). Precipitation also plays an important role in mosquito growth. For example, areas with greater amounts of precipitation are generally associated with *Aedes* mosquito abundance (Romero-Vivas and Falconar, 2005; Scott et al., 2000). Thus, three climate factors were used in the machine learning models, namely, minimum annual temperature, maximum annual temperature and annual cumulative precipitation. From the WorldClim database version 2.0 (http://www.wclim.org/), we obtained these global climate datasets with a 1 × 1 km spatial resolution, which were derived from monitoring data from world-wide meteorological stations based on ANUSPLIN-SPLINA software (Hijmans et al., 2005).

2.1.2. Environmental factors

The relationship between environmental conditions (i.e., relative humidity and NDVI) and DENV propagation has been thoroughly addressed in several studies (Colóngonzález et al., 2011; Bhatt et al., 2013). For example, the mosquito habitats could be protected from direct sunlight by the surrounding plants (Fuller et al., 2009), and greater relative humidity has also been found to promote DENV transmission (Brady et al., 2014). Considering the similarity of ZIKV and DENV, we assumed relative humidity and NDVI layers as potential environmental factors for ZIKV propagation. In the present research, we used mean annual NDVI and mean annual relative humidity dataset as two input data layers for machine learning models. The mean annual NDVI layer was derived from an advanced very high-resolution radiometer (AVHRR) NDVI dataset with an 8 × 8 km spatial resolution and a 15-day interval temporal resolution, which was developed by the Global Inventory Modeling and Mapping Studies (GIMMS) group

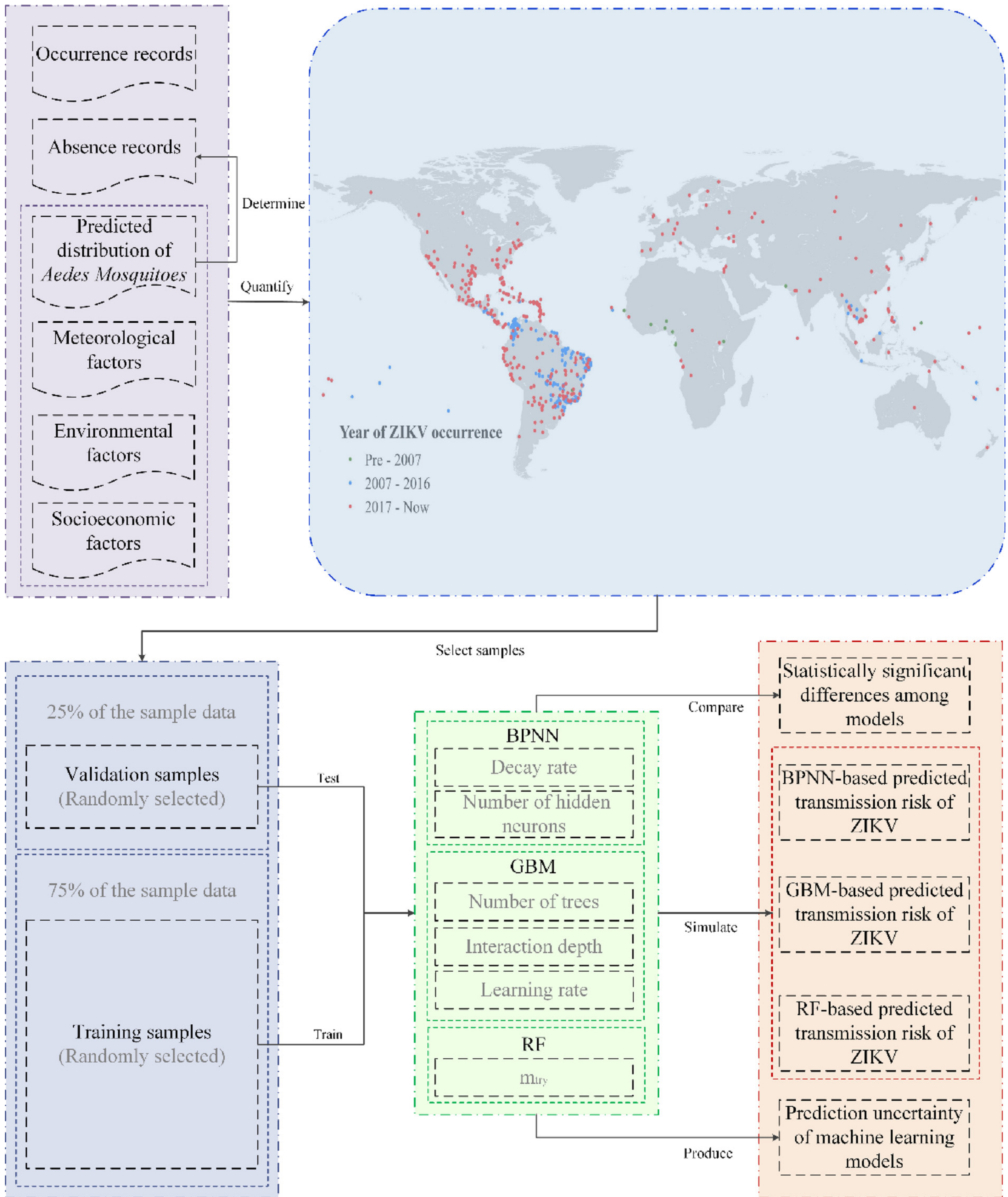


Fig. 1. Technical flow chart.

(http://glcf.umd.edu/) and acquired globally for years 1982 through 2015. From the NASA Surface Meteorology and Solar Energy (https://eosweb.larc.nasa.gov/), we obtained the global mean annual relative humidity dataset converted from a shapefile to raster layer.

