Identification of the High-Risk Area for Schistosomiasis Transmission in China Based on Information Value and Machine Learning—A Newly Data-Driven Modeling Try

Yan-Feng Gong
National Institute of Parasitic Diseases https://orcid.org/0000-0003-2540-6560
Ling-Qian Zhu
National Institute of Parasitic Diseases
Yin-Long Li
National Institute of Parasitic Diseases
Li-Juan Zhang
National Institute of Parasitic Diseases
Jing-Bo Xue
National Institute of Parasitic Diseases
Shang Xia
National Institute of Parasitic Diseases
Shan Lv
National Institute of Parasitic Diseases
Jing Xu
National Institute of Parasitic Diseases
Shi-Zhu Li (Lisz@chinacdc.cn)
National Institute of Parasitic Diseases

Research Article

Keywords: Schistosomiasis, Risk prediction, Information value, Machine learning, China.

Posted Date: April 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-445806/v1

Abstract

Objective

Information value (IV) and machine learning models were used to analyze and predict the high-risk distribution of schistosomiasis, in order to provide scientific evidence for disease surveillance and control in China.

Methods

The local case distribution from schistosomiasis surveillance data in China between 2005 and 2019 was assessed based on 19 variables including climate, geography, and social economy. Seven models were built in three categories including IV, three machine learning models (logistic regression, LR; random forest, RF; generalized boosted model, GBM), and three coupled models (coupled model of information value and logistic regression, IV + LR; coupled model of information value and random forest, IV + RF; coupled model of information value and generalized boosted model, IV + GBM). Accuracy, AUC (area under the curve), and F1-score were used to evaluate the prediction performance of the models. The best model was selected to predict the risk distribution for schistosomiasis.

Results

IV + GBM had the highest prediction effect (accuracy = 0.878, AUC = 0.902, F1 = 0.920). The results of IV + GBM showed that the risk area for transmission comprised 4.66% of China, mainly distributed in the coastal regions of the middle and lower reaches of the Yangtze River, the Poyang Lake region, and the Dongting Lake region. Risk areas can be divided into low-risk (2.47%), medium-risk (1.35%), and high-risk (0.84%). High-risk areas are primarily distributed in eastern Changde, western Yueyang, northeastern Yiyang, middle Changsha of the Hunan Province, southern Jiujiang, northern Nanchang, northeastern Shangrao, eastern Yichun in Jiangxi Province, southern Jingzhou, southern Xiantao, middle Wuhan in Hubei Province, southern Anqing, northwestern Guichi, eastern Wuhu in Anhui Province, middle Meishan, northern Leshan, and the middle of Liangshan in Sichuan Province.

Conclusions

The risk of schistosomiasis transmission in China still exists, with high-risk areas relatively concentrated within regions. Coupled models of IV and machine learning provide for effective analysis and prediction, forming a scientific basis for surveillance and control within key areas.

Background

As one of 20 neglected tropical diseases, schistosomiasis is a typical zoonotic parasitic disease that remains a major public health problem worldwide [1]. In the 1950s, schistosomiasis was endemic in 12 southern Chinese provinces in close proximity to the Yangtze River. China was one of the countries with the heaviest schistosomiasis burden with more than 10 million patients. Over the past 70 years of active control, China's schistosomiasis control program has achieved remarkable success [2]. By the end of 2020, 337 (74.89%) of the 450 schistosomiasis endemic counties in China had achieved the elimination standard, 97(21.56%) have achieved the transmission blocking standard and 16 (3.55%) have achieved transmission control [3]. China's "13th Five-Year Plan" for national schistosomiasis control identifies risk monitoring and early warning to be essential to reduce potential transmission risk. Prediction model design is an effective means by which to achieve accurate monitoring and precise control of schistosomiasis [4].

There are two development stages or methods for infectious disease risk prediction: a knowledge-driven method (qualitative method), and a data-driven method (quantitative method) [5]. There are four components to the process of development: epidemic data processing, environmental factor selection, model construction, and model evaluation. In particular, the application of GIS (geographic information system), RS (remote sensing), and GPS (global positioning system) in infectious disease research accelerates the development of quantitative risk prediction [6]. Commonly used qualitative methods are the analytic hierarchy process (AHP) and the Delphi method. For example, Ajakaye et al. [7] used AHP to evaluate the transmission risk of schistosomiasis in Nigeria. Yang et al. [8] used the Delphi method to establish a schistosomiasis early warning index in the middle and lower reaches of the Yangtze River. The results for early warning were consistent with epidemic levels based on a recent epidemiological survey. A single quantitative method or a combination of multiple quantitative methods is frequently used. Solano-Villarreal et al. [9] used a boosted regression tree to study the transmission risk of malaria in the Loreto area. Xia et al. [10] combined a variety of classification algorithms including random forest (RF) and a generalized boosted model (GBM) in BioMod2, to construct a combined model that predicted the potential distribution of *Oncomelania hupensis (O. hupensis)* in the Dongting Lake region. The combined model had greater prediction accuracy.

Information value (IV) is derived by statistical quantitative analysis of data based on information theory. A model is based on the influencing factors of an epidemic as well as an evaluation of risk for the region [11]. As an example, Rai [12] used IV to establish a malaria susceptibility index. IV has high modeling efficiency and can judge the weight of various influencing factors. Classification algorithms such as logistic regression (LR), RF, and GBM determine the weight of each influencing factor [5]. IV and classification algorithms can predict vector infectious diseases during the initial stage. For example, Chen et al. [13] used a coupled IV and LR model (IV + LR) to predict hot spots of HFRS (hemorrhagic fever with renal syndrome) in Hunan Province. The coupled model takes into account the merits of LR and IV, resulting in more reliable and practical prediction. Based on epidemic data and related environmental factors, we used IV combined with LR, RF, and GBM to evaluate and predict the risk for schistosomiasis transmission. The purpose of this study was to provide a methodological basis for monitoring and

control of epidemic schistosomiasis. In this manner, a theoretical knowledge of infectious disease prediction was established.

Materials And Methods

1.1 Study area

Schistosomiasis in China is caused by *Schistosoma japonicum (S. japonicum)*, which completes its life cycle through *O. hupensis*, the only intermediate host. The whole China is rich in geomorphic resources, with many lakes and beaches as well as a wide range of tropical and subtropical monsoon climates. Areas around lakes tend to have a gentle climate with abundant rainfall and vegetation suitable for the breeding of *O. hupensis*. This combination of factors increases the local residents' risk for schistosomiasis, especially in the south of the Yangtze River Basin.

