Impacts of sample ratio and size on the performance of random forest model to predict the potential distribution of snail habitats

Yuanhua Liu,[1]* Jun Zhang,[1]* Michael P. Ward,[2] Wei Tu,[3] Lili Yu,[4] Jin Shi,[1] Yi Hu,[1] Fenghua Gao,[5] Zhiguo Cao,[5] Zhijie Zhang[1]

[1]Key Laboratory of Public Health Safety of Ministry of Education, Department of Epidemiology and Health statistics, School of Public Health, Fudan University, Shanghai, China; [2]Sydney School of Veterinary Science, The University of Sydney, Sydney, Australia; [3]Department of Geology and Geography, Georgia Southern University, Statesboro, GA, USA; [4]Peace Center for Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA; [5]Anhui Institute of Schistosomiasis Control, Hefei, China

***These authors contributed equally**

Abstract

Few studies have considered the impacts of sample size and sample ratio of presence and absence points on the results of random forest (RF) testing. We applied this technique for the prediction of the spatial distribution of snail habitats based on a total of 15,000 sample points (5,000 presence samples and 10,000 control points). RF models were built using seven different sample ratios (1:1, 1:2, 1:3, 1:4, 2:1, 3:1, and 4:1) and the optimal ratio was identified via the Area Under the Curve (AUC) statistic. The impact of sample size was compared by RF models under the optimal ratio and the optimal sample size. When the sample size was small, the sampling ratios of 1:1, 1:2 and 1:3 were significantly better than the sample ratios of 4:1 and 3:1 at all four levels of sample sizes ($p<0.01$) and there was no significant difference among the ratios of 1:1, 1:2 and 1:3 ($p>0.05$). The sample ratio of 1:2 appeared to be optimal for a relatively large sample size with the lowest quartile deviation. In addition, increasing the sample size produced a higher AUC and a smaller slope and the most suitable sample size found in this study was 2400 (AUC=0.96). This study provides a feasible idea to select an appropriate sample size and sample ratio for ecological niche modelling (ENM) and also provides a scientific basis for the selection of samples to accurately identify and predict snail habitat distributions.

Correspondence: Zhijie Zhang, School of Public Health, Fudan University, 200032 Shanghai, China.
Tel.: +86.13817217362.
E-mail: epistat@gmail.com

Key words: sample ratio, sample size, random forest model, snail habitats distributions.

Conflict of interest: the authors declare no potential conflict of interest, and all authors confirm accuracy.

Availability of data and materials: all data generated or analyzed during this study are included in this published article.

Acknowledgements: this review was supported by the National Natural Science Foundation of China (81973102), the National Key Research and Development Program of China (2022YFC2602900), Public Health Talents Training Program of Shanghai Municipality (GWV-10.2-XD21), and the Shanghai New Three-year Action Plan for Public Health (GWV-10.1-XK16).

Received: 13 September 2022.
Accepted: 8 February 2023.

Geospatial Health 2023; 18:1151
doi:10.4081/gh.2023.1151

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Introduction

Schistosomiasis is a neglected parasitic disease caused by trematode parasites of the genus *Schistosoma*, which affects at least 206 million people worldwide according to a recent WHO report (Colley *et al.*, 2014). In China, *Oncomelania hupensis* is the sole intermediate host of *Schistosoma japonicum* (Zou & Ruan, 2015) and snail control is considered to be the most effective approach to control this disease (Zhang & Jiang, 2011). Effective snail control depends largely on how accurately the snail habitats can be located, however manual snail-searching is labour-intensive, expensive, and time-consuming (Guo *et al.*, 2005; Zhu *et al.*, 2015). Hence, we need more efficient and effective methods to identify potential snail habitats.

Ecological niche modelling (ENM), also known as species distribution modelling (SDM), is a class of methods that can be used to predict the potential distribution of species. ENM uses species distribution data and related environmental variables data to make a correlative model of the environmental conditions that meet a species' ecological requirements (Warren & Seifert, 2011).