2.1.3. Socioeconomic factors

There is evidence that some vector-borne diseases and *Aedes* mosquitoes are closely related to human habitat (Brown et al., 2011; Powell and Tabachnick, 2013; Medlock and Leach, 2015). Compared with natural rain-filled containers, artificial containers could provide more suitable habitats for the larval development of these mosquitoes (i.e., *Aedes aegypti*), which increases the risk of exposure to vector-borne pathogens (Ladeau et al., 2015; Morrison et al., 2004). To more accurately represent the geographic variation in human habitat, global urban areas and nighttime light layers were used in the present research. From the NASA Socioeconomic Data and Application Center (SEDAC) (http://sedac.ciesin.columbia.edu/) website, we downloaded the Global Urban Heat Island dataset (Center for International Earth Science Information Network - CIESIN - Columbia University, 2016) and extracted the global urban region distribution layer from the former. The global mean annual nighttime light layer was derived from stable light layers of nighttime light satellite imagery acquired from

Fig. 2. The AUC values of 50 simulation results, based on machine learning models with optimal parameters. (A) Accuracy evaluation based on training samples. (B) Accuracy evaluation based on validation samples.

1992 to 2013, which can be obtained from the NOAA Earth Observation Group (https://ngdc.noaa.gov/).

With the rapid increase in connectivity among the human population, close associations have been shown among international travel, trade routes, and ZIKV expansion (Kyeongah et al., 2016; Bogoch et al., 2016). The worldwide spread of ZIKV pathogens has become increasingly serious, but the data reflecting human movement on a global scale are always difficult to obtain. In this research, the urban accessibility dataset defining the travel time to the nearest city with a population of 50,000 people or more was used to represent possible human movement, considering that these data can be downloaded freely from the European Commission Joint Research Center website (http://forobs.jrc.ec.europa.eu/), and has also been found to be an important factor when predicting DENV and H7N9 transmission risk (Bhatt et al., 2013; Gilbert et al., 2014).

2.2. Sampling strategy

In the present research, we collected the known global occurrences of ZIKV in humans, which consisted of two parts. The first part of the dataset (https://doi.org/10.6084/m9.figshare.2573629.v1), compiled by Messina et al. through literature reviews and Google Maps searches (Jane and Freya, 2016), was downloaded from the Figshare website (https://figshare.com/), which has 326 occurrence records for ZIKV infection in humans for the years 1951-2016. The second part of the dataset was collected from informal online data sources, named HealthMap (http://www.healthmap.org/), which reported 1433 cases of human ZIKV infection from January to July 2017. It should be noted that the data quality of the former is much better than that of the latter. For the second part dataset, ZIKV infection cases reported outside of the region of predicted presence of *Aedes* mosquitoes were removed to improve the quality of data. In this research, we assumed that the regions where ZIKV infections have occurred in the past reflect appropriate conditions for ZIKV transmission. To match the spatial resolution of covariate datasets, we transformed the occurrence records from table to grid, at 5 × 5 km, based on the location information of these records. Finally, 612 unique occurrences were produced to reflect high-risk regions for ZIKV transmission.

To use machine learning models, we also need the absence records of ZIKV that represent where ZIKV is less likely to be transmitted to humans. However, it is difficult to obtain the actual absence records of ZIKV. According to previous studies, pseudo-absence records can be used instead of real absence records (Phillips et al., 2009). Additionally, it has been demonstrated that the areas where *Aedes* mosquitoes do not occur are currently in a low-risk state of ZIKV pathogen outbreak (Yakob and Walker, 2016; Santos and Meneses, 2017). Therefore, the optimally simulated distribution layers of *Ae. aegypti* and *Ae. albopictus* derived from our previous research (Ding et al., 2018) were selected as a limiting factor to ZIKV propagation and used to produce pseudo-absence records of ZIKV. In total, we randomly selected 612 pseudo-absence samples for ZIKV, which was an amount equal to the occurrence samples.

Table 2
Statistical comparison of prediction accuracy among different models.

Datasets	BPNN-GBM	BPNN-RF	GBM-RF
Based on training samples	0.0258 *	0.0001 ***	0.0768 NS
Based on validation samples	0.1620 NS	0.3258 NS	0.9922 NS

Note: * indicates $p < 0.05$; ** indicates $p < 0.01$; *** indicates $p < 0.001$; NS indicates not significant.

The performance of machine learning models could be affected by the samples, so we therefore repeated the process of selecting pseudo-absence samples 50 times. Based on the above data processing methods, we obtained 1224 sample data for ZIKV at a time. During each modeling process, 75% of the sample data were randomly selected as training data, and the remaining 25% were used as validation data. As the multidisciplinary datasets described previously have different units, we normalized them to facilitate the training step of the models using the following Eq. (1):