1.2 Data collection

Case and non-case data

Schistosomiasis data were derived from the national schistosomiasis survey of 2005 to 2019 [14]. Villages with local infection cases were selected as distribution points (Fig. 1). Longitude and latitude coordinates of the distribution points were identified with the Baidu map coordinate picking system (http://api.map.baidu.com/lbsapi/getpoint/index.html). Due to a lack of data for nonexistent points, and in order to increase the discrimination of environmental factors, this study randomly selected coordinate points for nonexistent points in non-endemic counties adjacent to schistosomiasis endemic counties based on a ratio of 1:2.

Environmental data

Environmental variables related to schistosomiasis and its vector snail distribution were collected. This included ten climate variables, six geographical variables, and three socio-economic variables, as shown in Table 1. Among the climate related variables, four types of background meteorological data were derived from the Resource and Environmental Science and Data Center of the Chinese Academy of Sciences (http://www.resdc.cn/) and represent conventional climate conditions. The other six bioclimatic variables were based on the high-resolution climate data website WorldClim (https://www.worldclim.org/). Those data included mean diurnal temperature range (BIO2), temperature annual range (BIO7), mean temperature of the warmest quarter (BIO10), mean temperature of the warmest quarter (BIO11), mean temperature of the coldest quarter (BIO16), precipitation of the wettest quarter, and precipitation of the driest quarter (BIO17). These represent extreme climatic conditions and limit the distribution range of *S. japonicum* and *O. hupensis.* Elevation and annual normalized vegetation index for the geographic environmental variables were from the Resource and Environment Science and Data Center of Chinese Academy of Sciences (http://www.resdc.cn/). Landform types and land use types are from the National Earth System Science Data Sharing Platform (http://www.geodata.cn). Distance to

waterways was obtained from WorldPop (https://www.worldpop.org/). Socio-economic variables including gross domestic product, population density, and night light were derived from a map of China. ArcGIS 10.2 software was used to trim all environmental variables to the same spatial range and then resampled to a spatial resolution of 1 km × 1 km.

Table 1
Summary of environmental variables related to the distribution of schistosomiasis and *O. hupensis*

Category	Variable name	Definition	source
Climate variables	AR	Aridity	http://www.resdc.cn/
	IM	Index of moisture	
	AAP	Average annual precipitation	
	AAT	Average annual temperature	
	BIO2	Mean diurnal temperature range	https://www.worldclim.org/
	BIO7	Temperature annual range	
	BIO10	Mean temperature of warmest quarter	
	BIO11	Mean temperature of coldest quarter	
	BIO16	Precipitation of wettest quarter	
	BIO17	Precipitation of driest quarter	
Geographic variables	LF	Landform	http://www.geodate.cn/
	LD	Land use	
	SLOPE	Slope	https://www.worldpop.org/
	DST	Distance to waterway	
	EL	Elevation	http://www.resdc.cn/
	ANDVI	Annual normalized difference vegetation index	
Socio -economic variables	GDP	Gross domestic product	http://www.resdc.cn/
	DP	Density of population	
	NTL	Night-time lights	https://www.worldpop.org/

1.3 Analytical modeling

Information value model (IV)

IV [13] uses the frequency or density of schistosomiasis occurrence to reflect the risk effect of different influencing factors and their sub-intervals. An IV is calculated that represents the contribution of different influencing factors related to the occurrence of schistosomiasis. A regional risk assessment for schistosomiasis transmission is realized through the spatial superposition of multi-factor information [13]. The formula is as follows:

$$I = \sum_{i=1}^{n} lg \frac{Ni/N}{Si/S}$$

When I is positive, the combination of multiple factors will increase the risk of schistosomiasis in grid cells, otherwise, it is not conducive to the occurrence of schistosomiasis. The IV model was implemented in R language (version 4.0.0; R Core Team 2020) using the "scorecard" package.

Machine learning

A logistic regression model (LR) [15] is a statistical nonlinear classification method based on logit transformation, which is widely used in classification and prediction tasks due to its simplicity, rapidity, and relative accuracy. LR takes the selected factors as independent variables and the occurrence of results (occurrence is 1, non-occurrence is 0) as dependent variables. LR uses the "H20" package to implement the modeling process in the R language (version 4.0.0; R Core Team 2020).

A random forest model (RF) [16] is a predictive model based on statistical analysis principles formed by the combination of multiple decision trees. The basic random forest concept is to use bootstrap to extract *k* samples from the original training set, with the sample size of each sample the same as the original training set. Next, *k* decision tree models are established for *k* samples to obtain *k* classification results. Finally, a vote on each record is used to determine a final classification according to *k* classification results. RF uses the "randomForest" package to implement the modeling process in the R language (version 4.0.0; R Core Team 2020).

A generalized boosted model or gradient boosting machine (GBM) [17] is based on two algorithms: regression trees and gradient boosting. It builds multiple regression trees on the basis of self-learning and multiple random selections. The process of multiple fittings can gradually reduce the error of model fitting, and in turn, the simulation accuracy of the regression tree is stably improved. For classification problems, the GBM algorithm only needs to limit the base classifier in the AdaBoost algorithm to a classification tree. The GBM model uses the "H20" package in the R language to implement the modeling process.

Model coupling

Using calculated information value “I” to replace the corresponding frequency ratio of LR, sample variable values for RF and GBM, and coupled models (IV + LR, IV + RF, and IV + GBM) are obtained.

1.4 Model evaluation

The sample data were randomly divided into two parts: 75% as training samples for model construction, and 25% as test samples to evaluate the accuracy. A confusion matrix was used to reflect the comprehensive performance of the models. The accuracy, AUC (area under the curve), and F1-score derived from the confusion matrix were used to evaluate the prediction effect.

Table 2
Confusion matrix of binary classification results

Predicted result	Predicted presence	Predicted absence
Investigated presence	a	b
Investigated absence	c	d
Note: a. True predicted presence; b. False predicted presence; c. False predicted absence; d. True predicted absence.		