ENM has been increasingly used to predict regions of occurrence of vectors and pathogens, which are frequently associated with human diseases in new areas where investigations have not been previously carried out (Sage *et al.*, 2017; Chalghaf *et al.*, 2018). More specifically, ENM has been used to predict the potential spatial distribution of the intermediate host snails of *Schistosoma* (Pedersen *et al.*, 2014; Scholte *et al.*, 2012; Zhu *et al.*, 2017). Recent studies found that machine-learning models, especially random forest (RF; a type of ENM), performed better in predicting potential snail habitats (Xia *et al.*, 2019; Zhang *et al.*, 2020). However, few RF-based ENM studies have explored the impacts of sample size and sample ratio (the ratio of positive samples and control samples) on the results even though both issues have been frequently addressed in previous studies using traditional ENM, such as Maxent and GARP (Bean *et al.*, 2012; Hernandez *et al.*, 2006; Stockwell & Peterson, 2002).

Furthermore, to the best of our knowledge, no studies have investigated the impact of sample size on ENM that had been designed to predict the distribution of vectors (such as snails) and pathogens associated with human diseases. Generally, the larger the sample size, the greater will be the estimated accuracy of ENM, but also the higher the sample survey cost (Peterson *et al.*, 2007). Thus, an optimal sample size and ratio are vital for predictive models considering cost-effectiveness.

In this study, we attempted to identify an appropriate sample size and ratio of ENM for predicting snail habitats based on the RF type of ENM, which was found to perform the best in our previous researches (Xia *et al.*, 2019; Zhang *et al.*, 2020). We first compared different sample ratios (presence samples versus absence samples in the case of the same total sample size) to select an optimal ratio, and then identified the most appropriate sample size based on the optimal sample ratio. Finally, we explored the impacts of sample ratio and sample size on the model performance for predicting the potential distribution of snail habitats.

Materials and Methods

Data sources

The snail data used in our study were from a snail survey conducted from March to May 2016, covering the whole Anhui Province, a traditional schistosomiasis-endemic area in eastern China (Cao *et al.*, 2018). For the survey, a 0.11m^2 frame was set in every 10*20m area selected by systematic sampling in the study area. The exact snail locations (by geographical coordinates) were recorded by a handheld global positioning system (GPS) instrument (Garmin GPSMAP 64s). After the survey, parts of the area where snails were found in the survey were randomly selected as the sampling point for this study. A total of 5,000 snail sample points (presence samples) were collected in the lake and marshland areas, while 10,000 control points were generated by random sampling throughout the entire study area excluding areas within 100 m of presence points, and these control points were used as the points for which absence of snails in modelling (further details in Zhang *et al.*, 2020). Figure 1 shows the distribution of the snail sample points in our study area. To predict potential snail habitats, 19 kinds of environmental factors - including climatic, soil and geomorphological factors - were screened and selected from 36 environmental factors (*Supplementary Table S1*). Among these environmental variables, the land surface temperature (LST) and normalized difference vegetation index (NDVI) were calculated from remotely sensed image data collected by Landsat 7 equipped with the Enhanced Thematic Mapper Plus (ETM+). The slope orientation (Asp) and slope were extracted from the elevation module (DEM) in the United States Geological Survey (USGS) Global Land Information System (GLIS). The climate variables Bio1 to Bio19 were obtained from WorldClim. We calculated the distances to the nearest water body (water) from the water body data