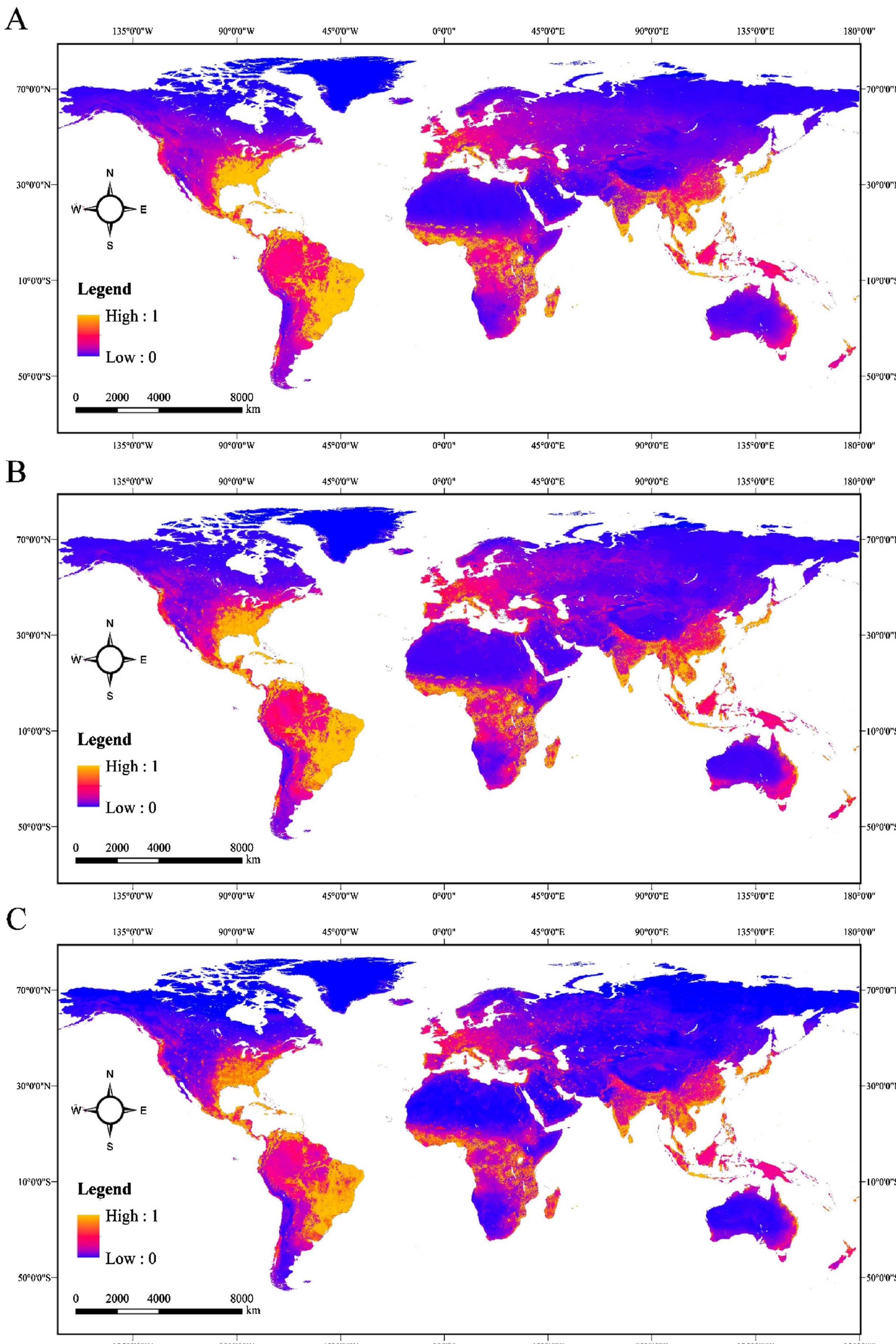
$$X_j' = \frac{X_j - X_{\min}}{X_{\max} - X_{\min}} (j = 1, 2, 3..., n) \quad (1)$$

where X_j' is the normalized value between 0 and 1; X_j is the value of the factor; $X_{\min}$ is the minimum value; $X_{\max}$ is the maximum value; and n is the number of data.

2.3. Modeling

We adopted version 3.3.3 of the 64-bit version of R language for mapping transmission risk of ZIKV, which is open-source software used to build models, tune parameters and assess accuracy. The caret package was employed to develop BPNN, GBM and RF models because the package provided a consistent environment for training machine learning models and tuning their associated parameters. The 10-fold cross-validation method was applied to train machine learning models, and the area under the curve (AUC) was used to evaluate the predictive performance of these models.

For the BPNN model, the number of hidden neurons and the decay rate have a greater effect on the model performance. Increasing the former makes the model more suitable for a particular task domain, which is believed to have influence on the model accuracy, whereas increasing the latter parameter affects convergence rate, which may also result in non-convergence (Rumelhart et al., 1988). Therefore, we needed to fine-tune the value of the decay rate and the number of hidden units when training a BPNN model with sigmoid activation function. For simulation built with a GBM model, the number of trees, the interaction depth value and the learning rate have been shown to influence the predictive performance of the model (Ridgeway, 2006). In the present research, the three model tuning parameters for the GBM model needed to be fine-tuned when using GBM model to simulate transmission risk of ZIKV. For RF-based modeling process, the default value (500) is usually adopted since values larger than default cannot significantly improve model performance (Liaw and Wiener, 2002). The only adjustable tuning parameter, the m_{try} parameter, influenced the model accuracy because it controlled the number of variables randomly



(caption on next page)

Fig. 3. Maps of the global transmission risk for ZIKV at 5 × 5 km spatial resolution based on machine learning models: (A) BPNN-based prediction; (B) GBM-based prediction; (C) RF-based prediction.