Accuracy = (a + d) / (a + b + c + d); F1 = (2(a / (a + b) * a / (a + c)) / (a / (a + b) + a / (a + c)). The higher the accuracy and F1, the better the prediction effect of the model [18]. The AUC is derived from the receiver operating characteristic curve, which takes the true positive rate (a / a + c) as the ordinate and the false positive rate (b / b + d) as the abscissa according to a series of different dichotomies. The AUC threshold is (0.5, 1), where 0.5 represents a completely random classification, and 1 represents completely correct classification, so the larger the AUC value, the better the performance of the model [19].

1.5 Risk visualization analysis

We selected the optimal model based on the evaluation indicator and calculated the transmission risk index for the study area. Then, the area was divided into four levels: no-risk area (0.00 to 0.40), low-risk area (0.41 to 0.60), medium-risk area (0.61 to 0.80), and high-risk area (0.81 to 1.00) [20].

Results

2.1 Correlation analysis among schistosomiasis and environmental factors

Based on the principle of chi-square binning, the upper limit of binning is set to 8, and the IV of different levels of influencing factors is calculated according to the binning situation (Table 3). When annual average temperature is 11.5–19.0°C, the annual average rainfall is 1000–1550 mm, the dryness is 66%-92%, and the wetness index is 45%-70%, schistosomiasis is more likely to occur. In this geographic environment, the risk of schistosomiasis transmission is higher when the distance from waterways is less than 2.5 km, the altitude is less than 100 m, the land use is paddy field, grassland, and water area, and the landform type is plain. Among the socio-economic variables, when population density is above 200, the GDP is over 100, and Night-time lights are above 0.12, then a situation exists that is more disease epidemic-prone. Extreme climate and geographic conditions are not conducive to the spread of

schistosomiasis: for example, annual rainfall of less than 1000 mm or more than 1550 mm, annual average temperature of less than 11.5°C or more than 19°C, average temperature during the hottest season of less than 27°C, rainfall in the wettest season of less than 500 mm, and distance to the waterway of more than 3 km, with a slope greater than six (Table 4).

Table 3
Number and meaning of environmental factor classification based on the principle of chi-square binning

Factors	Number	Classification index
AAP (mm)	8	< 850;850–950;950–1000;1000–1350;1350–1450;1450–1500;1500–1550;>1550
AAT (°C)	8	< 11.5;11.5–16.0;16.0–17.0;17.0-17.5;17.5–18.0;18.0-18.5;18.5–19.0;>19.0
IM (%)	8	< 45;45–50;50–55;55–60;60–65;65–70;70–90;>90
AR (%)	8	< 62;62–66;66–68;68–72;72–74;74–92;92–95;>95
BIO2	8	< 7.3;7.3–7.8;7.8–7.9;7.9–8.2;8.2–8.6;8.6–9.3;9.3–9.9;>9.9
BIO7	8	< 24;24-27.5;27.5–29;29–31;31-31.5;31.5–33;33-33.5;>33.5
BIO10 (°C)	8	< 17;17–20;20–22;22–25;25-26.5;26.5–27;27–28;>28
BIO11 (°C)	8	< 5.8;5.8-6.0;6.0-6.2;6.2–6.4;6.4–6.6;6.6–7.6;7.6–8.6;>8.6
BIO16(mm)	8	< 440;440–460;460–480;480–500;500–520;520–540;540–560;>560
BIO17(mm)	8	< 20;20–50;50–130;130–140;140–155;155–160;160–175;>175
LF	6	Plains; terraces; hills; small undulating mountains; medium undulating mountains; large undulating mountains
LD	7	Paddy field; dry land; woodland; grassland; water area; urban and rural residential land; unused land
EL(m)	7	< 50;50–100;100–450;450–700;700–2150;2150–2500;>2500
SLOPE (°)	8	< 2;2–3;3–6;6–9;9–13;13–22;22–29;>29
DST (km)	8	< 0.5;0.5-1.0;1.0-1.5;1.5-2;2-2.5;2.5-3;3-3.5;>3.5
ANDVI	8	< 0.78;0.78–0.79;0.79–0.8;0.8–0.81;0.81–0.82;0.82–0.83;0.83–0.84;>0.84
GDP(10,000/km2)	7	< 50;50–100;100–150;150–250;250–350;350–800;800–1000;>1000
DP(Person/km2)	8	< 100;100–150;150–200;200–250;250–400;400–450;450–550;>550
NTL	8	< 0.08;0.08–0.10;0.10–0.12;0.12–0.14;0.14–0.16;0.16–0.18;0.18–0.54;>0.54

Table 4
Results for grading IV by environmental influencing factors

Grade	1	2	3	4	5	6	7	8
AAP	-1.435	-0.941	-0.789	0.223	0.901	1.219	0.118	-0.811
AAT	-2.970	0.411	0.647	0.693	0.544	1.067	0.582	-0.305
IM	-0.498	0.916	0.095	0.693	0.818	1.587	-0.288	-1.466
AR	-1.224	-0.693	0.836	0.773	0.383	0.228	-0.801	-1.447
BIO2	0.547	1.176	0.319	0.323	0	-0.553	-1.194	-1.269
BIO7	-0.406	-0.651	-1.504	-1.355	0.357	0.774	0.568	-1.082
BIO10	-2.773	-0.838	-0.693	-1.674	-0.827	-1.584	0.894	1.192
BIO11	-1.064	1.121	1.121	1.118	0.847	0.228	-0.773	-0.406
BIO16	-0.916	-0.074	-0.442	-1.065	0.598	0.223	0.811	0.180
BIO17	-2.110	-0.887	-0.203	0.499	1.421	0.767	0.534	-0.095
LF	0.950	1.068	0.766	-0.300	-0.742	-1.789		
LD	0.347	0.169	-0.266	0.342	0.234	0.123	-1.634	
SLOPE	0.841	0	-0.187	-1.099	-2.485	-0.821	-0.949	-1.946
DST	0.395	0.560	0.821	0.442	0.147	-0.406	0	-0.515
EL	0.959	0.195	-1.126	-0.167	-0.975	-0.651	-2.169	
ANDVI	0.227	-0.105	-0.486	0.223	-0.452	-0.223	-1.299	-0.256
GDP	-1.052	-0.065	0.035	0.773	0.619	0.218	0.211	0.511
DP	-0.946	-0.102	-1.179	0.560	0.431	0.368	1.099	0.621
NTL	-0.887	-0.674	0.111	-0.827	0.143	0.470	0.186	0.450