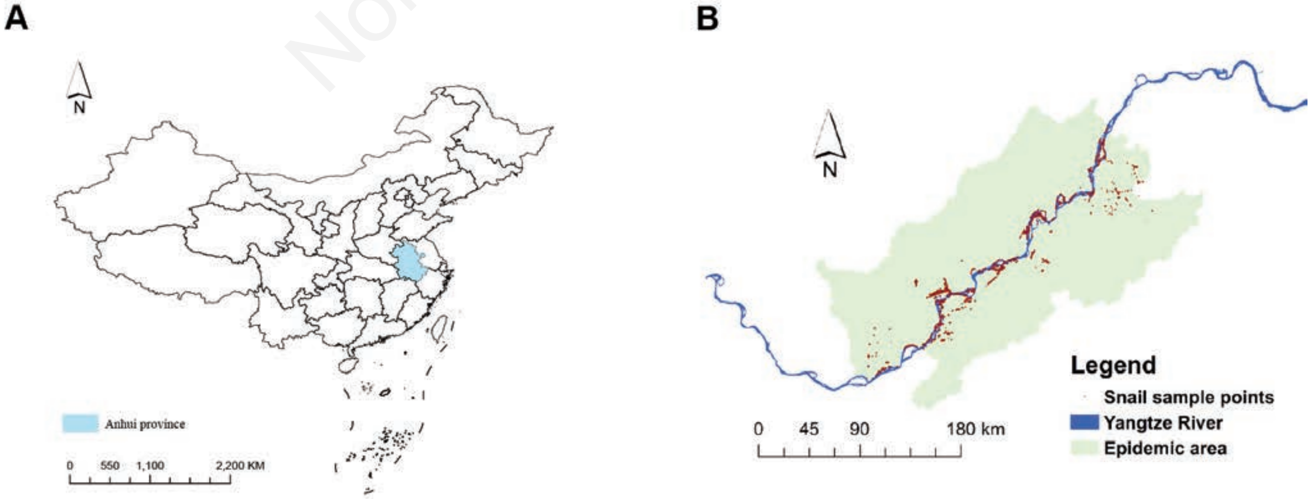

Figure 1. The epidemic areas and sample points in Anhui Province: A) location of Anhui Province (in blue shade) in Chaina; B) the epidemic areas and sample points in Anhui Province shown where traversed by the Yangtze River. Snail habitat sample points are marked in red. The map was created using the ArcGIS 10.0 software (ESRI Inc., Redlands, CA, USA).

obtained from conservation science datasets of the World Wildlife Fund (https://www.worldwildlife.org/pages/conservation-science-data-and-tools). Climatic variables and data on soil, such as geomorphic type (Geo), vegetation type (Veg), ecosystem type (Eco) and land use type (Lucc) came from the Data Center for Resources and Environmental Sciences of Chinese Academy of Sciences.

To screen the environmental variables and avoid potential multi-collinearity, we conducted a correlation analysis for all climatic variables. We excluded one of the variables if the correlation coefficient of a pair of variables >0.7. The screening process methodology for environmental factors was based on a previous study (Escobar & Craft, 2016). The selected raw environmental data were standardized to be in the same range and rasterized on the study area using a cell size of 100×100 m.

Modelling and evaluation

The snail location dataset was split randomly into two parts, 80% of the data were used as the training set for model development and the remaining 20% served as the test set for model evaluation. To compare the predictive capability of different sample ratios, we built RF models using 7 different sample ratios of presence points and absence points (1:1, 1:2, 1:3, 1:4, 2:1, 3:1, 4:1) with 4 total sample sizes (100, 500, 1000, 5000). We chose 4 samples sizes because we found that different sample ratios behaved better in the case of different total sample sizes when the models were built. It is a common practice to use 20% of the total dataset for evaluating the performance of all models built via AUC (Escobar & Craft, 2016). Models for each sample ratio were repeated 100 times with each one of four sample sizes to obtain the distribution of AUC to address model uncertainty. A higher AUC value indicates better model performance. The optimal sample ratio was then determined and used in the RF models with different sample sizes. To accurately evaluate the impact of sample size on the models, we built models using different sample sizes from 120 to 12,000 with fixed increments of 120, which was determined based on the optimal sample ratio. The impacts of sample size on the model performance were evaluated using four different indicators: AUC, sensitivity, specificity, Kappa and percent correctly classified (PCC). AUC, sensitivity and specificity were used to evaluate the accuracy of the models, while Kappa and PCC were used to evaluate the accuracy of the prediction results. We set a threshold of 0.5 for sensitivity and specificity, in which a value greater than (or equal to) 0.5 represents the potential positive area (snails are present) and a value less than 0.5 represents the potential negative area (snails are absent) (Liu *et al.*, 2011). In our study, all RF models were built using the *Biomod2* package for R (Thuiller *et al.*, 2009).

Results

Sample ratio

Figure 2 shows the AUC of models built using 7 sample ratios at 4 different sample sizes. Figure 2 indicates that models built using more absence samples were overall better than those with more presence samples. In addition, the ratio of 1:2 and 1:3 were better than the other ratios. The results of a pair-wise comparison (*Supplementary Table S2*) show that there was no significant statistical difference between the sample ratio of 1:1, 1:2, and 1:3 (p>0.05) and that these three ratios were significantly better than the sample ratios of 4:1 and 3:1 at all the four size levels sizes (p<0.01). The sample ratio of 1:2 appeared to be better in the case of a large sample size, and there was no significant difference between 1:1, 1:2 and 1:3 when the sample size was small.