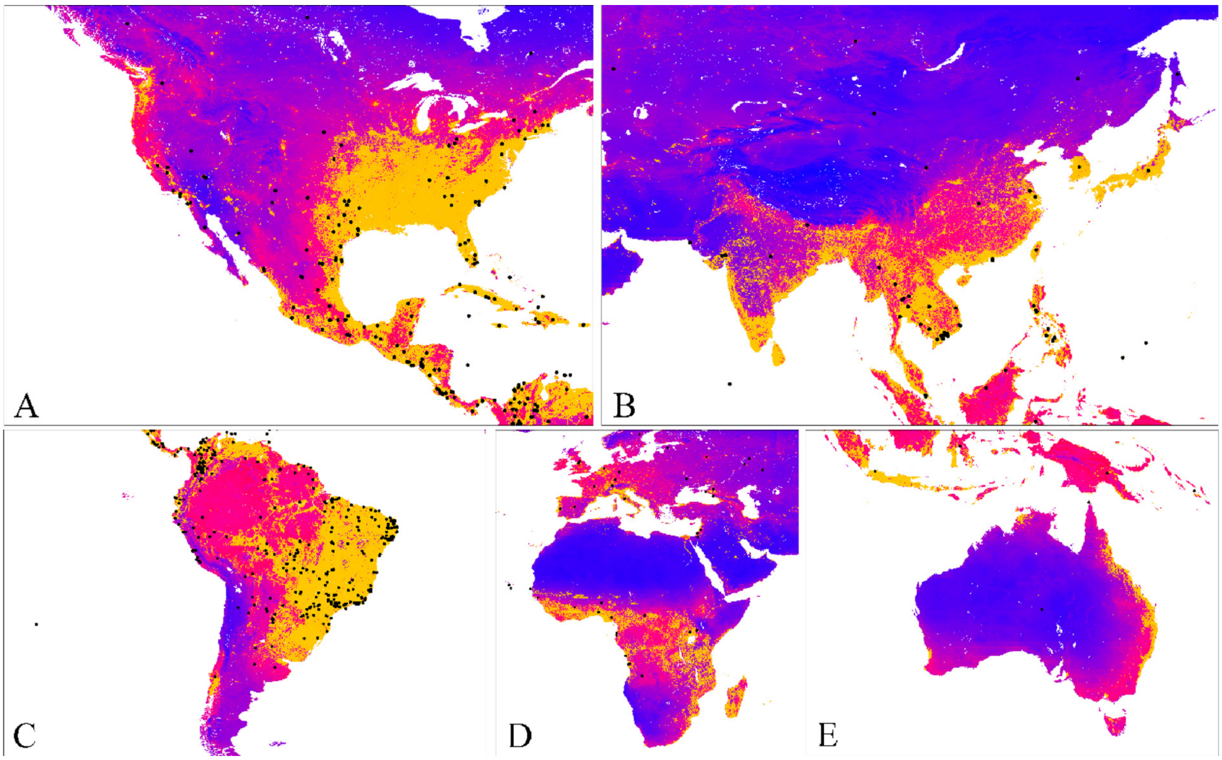


Fig. 4. The spatial distribution of 612 unique ZIKV occurrence units: (A) North America; (B) Asia; (C) South America; (D) Africa and Europe; (E) Oceania.

sampled, as candidates at each split.

3. Results

3.1. Parameter fine-tuning and accuracy comparison

Several decay rate values (0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6 and 0.8) and the number of hidden neurons (10, 15, 20, 25, 30 and 35) were examined for the prediction of ZIKV transmission risk based on the BPNN model. For the modeling process using the GBM model, a total of 8 values for interaction depth parameter (3, 4, 5, 6, 7 and 8), 6 values for number of trees (50, 100, 150, 200, 250, 300, 350 and 400) and 4 values for learning rate (0.01, 0.05, 0.1 and 0.2) were examined. When we used the RF model to simulate the global transmission risk of ZIKV, a total of 10 values for the m_{try} parameter (1, 2, 3, 4, 5, 6, 7, 8, 9 and 10) were examined. In the present research, we employed the BPNN, GBM and RF models with optimized tuning parameter values to simulate transmission risk, respectively, in each modeling process.

Fig. 2 shows the performance of three machine learning models applied to the training dataset and validation dataset in the 50 simulations process. Based on the training dataset, we found that the BPNN models obtained the highest predictive accuracy (10-fold cross-validation AUC = 0.966 [95% CI 0.965-0.967]), followed by the GBM model (10-fold cross-validation AUC = 0.964[0.963-0.965]) and the RF model (10-fold cross-validation AUC = 0.963[0.962-0.964]) (Fig. 2A). When we used validation samples to evaluate the performance of machine learning models, the same trends were observed (Fig. 2B): the BPNN model provided higher AUC (0.966 [0.963-0.968]) values than GBM (0.963[0.960-0.966]) and RF (0.963[0.960-0.966]). Generally, those goodness of fit metrics indicated that three machine learning models not only achieved high performance of prediction close to the training samples but also provided insight to other geographical areas such as validation areas.

The *t*-test was adopted to compare significant differences between predictive accuracies among different models. It should be noted that the F-test was employed before using *t*-test, which illustrated that the variance differences between prediction accuracies of these machine learning models were not significant at the 5% level. Table 2 revealed that compared to BPNN-based prediction accuracies, significant differences ($p = 0.0258^*$ and $p = 0.0001^{***}$, respectively) were observed for prediction accuracies achieved by the GBM and RF models applied to training samples. Additionally, the accuracy difference between the GBM and RF models was not significant ($p = 0.0768$) when using training samples to evaluate accuracy. Nonetheless, the differences between predictive accuracies among different models were not significant ($p > 0.05$) when evaluating accuracy based on validation samples.

3.2. Simulated global transmission risk of ZIKV

Fig. 3 depicts the global transmission risk maps for ZIKV from 0 (Low risk) to 1 (High risk) based on three examined machine learning models, which are the average risk distribution maps of 50 prediction results. In general, these maps show a reasonably accurate visual depiction of the areas susceptible to ZIKV outbreak, which are similar to each other. The predicted high-risk areas for ZIKV transmission are concentrated in Southeast North America (Cuba, Mexico, Honduras and Southeast United States), and Eastern and Northern South America (Colombia, Venezuela and large portions of Brazil). The potential risk for ZIKV transmission is also high in Africa, extending from western regions (Ivory Coast, Ghana and Nigeria) to central regions (Zaire) and southeast regions (Tanzania, Mozambique and Madagascar). Throughout Europe, high-risk areas are concentrated in Italy, followed by France and Slovenia, extending along the Mediterranean coast.