2.2 Comparison of prediction results based on the seven models

Prediction results for IV, by three machine learning models (LR, RF, GBM), and three coupled models (IV + LR, IV + RF, IV + GBM) are shown in Appendix 1, Figs. 1–7. IV shows that the schistosomiasis risk is widely distributed throughout the Yangtze River Basin and its southern areas. High-risk areas are mainly distributed in southern Hubei, northern Hunan, northwestern Jiangxi, and central Anhui. Prediction results for the three machine learning models had similarities and differences. The possibility for schistosomiasis transmission was mainly concentrated in the middle and lower reaches of the Yangtze River by three machine learning models. LR indicated the risk was also distributed in northern Xinjiang and southwestern Tibet. RF showed a lower risk in southern Guangzhou. GBM showed a lower risk in northern Xinjiang. Prediction results for the three coupled models were better than those for the single

models. North of the Yangtze River there was no obvious abnormal risk, although small detail differences in risk areas were observed. For example, IV + RF showed no obvious risk area in central Sichuan or northwestern Yunnan, as opposed to IV + GBM.

The predicted performance for schistosomiasis by the seven models as judged by transmission risk, accuracy, AUC, and F1 for each model was calculated (Table 5). Sorted model prediction results were ordered as follows: AUC, IV + GBM > IV + RF > GBM > IV + LR > IV > RF > LR. Overall, the coupled models had the best results, followed by the three machine models, and then the information model. The best of the three machine learning models was GBM, and the best of the three coupled models was IV + GBM (accuracy = 0.878, AUC = 0.902, F1 = 0.920).

Table 5
Predictive performance indicators for the seven models

Model	IV	LR	IV + LR	RF	RF + IV	GBM	IV + GBM
Accuracy	0.732	0.790	0.815	0.785	0.820	0.849	0.878
AUC	0.750	0.827	0.853	0.840	0.872	0.859	0.902
F1	0.705	0.867	0.871	0.854	0.875	0.903	0.920

2.3 Risk prediction of schistosomiasis transmission in China based on the optimal coupled model

Prediction results for GBM + IV showed the risk of schistosomiasis in China to be scattered through a large spatial range, although clusters appeared in southeastern Hubei province, northeastern Hunan province, northern Jiangxi Province, central Anhui province, central Sichuan province, northwestern Yunnan province, and southern Jiangsu province. Superimposed on the national river map, risk areas were concentrated in the coastal areas of the middle and lower reaches of the Yangtze River, Poyang Lake region, and Dongting Lake region.

Classification of transmission risk shows that 4.66% of China is in an at-risk area and 95.34% is not. Risk areas can be divided into low-risk (2.47%), medium-risk (1.35%), and high-risk areas (0.84%). High-risk areas are primarily distributed in eastern Changde, western Yueyang, northeastern Yiyang, middle Changsha of the Hunan Province, southern Jiujiang, northern Nanchang, northeastern Shangrao, eastern Yichun in Jiangxi Province, southern Jingzhou, southern Xiantao, middle Wuhan in Hubei Province, southern Anqing, northwestern Guichi, eastern Wuhu in Anhui Province, middle Meishan, northern Leshan, and the middle of Liangshan in Sichuan Province (Fig. 3). Medium-risk areas and low-risk areas are distributed in areas adjacent to high-risk areas, as well as southern Jiangsu and northwestern Yunnan.

Discussion

Due to the unique life history of *S. japonicum* and *O. hupensis*, as well as the numerous terminal hosts of *S. japonicum*, the epidemic process for schistosomiasis is exceedingly complex. Geographic, climatic, socio-economic, and other factors affect the scope and degree of schistosomiasis [21]. In this study,

coupled models for IV and machine learning were used to evaluate factors that interfere with schistosomiasis transmission. A spatial distribution pattern of potential risks provided a support tool for the formulation of macroscopic schistosomiasis control strategies and the development of a quantitative risk assessment model for communicable diseases.

To the best of our knowledge, this is the first time that coupled models of IV and machine learning were applied to schistosomiasis transmission risk. Coupled models were used to establish statistical relationships among case distribution and environmental factors, providing a new method for analysis and prediction of hot spots of schistosomiasis transmission. By comparing the seven model indicators, we found that coupled models have better prediction accuracy than IV and machine learning models alone. The prediction results more accurately reflected the spatial distribution of risk for schistosomiasis. Differences in prediction results and goodness of fit were found for the seven models, reflecting model uncertainty. A final, optimal model, GBM + IV, was selected to predict the risk for schistosomiasis transmission. That model reduced the errors associated with the other models. Machine learning algorithms cannot express the relationships among the influencing factor's internal levels and the occurrence of schistosomiasis. IV does not consider differences in the weight contribution of influencing factors [22]. The higher success rate for the coupled model is that it considers the internal level of influencing factors and the weight of different influencing factors in relationship to schistosomiasis [23]. Therefore, risk prediction results are more scientific and reasonable.

Predicted middle-risk and high-risk areas based on the optimal coupled model were consistent with the areas of schistosomiasis transmission control and blocking in China [24]. Combined with the distribution of water areas in China, the coastal areas of middle and lower reaches of the Yangtze River, the Poyang Lake region, and the Dongting Lake region are the high-risk areas for schistosomiasis spread. This is likely due to the wide distribution and high density of *O. hupensis* in those areas [25]. Further, there are numerous water conservancy projects, frequent population flow, developed animal husbandry industries, and increased opportunities for human and animal contact, placing these regions at risk for schistosomiasis rebound [26–27]. Comprehensive control strategies have focused on infection control, the distribution pattern of intermediate schistosomiasis hosts, composition, and distribution of infection sources. However, the pattern of population activities has undergone significant changes due to floods, disasters [28], wetland construction [29], and global warming [30], which have increased the risk of snail spreading. Hence, there is a greater risk for infection in the areas described above. In the epidemic risk areas, we recommend *O. hupensis* monitoring, strengthened infection control of domestic and wild animals, and timely assessment of epidemic schistosomiasis. In this manner, the goal of schistosomiasis elimination by 2030 will be achieved [31].