The exact quartile deviations of the 7 sample ratios at the 4 sample sizes were also calculated (Table 1). It was found that the larger the sample size, the smaller the standard deviation. Moreover, the quartile deviations of the sample ratio of 1:2 were generally found to be the lowest among all the ratios (*Supplementary Table S3*).

Sample size

According to the quartile deviation of the sample ratios, we used 1:2 as the fixed (optimal) ratio to evaluate the effect of sample size on model performance. The effect of sample size is shown in Figure 3. It is obvious that a larger sample size corresponded to higher AUC as well as smaller slope (Figure 3 A), and a similar trend was also found for specificity (Figure 3 C). Moreover, sensitivity increased first and then decreased with increasing sample size (Figure 3 B). Other indicators showed trends similar to AUC and specificity (*Supplementary Figure S1*). The receiver operating characteristic (ROC) curves were drawn based on three typical sample sizes, large (12000), median (2400), and small (120) (Figure 4). The AUC of the models built by using these three sample sizes were 0.98, 0.96, and 0.92, respectively.

Discussion

ENM has been widely used to predict the distribution of various diseases or pathogens but most previous studies just investigated the impacts of the sample size on ENM, while only few studies focused on presence-only models, such as Maxent and GARP

Table 1. The quartile deviation of models built with seven sample ratios and four sample sizes.

Sample ratio*			Sample size	
	100	500	1,000	5,000
1:1	0.046	0.006	0.005	0.002
1:2	0.034	0.007	0.004	0.002
1:3	0.040	0.009	0.005	0.002
1:4	0.049	0.011	0.005	0.002
2:1	0.056	0.009	0.006	0.002
3:1	0.054	0.012	0.007	0.002
4:1	0.063	0.013	0.007	0.004

*Sample ratio represents the ratio of presence and absence samples used in the models.