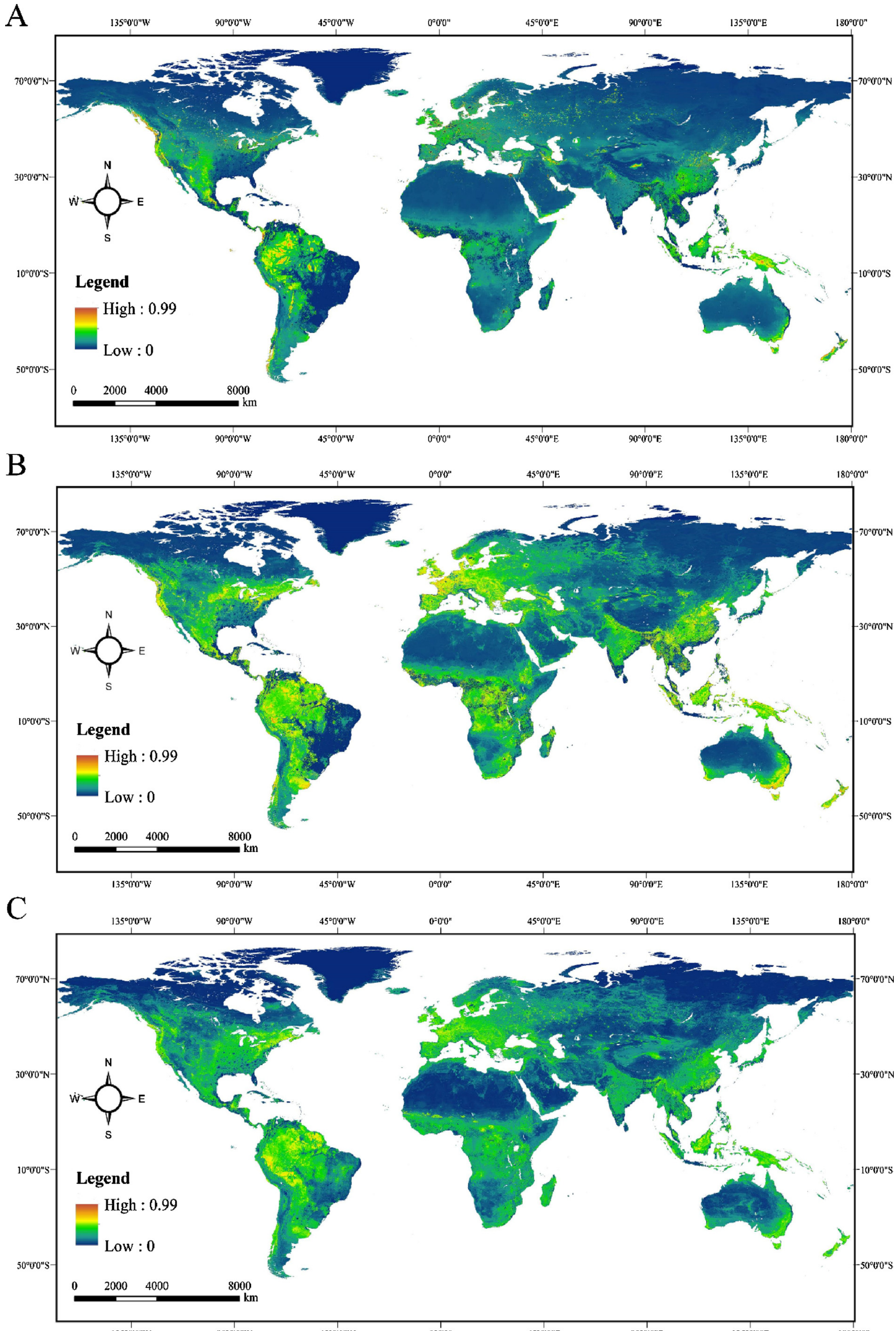


Fig. 5. Quantification of the uncertainty of machine learning models in predicting global ZIKV transmission risk: (A) BPNN; (B) GBM; (C) RF.

Meanwhile, we also find that the South, Southeast and East Asia regions (India, Thailand, China, Korea and Japan) all have large areas of high transmission risk for ZIKV. In Oceania, several areas identified as high transmission risk for ZIKV are mainly located on the eastern coast of Australia and Northern New Zealand. From the local visual effects point of view, however, the major visual difference in thematic maps derived from machine learning models is mainly located in the eastern United States. For example, the area in Fig. 3C is darker than that in Fig. 3A and Fig. 3B, which reflects that the RF-based predicted transmission risk for ZIKV in the eastern United States is lower than that derived from the BPNN and GBM models. Additionally, the predicted high-risk level area produced by the BPNN model is in good agreement with the units where ZIKV infection cases have occurred (Fig. 4).

3.3. The uncertainty of prediction

To analyze the influence of samples on model prediction, Fig. 5 was produced based on 50 prediction results, which show that uncertainty around global ZIKV-predicted transmission risk ranges from 0 to 0.99. Overall, compared with moderate risk areas (0.3–0.7), the uncertainties of high-risk areas (0.7–1) and low-risk areas (0–0.3) are relatively low. The prediction uncertainty of the BPNN model is higher than that of other models in several local regions, including the northwestern part of South America and the eastern part of Indonesia. However, on the whole, Fig. 5A overall is darker than Fig. 5B and C, which indicates the prediction uncertainty of the BPNN model, which is lower than that of the GBM and RF models, especially in Central Africa, Southeastern North America and Southern Europe.