The relationships among the spatial change of schistosomiasis risk and environmental factors can be explained by a biological knowledge of *S. japonicum* and snails [32]. Suitable climatic conditions, small slopes, and proximity to rivers are conducive to the growth and reproduction of *S. japonicum* and snails [33], which in turn leads to the prevalence of schistosomiasis. This study demonstrates that temperature, rainfall, altitude, and the risk of schistosomiasis transmission are closely related. Abnormal climatic

conditions will have a negative impact on an epidemic, which confirms previous studies using different methods [34]. Certainly, environmental factors determine the transmission dynamics of schistosomiasis. Previous studies [35] have shown that land use greatly affects the distribution and density of snails in rice fields. When water is high and in proximity to a river, there is an increased risk for infection. This may be due to the increased risk of swimming, fishing, and agricultural activities in contact with water bodies containing cercariae [36]. This study did not find a high risk for schistosomiasis transmission in economically backward areas, which may be due to the large scope of the study and the lack of conditions for schistosomiasis prevalence in backward areas such as Xinjiang and Tibet. Further, results were based on surveillance data from 2005 to 2019 in China, which is accurate and reliable. However, there may be errors in the analysis of relationships among influencing factors and transmission risk due to insufficient case numbers.

This study has limitations. Firstly, although IV + GBM provided high goodness of fit, the potential risk for schistosomiasis remains uncertain, because of other associated factors such as snail control, cattle grazing, water conservancy construction, and behaviors [37–39]. Secondly, risk prediction based on IV + GBM identified sporadic high risk in northern Zhejiang, which is inconsistent with the known elimination of schistosomiasis in Zhejiang. That area is very similar to that of the case distribution in this study but schistosomiasis is no longer endemic in that area due to human intervention, which resulted in a false positive. For the future, more variables related to disease transmission should be collected, which would enrich the data set. Further, IV combined with more classification algorithms would improve assessment. These approaches would result in better predictive model performance and provide guidance for monitoring and early warning of disease in key areas.

Conclusions

This study confirmed that a model that combines IV and machine learning is better than a single model. Among the models, the optimal coupling model had a better predictive performance for schistosomiasis risk assessment, roughly consistent with the actual situation. These results can guide monitoring and control of schistosomiasis and serve as a reference for predicting the risk of other vector-mediated infectious diseases.

Abbreviations

IV: information value; LR: logistic regression; RF, random forest; GBM, generalized boosted model; IV+LR, coupled model of information value and logistic regression; IV+RF, coupled model of information value and random forest; IV+GBM, coupled model of information value and generalized boosted model; AUC, area under the curve; GIS, geographic information system; RS, remote sensing; GPS, global positioning system; AHP, analytic hierarchy process; HFRS, hemorrhagic fever with renal syndrome; *O. hupensis, Oncomelania hupensis; S. japonicum, Schistosoma japonicum.*

Declarations

Acknowledgments

Not applicable

Authors' contributions

YF Gong and SZ Li did the search of literature and wrote the first draft. YF Gong and LQ Zhu performed data analyses. YL Li, LJ Zhang, JB Xue, S Xia, S Lv and J Xu revised the manuscript and provided important intellectual content. LJ Zhang, YL Li, J Xu participated in data collection. SZ Li and YF Gong participated in manuscript writing design. All authors have approved the final manuscript for publication.

Funding

Supported by the National Special Science and Technology Project for Major Infection Diseases of China (No. 2016ZX10004222-004), the Fifth Round of Three-Year Public Health Action Plan of Shanghai (No. GWV-10.1-XK13). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

This paper was based on an analysis of routinely collected surveillance data from national institute of parasitic diseases, China CDC. No individual information was revealed.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

References

1. Li EY, Gurarie D, Lo NC, Zhu X, King CH. Improving public health control of schistosomiasis with a modified WHO strategy: a model-based comparison study. Lancet Glob Health. 2019;7(10):e1414–22.
2. Xu J, Yu Q, Tchuenté LA, Bergquist R, Sacko M, Utzinger J, et al. Enhancing collaboration between China and African countries for schistosomiasis control. Lancet Infect Dis,2016,16(3):376–83.

3. Lv S, Lv C, Li YL, Xu J, Hong QB, Zhou J, et al. Expert consensus on the strategy and measures to interrupt the transmission of schistosomiasis in China [J]. Chin J Schisto Control,2021,33(01):10–14. (in Chinese).
4. Xu J, Li SZ, Chen JX, Wen LY, Zhou XN. Playing the guiding roles of national criteria and precisely eliminating schistosomiasis in P. R. China[J]. Chin J Schisto Control,2017,29(01):1–4. (in Chinese).
5. Zhang ZP, Wei ZH. Landslide susceptibility assessment based on weighted information values model: take Baqiao district as an example[J]. Science Technology and Engineering,2020,20(9):3492–3500. (in Chinese).
6. Yang GJ, Vounatsou P, Zhou XN, Utzinger J, Tanner M. A review of geographic information system and remote sensing with applications to the epidemiology and control of schistosomiasis in China. Acta Trop. 2005;96(2–3):117–29.
7. Ajakaye OG, Adedeji OI, Ajayi PO. Modeling the risk of transmission of schistosomiasis in Akure North Local Government Area of Ondo State, Nigeria using satellite derived environmental data. PLoS Negl Trop Dis. 2017;11(7):e0005733.
8. Yang K, Xu JF, Zhang JF, Li W, He J, Liang S, et al. Establishing and applying a schistosomiasis early warning index (SEWI) in the lower Yangtze River Region of Jiangsu Province, China. PLoS One. 2014;9(4):e94012.
9. Solano-Villarreal E, Valdivia W, Pearcy M, Linard C, Pasapera-Gonzales J, Moreno-Gutierrez D, et al. Malaria risk assessment and mapping using satellite imagery and boosted regression trees in the Peruvian Amazon. Sci Rep. 2019;9(1):15173.
10. Xia C, Hu Y, Ward MP, Lynn H, Li S, Zhang J, et al. Identification of high-risk habitats of *Oncomelania hupensis*, the intermediate host of schistosoma japonium in the Poyang Lake region, China: A spatial and ecological analysis. PLoS Negl Trop Dis. 2019;13(6):e0007386.
11. Tan Y, Guo D, Bo Xu A. Geospatial information quantity model for regional landslide risk assessment. Nat Hazards. 2015;79:1385–98.
12. Rai PK, Nathawat MS, Rai S. Using the information value method in a geographic information system and remote sensing for malaria mapping: a case study from India. Inform Prim Care. 2013;21(1):43–52.
13. Chen Z, Liu F, Li B, Peng X, Fan L, Luo A. Prediction of hot spot areas of hemorrhagic fever with renal syndrome in Hunan Province based on an information quantity model and logistical regression model. PLoS Negl Trop Dis. 2020;14(12):e0008939.
14. Xu J, Li SZ, Zhang LJ, Bergquist R, Dang H, Wang Q, et al. Surveillance-based evidence: elimination of schistosomiasis as a public health problem in the Peoples' Republic of China. Infect Dis Poverty. 2020;9(1):63.
15. Xu JF, Xu J, Li SZ, Jia TW, Huang XB, Zhang HM, et al. Transmission risks of schistosomiasis japonica: extraction from back-propagation artificial neural network and logistic regression model. PLoS Negl Trop Dis. 2013;7(3):e2123.