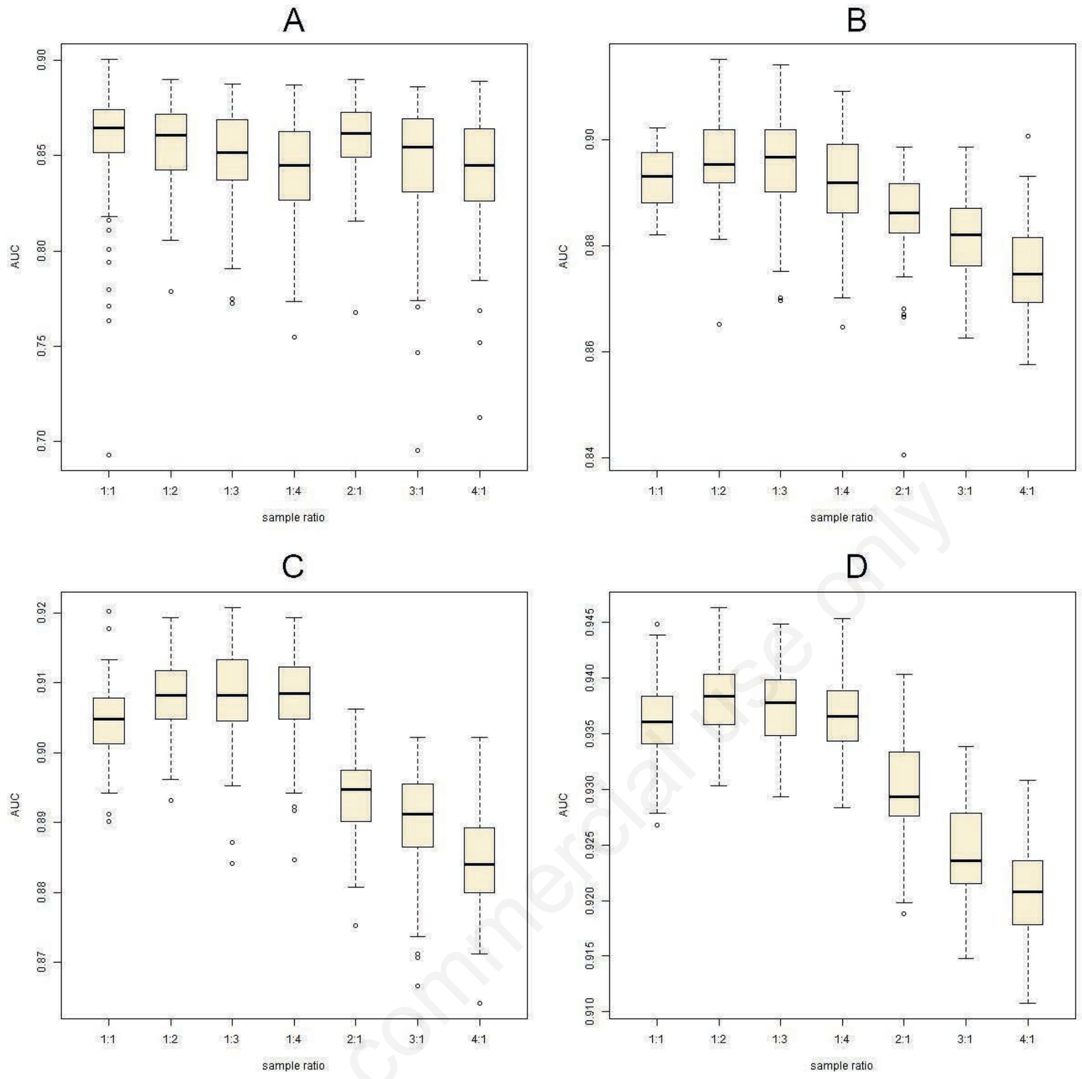

Figure 2. The AUC of models built using seven sample ratios and four sample sizes: A) represents the total size of 100 sample points; B) 500 sample points; C 1,000 sample points; D 5,000 sample points. Each panel of the figure corresponds with the box plot of the AUC of the model built using 7 sample ratios.

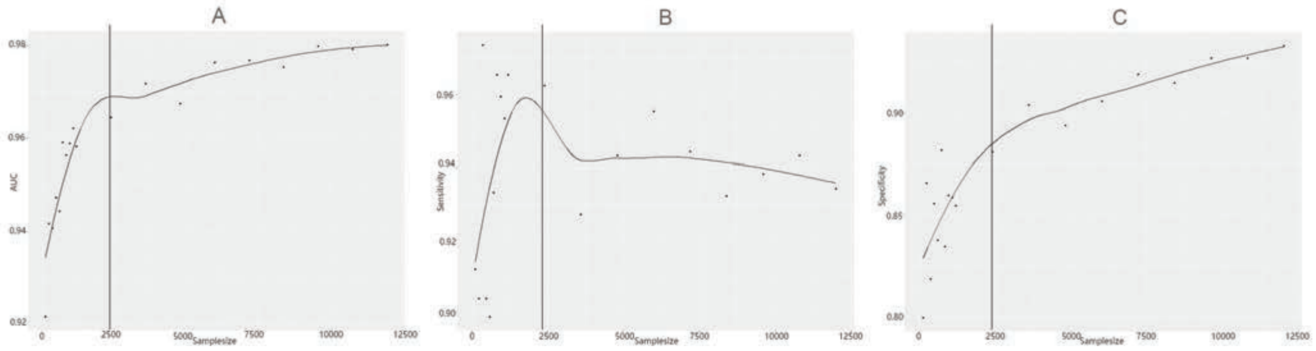

Figure 3. The relationship between sample size and AUC, sensitivity and specificity. The sample ratio was set as 1:2; the vertical line marks a sample size of 2,400.