4. Discussion and conclusion

Three relatively popular and robust machine learning models were adopted to predict the transmission risk of ZIKV at a 5×5 km spatial resolution, pairing high-dimensional multidisciplinary covariate layers with comprehensive location data on recorded ZIKV infection in humans. Based on accuracy verification and uncertainty analysis, we found that BPNN model performed better than GBM and RF models, in general, which could make more accurate and scientific prediction for quantifying the areas susceptible to ZIKV outbreak. However, it should be noted that the BPNN model is more like a black box and less interpretable when compared with GBM and RF models.

The risk of ZIKV epidemic spread has been discussed in previous studies. GIS is always adopted to analyze the distribution patterns of reported ZIKV cases on an urban scale, which is mainly based on some geostatistics and spatial superposition methods (Rodriguezmorales et al., 2017; Rodriguez-Morales et al., 2017, 2016). On regional and global scales, niche modeling techniques are often employed by researchers, including Maximum Entropy (Santos and Meneses, 2017; Alaniz et al., 2017) and BRT (Teng et al., 2017; Messina et al., 2016b). The former links explicit environmental layers with records of ZIKA vector species, regarding the potential distribution area of the vector species, as the risk areas of ZIKV, but the input factors, which are not comprehensive enough, are confined to geographical and environmental conditions. The latter is actually a machine learning model, input factors, which are more flexible than that of the former. The predictive uncertainty of ZIKV transmission risk was always overlooked when predicting the potential risk of ZIKV based on the BRT model. Compared with the above studies, our research strengths are as follows: First, we not only considered meteorological and environmental covariates and the demographic distribution of *Aedes* mosquitoes but also incorporated socioeconomic factors reflecting human movements and urbanization into the prediction models to map the transmission of ZIKV. The results derived from F-score method revealed that the discriminatory power of predicted distribution of *Ae. aegypti* was the highest (4.00), followed by predicted distribution of *A. albopictus* (3.80), nighttime lights (0.95), minimum annual temperature (0.88), urbanicity (0.85), NDVI (0.67), maximum annual temperature (0.64), annual cumulative precipitation (0.64), urban accessibility (0.45) and relative humidity (0.17). Second, we adopted three relatively popular and robust machine learning models to simulate the global probabilistic risk of ZIKV transmission and quantified the prediction uncertainty of these models, which could identify the high-risk areas of ZIKV more accurately and scientifically. Last but not least, the significant differences between these predictions' accuracy produced by machine learning models were compared for the first time in the prediction of the global risk level of ZIKV outbreak, which provided some reference information for the selection of models in the field of epidemiological cartography.

However, this study has limitations that should be noted. Although all data used in the present study are generally accepted and accessible to the public, there are several mismatches in temporal and spatial scales between predictor variables and presence points as well as among predictor variables, which may have an impact on predicting the risk level of ZIKV transmission. The multi-year average data can reflect the geographical distribution of covariates, but cannot express the likely difference in nighttime light and urban access between 1951 and 2017, in addition to climate change. Therefore, we will attempt to simulate the global spread risk of ZIKV on a time scale in our next study by combining long time series covariate data. Considering that the world is warming up, machine learning methods will be combined with a variety of climate models to predict transmission risk distribution patterns of ZIKV in future scenarios.

Competing financial interests

The authors declare no competing financial interests.

Author contributions

D.J. and F.D. contributed to all aspects of this work; D.J. and F.D. wrote the main manuscript text; M.H. J.F. and M.L. gave some useful comments and suggestions to this work. All authors reviewed the manuscript.

Acknowledgments

We thank Qiaoling Zhu for providing the valuable suggestions. This research was supported and funded by the Ministry of Science and Technology of China (2016YFC1201300).

References

Alaniz, A.J., Bacigalupo, A., Cattan, P.E., 2017. Spatial quantification of the world population potentially exposed to Zika virus. Int. J. Epidemiol. 0–10.
Attaway, D.F., Waters, N.M., Geraghty, E.M., Jacobsen, K.H., 2017. Zika virus: endemic and epidemic ranges of Aedes mosquito transmission. J. Infect. Public Health 10, 120.
Benelli, G., Mehlhorn, H., 2016. Declining malaria, rising of dengue and Zika virus: insights for mosquito vector control. Parasitol. Res. 115, 1747–1754.
Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., et al., 2013. The global distribution and burden of dengue. Nature 496, 504–507.
Bogoch, I.I., Brady, O.J., Kraemer, M.U.G., German, M., Creatore, M.I., et al., 2016. Potential for Zika virus introduction and transmission in resource limited countries in Africa and Asia-Pacific: a modeling study. Lancet Infect. Dis. 16, 1237.
Boorman, J.P., Draper, C.C., 1968. Isolations of arboviruses in the Lagos area of Nigeria, and a survey of antibodies to them in man and animals. Trans.R. Soc. Trop. Med. Hygiene 62, 269–277.
Brady, O.J., Golding, N., Pigott, D.M., Kraemer, M.U., Messina, J.P., et al., 2014. Global temperature constraints on Aedes aegypti and Ae. albopictus persistence and competence for dengue virus transmission. Parasites Vectors 7, 338.
Brown, J.E., McBride, C.S., Johnson, P., Ritchie, S., Paupy, C., et al., 2011. Worldwide patterns of genetic differentiation imply multiple 'domestications' of Aedes aegypti, a major vector of human diseases. Proc. R. Soc. Lond. B: Biol. Sci. 2446–2454.
Campos, G.S., Bandeira, A.C., Sardi, S.I., 2015. Zika virus outbreak, Bahia, Brazil. Emerg. Infect. Dis. 21, 1885–1886.
Cauchemez, S., Besnard, M., Bompard, P., Dub, T., Guillemette-Artur, P., et al., 2016. Association between Zika virus and microcephaly in French Polynesia, 2013–15: a retrospective study. Lancet 387, 2125–2132.