16. Liang R, Lu Y, Qu X, Su Q, Li C, Xia S, et al. Prediction for global African swine fever outbreaks based on a combination of random forest algorithms and meteorological data. Transbound Emerg Dis. 2020;67(2):935–46.
17. Teng Y, Bi D, Xie G, Jin Y, Huang Y, Lin B, et al. Model-informed risk assessment for Zika virus outbreaks in the Asia-Pacific regions. J Infect. 2017;74(5):484–91.
18. Kim M, Chae K, Lee S, Jang HJ, Kim S. Automated Classification of Online Sources for Infectious Disease Occurrences Using Machine-Learning-Based Natural Language Processing Approaches. Int J Environ Res Public Health. 2020;17(24):9467.
19. Assaf D, Gutman Y, Neuman Y, Segal G, Amit S, Gefen-Halevi S, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med. 2020;15(8):1435–43.
20. Hu XK, Hao YW, Xia S, Guo YH, Xue JB, Zhang Y, et al. Detection of schistosomiasis transmission risks in Yunnan Province based on ecological niche modeling. Chin J Parasitol Parasit Dis. 2020;38(1):80–6,94. (in Chinese).
21. Cheng G, Li D, Zhuang D, Wang Y. The influence of natural factors on the spatio-temporal distribution of Oncomelania hupensis. Acta Trop. 2016;164:194–207.
22. Hembram T, Paul G, Saha S. Spatial prediction of susceptibility to gully erosion in Jainti River basin, Eastern India: a comparison of information value and logistic regression models. Model Earth Syst Environ. 2019;5:689–708.
23. Li ZT, Wang T, Zou Y, Liu JM, Xin P. Landslide susceptibility assessment based on information value model, logistic regression model and their integrated model: a case in Shatang River Basin, Qinghai Province[J]. Geoscience. 2019;33(1):235–45. (in Chinese).
24. Zhang LJ, Xu ZM, Guo JY, Dai SM, Dang H, Lv S, et al. Endemic status of schistosomiasis in People's Republic of China in 2018. Chin J Schisto Control. 2019;31(06):576–82. (in Chinese).
25. Song LG, Wu XY, Sacko M, Wu ZD. History of schistosomiasis epidemiology, current status, and challenges in China: on the road to schistosomiasis elimination. Parasitol Res. 2016;115(11):4071–81.
26. Hu F, Ge J, Lv SB, Li YF, Li ZJ, Yuan M, et al. Distribution pattern of the snail intermediate host of schistosomiasis japonica in the Poyang Lake region of China. Infect Dis Poverty. 2019;8(1):23.
27. Li FY, Hou XY, Tan HZ, Williams GM, Gray DJ, Gordon CA, et al. Current Status of Schistosomiasis Control and Prospects for Elimination in the Dongting Lake Region of the People's Republic of China. Front Immunol. 2020;11:574136.
28. Li YS, Raso G, Zhao ZY, He YK, Ellis MK, McManus DP. Large water management projects and schistosomiasis control, Dongting Lake region, China. Emerg Infect Dis. 2007;13(7):973–9.
29. Anthonj C, Diekkrüger B, Borgemeister C. Thomas Kistemann. Health risk perceptions and local knowledge of water-related infectious disease exposure among Kenyan wetland communities. Int J Hyg Environ Health. 2019;222(1):34–48.
30. Stensgaard AS, Vounatsou P, Sengupta ME, Utzinger J. Schistosomes, snails and climate change: Current trends and future expectations. Acta Trop. 2019;190:257–68.

31. World Health Organization. (2020). Ending the neglect to attain the Sustainable Development Goals: a road map for neglected tropical diseases 2021–2030. World Health Organization. [EB/OL].(2021-03-20). https://apps.who.int/iris/handle/10665/338565.
32. Hu Y, Ward MP, Xia C, Li R, Sun L, Lynn H, et al. Monitoring schistosomiasis risk in East China over space and time using a Bayesian hierarchical modeling approach. Sci Rep. 2016;6:24173.
33. Olkeba BK, Boets P, Mereta ST, Yeshigeta M, Akessa GM, Ambelu A, et al. Environmental and biotic factors affecting freshwater snail intermediate hosts in the Ethiopian Rift Valley region. Parasit Vectors. 2020;13(1):292.
34. Yang J, Zhao Z, Li Y, Krewski D, Wen SW. A multi-level analysis of risk factors for Schistosoma japonicum infection in China. Int J Infect Dis. 2009;13(6):e407-12.
35. Niu Y, Li R, Qiu J, Xu X, Huang D, Qu Y. Geographical Clustering and Environmental Determinants of Schistosomiasis from 2007 to 2012 in Jianghan Plain, China. Int J Environ Res Public Health. 2018;15(7):1481.
36. Angora EK, Boissier J, Menan H, Rey O, Tuo K, Touré AO, et al. Prevalence and Risk Factors for Schistosomiasis among Schoolchildren in two Settings of Côte d'Ivoire. Trop Med Infect Dis. 2019;4(3):110.
37. Yang Y, Zhou YB, Song XX, Li SZ, Zhong B, Wang TP, et al. Integrated Control Strategy of Schistosomiasis in The People's Republic of China: Projects Involving Agriculture, Water Conservancy, Forestry, Sanitation and Environmental Modification. Adv Parasitol. 2016;92:237–68.
38. Cao CL, Zhang LJ, Deng WP, Li YL, Lv C, Dai SM, et al. Contributions and achievements on schistosomiasis control and elimination in China by NIPD-CTDR. Adv Parasitol. 2020;110:1–62.
39. Qiu J, Li R, Zhu H, Xia J, Xiao Y, Huang D, et al. The effect of ecological environmental changes and molluscicide on snail intermediate host of Schistosoma in Qianjiang city of China from 1985 to 2015. Parasit Vectors. 2020;13(1):397.

Figures

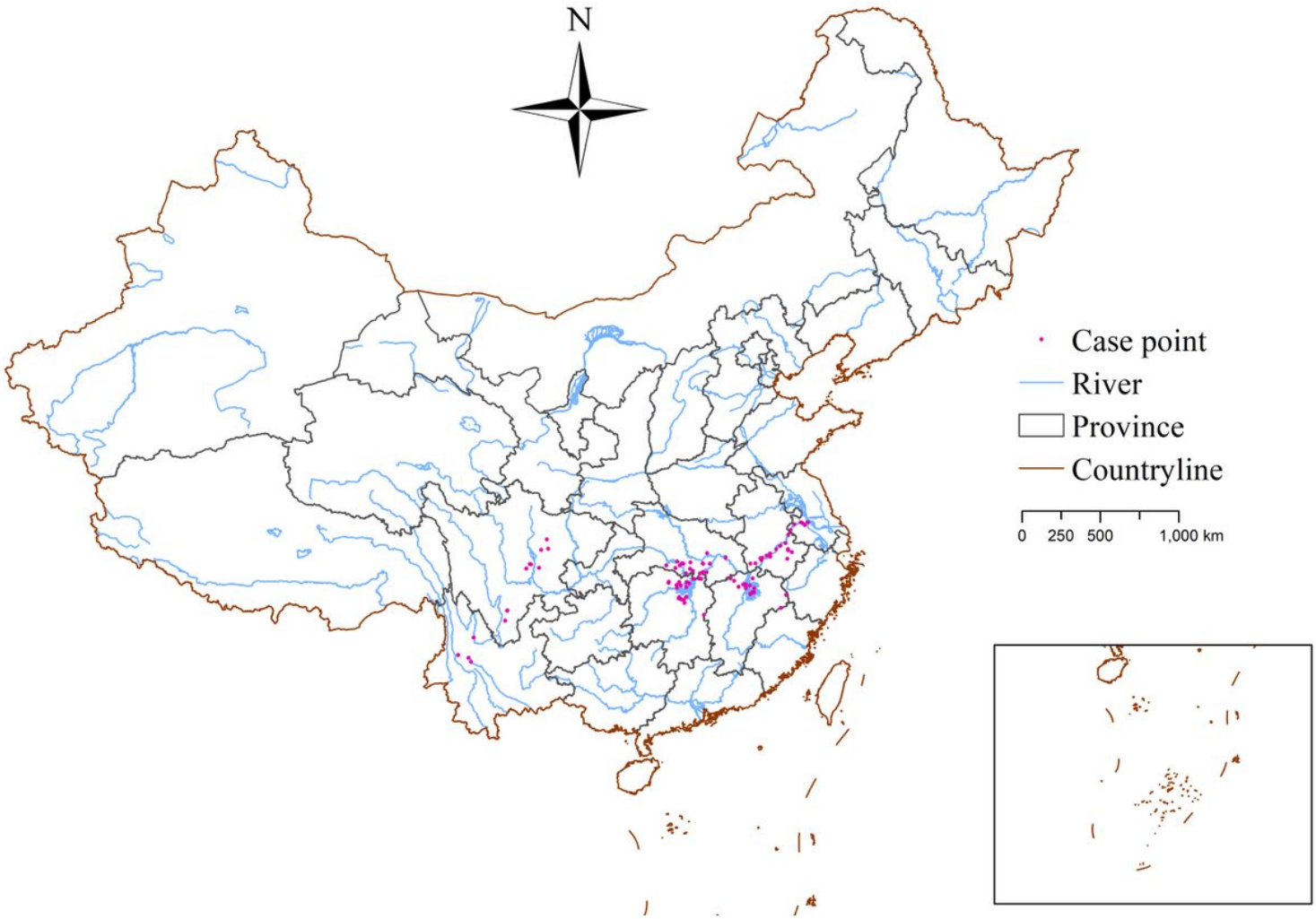

Figure 1

Study area and case distribution Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