Center for International Earth Science Information Network - CIESIN - Columbia University, 2016. Global Summer Land Surface Temperature (LST) Grids, 2013. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, NY.
Chan, J.F., Yip, C.C., Tee, K.M., Zhu, Z., Tsang, J.O., et al., 2017. Improved detection of Zika virus RNA in human and animal specimens by a novel, highly sensitive and specific real-time RT-PCR assay targeting the 5'-untranslated region of Zika virus. Trop. Med. Int. Health 22, 594.
Colóngonzález, F.J., Lake, I.R., Bentham, G., 2011. Climate variability and dengue fever in warm and humid Mexico. Am. J. Trop. Med. Hygiene 84, 757–763.
Cordeiro, M.T., Pena, L.J., Brito, C.A., Gil, L.H., Marques, E.T., 2016. Positive IgM for Zika virus in the cerebrospinal fluid of 30 neonates with microcephaly in Brazil. Lancet 387, 1811–1812.
Dick, G.W.A., 1952. Zika virus (II). Pathogenicity and physical properties. Trans. R. Soc. Trop. Med. Hygiene 46, 521.
Dick, G.W.A., Kitchen, S.F., Haddow, A.J., 1952. Zika virus (I). Isolations and serological specificity. Trans. R. Soc. Trop. Med. Hygiene 46, 509–520.
Ding, F., Fu, J., Jiang, D., Hao, M., Lin, G., 2018. Mapping the spatial distribution of Aedes aegypti and Aedes albopictus. Acta Trop. 178, 155–162.
Duffy, M.R., Chen, T.H., Hancock, W.T., Powers, A.M., Kool, J.L., et al., 2009. Zika virus outbreak on Yap Island, Federated States of Micronesia. New Engl. J. Med. 360, 2536–2543.
Enfissi, A., Codrington, J., Roosblad, J., Kazanji, M., Rousset, D., 2016. Zika virus genome from the Americas. Lancet 387, 227–228.
Faye, O., Freire, C.C.M., Iamarino, A., Faye, O., Oliveira, J.V.C.D., et al., 2014. Molecular evolution of Zika virus during its emergence in the 20th century. PLoS Negl. Trop. Dis. 8, e2636.
Fuller, D.O., Troyo, A., Beier, J.C., 2009. El Niño Southern Oscillation and vegetation dynamics as predictors of dengue fever cases in Costa Rica. Environ. Res. Lett. 4, 140111–140118.
Gardner, L.M., Chen, N., Sarkar, S., 2016. Global risk of Zika virus depends critically on vector status of Aedes albopictus. Lancet Infect. Dis. 16, 522–523.
Gilbert, M., Golding, N., Zhou, H., Wint, G.R., Robinson, T.P., et al., 2014. Predicting the risk of avian influenza a H7N9 infection in live-poultry markets across Asia. Nat. Commun. 5, 4116.
Hancock, W.T., Marfel, M., Bel, M., 2014. Zika virus, French Polynesia, South Pacific, 2013. Emerg. Infect. Dis. 20, 1960.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978.
Jane, M., Freya, S., 2016. Global Compendium of Human Zika Virus Occurrence.
Kyeongah, N., Kenji, M., Yuichiro, M., Yohei, Y., Ryo, K., et al., 2016. Estimating risks of importation and local transmission of Zika virus infection. Peerj 4, e1904.
Ladeau, S.L., Allan, B.F., Leisnham, P.T., Levy, M.Z., 2015. The ecological foundations of transmission potential and vector-borne disease in urban landscapes. Funct. Ecol. 29, 889–901.
Lanciotti, R.S., Kosoy, O.L., Laven, J.J., Velez, J.O., Lambert, A.J., et al., 2008. Genetic and serologic properties of zika virus associated with an epidemic, Yap State, Micronesia, 2007. Emerg. Infect. Dis. 14, 1232.
Li, C., Xu, D., Ye, Q., Hong, S., Jiang, Y., et al., 2016. Zika virus disrupts neural progenitor development and leads to microcephaly in mice. Cell Stem Cell 19, 120–126.
Liaw, A., Wiener, M., 2002. Classification and regression by random Forest. R News 2, 18–22.
Liu, Y., Liu, J., Du, S., Shan, C., Nie, K., et al., 2017. Evolutionary enhancement of Zika virus infectivity in Aedes aegypti mosquitoes. Nature 545, 482.
Macnamara, F.N., 1954. Zika virus: a report on three cases of human infection during an epidemic of jaundice in Nigeria. Trans. R. Soc. Trop. Med. Hygiene 48, 139–145.
Mansuy, J.M., Dutertre, M., Mengelle, C., Fourcade, C., Marchou, B., et al., 2016a. Zika virus: high infectious viral load in semen, a new sexually transmitted pathogen? Lancet Infect. Dis. 16, 405.
Mansuy, J.M., Suberbielle, E., Chapuy-Regaud, S., Mengelle, C., Bujan, L., et al., 2016b. Zika virus in semen and spermatozoa. Lancet Infect. Dis. 16, 1106.
Medlock, J.M., Leach, S.A., 2015. Effect of climate change on vector-borne disease risk in the UK. Lancet Infect. Dis. 15, 721.
Medlock, J.M., Vaux, A.G., Cull, B., Schaffner, F., Gillingham, E., et al., 2017. Detection of the invasive mosquito species Aedes albopictus in southern England. Lancet Infect. Dis. 17, 140.
Messina, J.P., Kraemer, M.U., Brady, O.J., Pigott, D.M., Shearer, F.M., et al., 2016a. Mapping global environmental suitability for Zika virus. Elife 5. http://dx.doi.org/10.7554/eLife.15272.
Messina, J.P., Kraemer, M.U., Brady, O.J., Pigott, D.M., Shearer, F.M., et al., 2016b. Mapping global environmental suitability for Zika virus. Elife 5, e15272.
Morrison, A.C., Gray, K., Getis, A., Astete, H., Sihuincha, M., et al., 2004. Temporal and geographic patterns of Aedes aegypti (Diptera: Culicidae) production in Iquitos, Peru. J. Med. Entomol. 41, 1123–1142.
Musso, D., Gubler, D.J., 2016. Zika virus. Clin. Microbiol. Rev. 29, 487.
Musso, D., Nilles, E.J., Caolormeau, V.M., 2014. Rapid spread of emerging Zika virus in the Pacific area. Clin. Microbiol. Infection 20, 595–596.
Pacheco, O., Beltrán, M., Nelson, C.A., Valencia, D., Tolosa, N., et al., 2016. Zika virus disease in Colombia - preliminary report. New Engl. J. Med.
Parra, B., Lizarazo, J., Jiménez-Arango, J.A., Zea-Vera, A.F., González-Manrique, G., et al., 2016. Guillain-Barré syndrome associated with Zika virus infection in Colombia. New Engl. J. Med. 387 1482-1482.
Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., et al., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19, 181–197.
Powell, J.R., Tabachnick, W.J., 2013. History of domestication and spread of Aedes aegypti-a review. Memorias do Instituto Oswaldo Cruz 108, 11–17.
Ridgeway, G., 2006. Gbm: Generalized Boosted Regression Models. R Package Version 1. pp. 55.
Rodriguezmorales, A.J., Galindomarquez, M.L., Garcíaloaiza, C.J., Sabogalroman, J.A., Marinloaiza, S., et al., 2017. Mapping Zika virus disease incidence in Valle del Cauca. Infection 45, 1–10.
Rodriguez-Morales, A.J., García-Loayza, C.J., Galindo-Marquez, M.L., Sabogal-Roman, J.A., Marin-Loayza, S., et al., 2016. Zika infection GIS-based mapping suggest high transmission activity in the border area of La Guajira, Colombia, a northeastern coast Caribbean department, 2015-2016: implications for public health, migration and travel. Travel Med. Infectious Dis. 14, 286–288.
Rodriguez-Morales, A.J., Ruiz, P., Tabares, J., Ossa, C.A., Yepes-Echeverry, M.C., et al., 2017. Mapping the ecoepidemiology of Zika virus infection in urban and rural areas of Pereira, Risaralda, Colombia, 2015–2016: implications for public health and travel medicine. Travel Med. Infectious Dis. 57–66.
Romero-Vivas, C.M., Falconar, A.K., 2005. Investigation of relationships between Aedes aegypti egg, larvae, pupae, and adult density indices where their main breeding sites were located indoors. J. Am. Mosq. Control Assoc. 21, 15–21.
Roth, A., Mercier, A., Lepers, C., Hoy, D., Duituturaga, S., et al., 2014. Concurrent outbreaks of dengue, chikungunya and Zika virus infections - an unprecedented epidemic wave of mosquito-borne viruses in the Pacific 2012-2014. Euro Surveill.: Bulletin Européen sur les maladies transmissibles = European communicable disease bulletin 19.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1988. Learning representations by back-propagating errors. Cogn. Model. 5, 1.
Santos, J., Meneses, B.M., 2017. An integrated approach for the assessment of the Aedes aegypti and Aedes albopictus global spatial distribution, and determination of the zones susceptible to the development of Zika virus. Acta Trop. 168, 80.
Scott, T.W., Amerasinghe, P.H., Morrison, A.C., Lorenz, L.H., Clark, G.G., et al., 2000. Longitudinal studies of Aedes aegypti (Diptera: Culicidae) in Thailand and Puerto Rico: blood feeding frequency. J. Med. Entomol. 37, 89.
Simpson, D.I., 1964. Zika virus infection in man. Trans. R. Soc. Trop. Med. Hygiene 58, 339–344.
Song, B.H., Yun, S.I., Woolley, M., Lee, Y.M., 2017. Zika virus: history, epidemiology, transmission, and clinical presentation. J. Neuroimmunol. 308.
Teng, Y., Bi, D., Xie, G., Jin, Y., Huang, Y., et al., 2017. Model-informed risk assessment for Zika virus outbreaks in the Asia-Pacific regions. J. Infect. 74, 484.
Waddell, L.A., Greig, J.D., 2016. Scoping review of the Zika virus literature. PloS One 11, e0156376.
Wikan, N., Smith, D.R., 2016. Zika virus: history of a newly emerging arbovirus. Lancet Infect. Dis. 16, e119.
Wikan, N., Smith, D.R., 2017. First published report of Zika virus infection in people: Simpson, not MacNamara. Lancet Infect. Dis. 17, 15.
World Health Organization, 2016. Who statement on the first meeting of the International Health Regulations (2005) (IHR 2005) emergency committee on Zika virus and observed increase in neurological disorders and neonatal malformations. Int. Legal Mater. 55, 1010–1011.
Yakob, L., Walker, T., 2016. Zika virus outbreak in the Americas: the need for novel mosquito control methods. Lancet Glob. Health 4, e148–e149.
Zanluca, C., Melo, V.C.A.D., Mosimann, A.L.P., Santos, G.I.V.D., Santos, C.N.D.D., et al., 2015. First report of autochthonous transmission of Zika virus inBrazil. Memórias Do Instituto Oswaldo Cruz 110, 569–572.