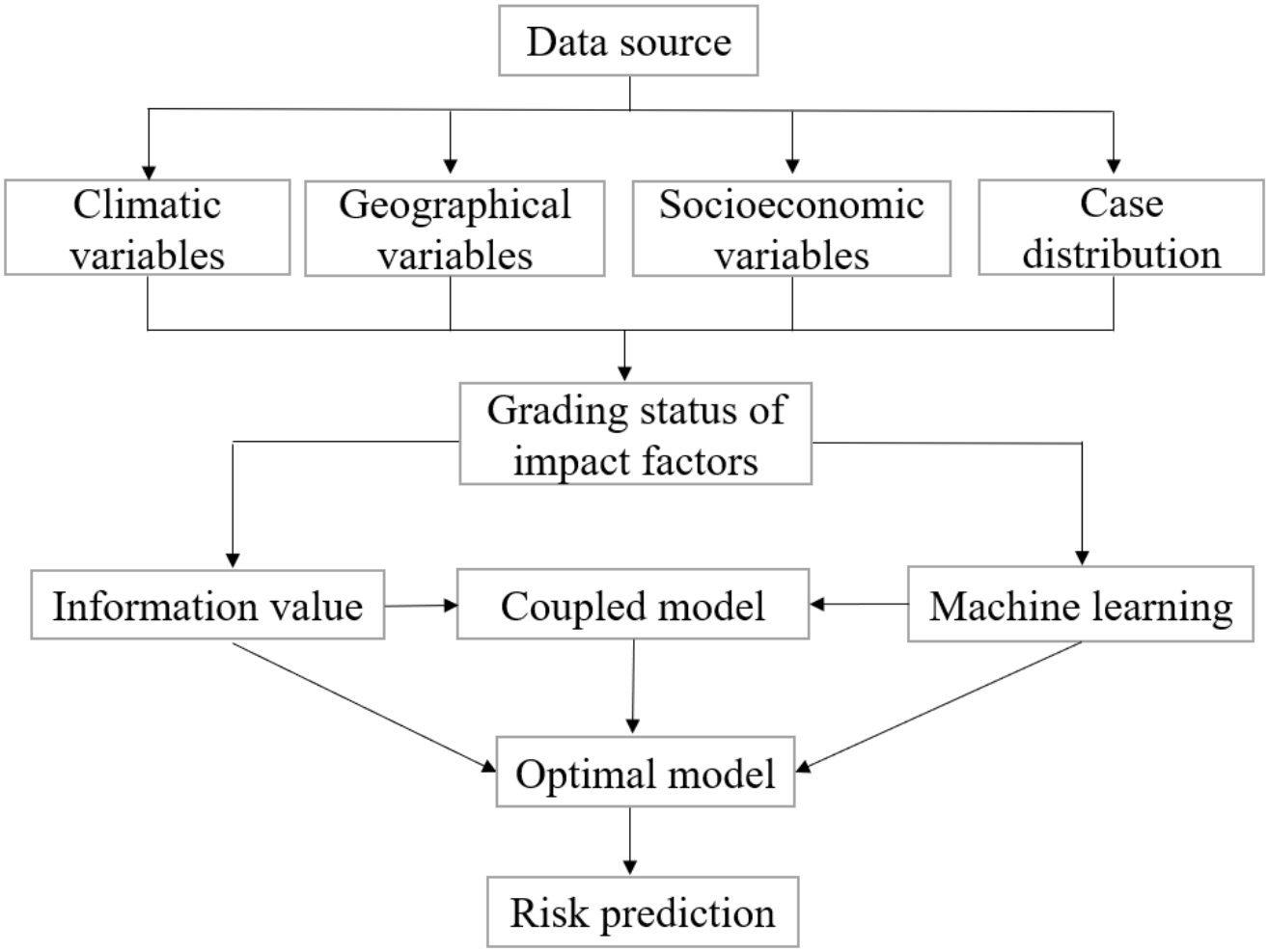

Figure 2

Implementation path of model building

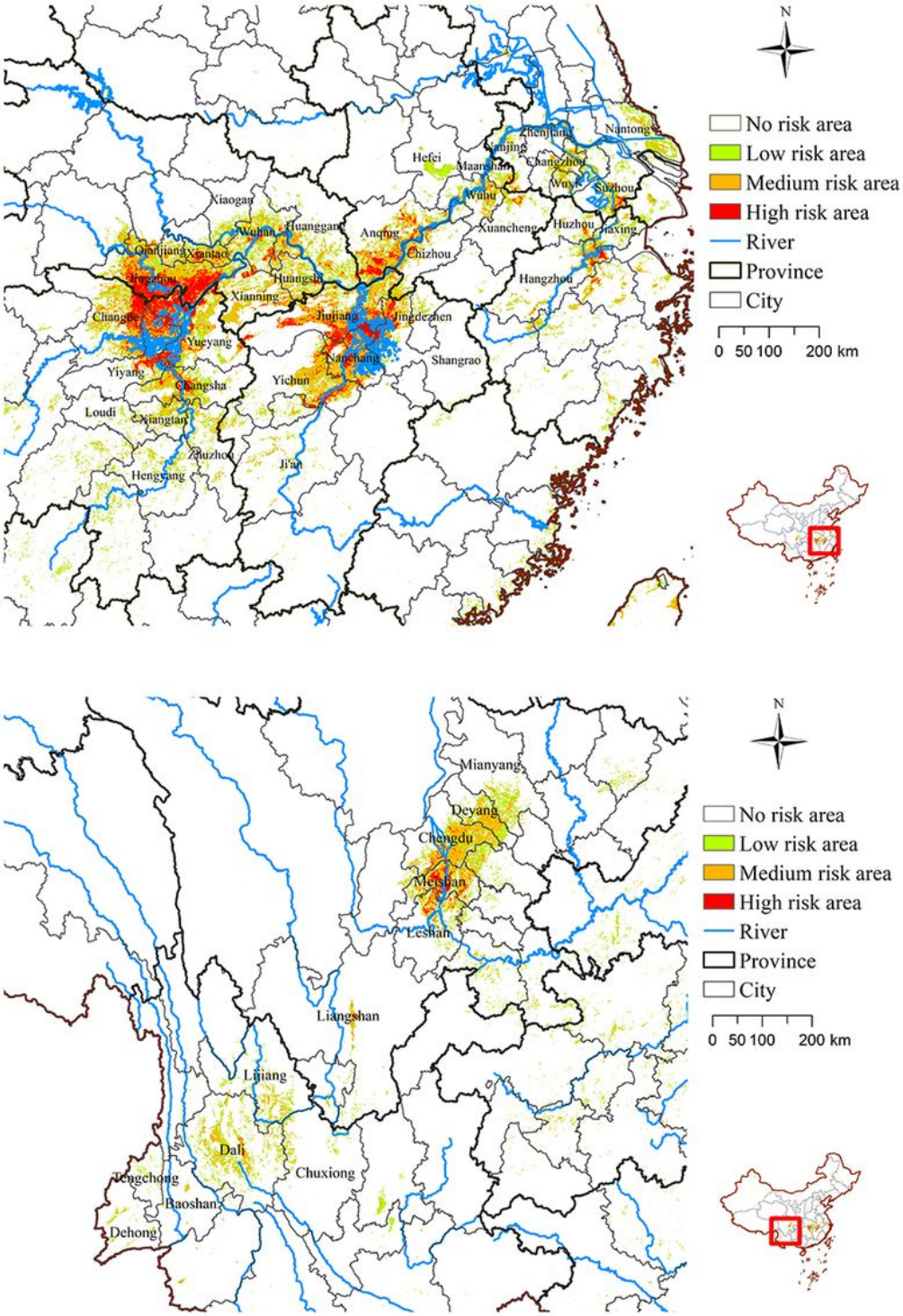

Figure 3

Current risk prediction for schistosomiasis in China based on the optimal coupled model Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix1.Figures1.jpg
- Appendix1.Figures2.jpg
- Appendix1.Figures3.jpg
- Appendix1.Figures4.jpg
- Appendix1.Figures5.jpg
- Appendix1.Figures6.jpg
- Appendix1.Figures7.jpg