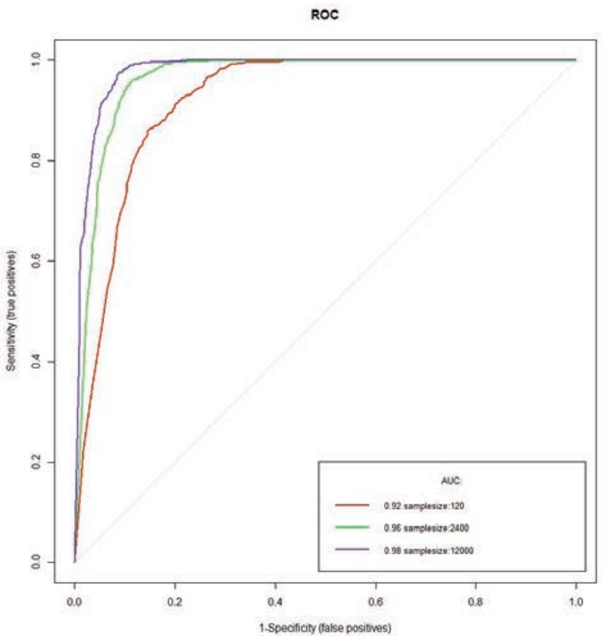

Figure 4. The ROC curves of the model with three different sample sizes.

(Hernandez *et al.*, 2006). There is a lack of studies about the sample size and sample ratio with ENM in disease-related field studies and we aimed to bridge this gap. If the sample size is not chosen wisely, ENM can produce incorrect results or cause waste of resources. In this study we explored both the impacts of sample size and sample ratio on RF model results built with a presence-absence schistosomiasis dataset to predict the potential distribution of snail habits. Study results can help to better predict potential snail habitats, so that schistosomiasis can be more efficiently and effectively prevented and controlled.

No significantly statistical difference was found among the sample ratios of 1:1, 1:2 and 1:3. This could be due to the small differences between these ratios or that we did not build sufficiently effective models. However, the sample ratios of 1:1, 1:2, and 1:3 were all significantly better than the ratio of 4:1 and 3:1, which suggests that a sample ratio of approximately 1:2 might produce better model performance. In addition, the results of the quantile deviation showed that model prediction at different sample sizes was most consistent when the sample ratio was either 1:2 or 1:3. This is in agreement with sensitivity, specificity and AUC results.

Furthermore, we found that 2,400 was the most suitable sample size in our study area, because the improvement of AUC became significantly smaller beyond this size and the sensitivity began to decrease. The decrease of slope of AUC with the increase in sample size suggests that an excessively large sample size should be discouraged. The reason is that the increasing survey costs would add little improvement in model prediction. This argument is also supported by the decreasing sensitivity beyond a sample size of 2,400, which may be more important than specificity when predicting snail habits for schistosomiasis prevention.

Compared with the sample ratio, the sample size appeared to be more important for RF models in our study. There was no significant difference between the sample ratio of 1:1, 1:2, and 1:3 for a given sample size. However, the optimal sample size is important because this determines the resources that will need to be invested in research projects.

In our study, we chose RF to explore the impacts of sample size and ratio on the prediction of snail habitats. This study provides a new approach for the selection of sample size and sample ratio for ENM, especially the models which use both presence data and absence data. Furthermore, it also provides a scientific basis for the selection of samples for the identification of snail habitats, so that resources can be planned and allocated more rationally. Still. other types of ENM are also worth studying in the future.

References

Bean WT, Stafford R, Brashares JS, 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. Ecography 35:250-8.

Cao ZG, Li S, Zhao YE, Wang TP, Bergquist R, Huang YY, Gao FH, Hu Y, Zhang ZJ, 2018. Spatio-temporal pattern of schistosomiasis in Anhui Province, East China: Potential effect of the Yangtze River - Huaihe River Water Transfer Project. Parasitol Int 67:538-46.

Chalghaf B, Chemkhi J, Mayala B, Harrabi M, Benie GB, Michael E, Ben Salah A, 2018. Ecological niche modeling predicting the potential distribution of Leishmania vectors in the Mediterranean basin: impact of climate change. Parasites Vectors 11:461.

Colley DG, Bustinduy AL, Secor WE, King CH, 2014. Human schistosomiasis. Lancet 383:2253-64.

Escobar LE, Craft ME, 2016. Advances and limitations of disease biogeography using ecological niche modeling. Front Microbiol 7:1174. eCollection 2016.

Guo JG, Vounatsou P, Cao CL, Utzinger J, Zhu HQ, Anderegg D, Zhu R, He ZY, Li D, Hu F, Chen MG, Tanner M, 2005. A geographic information and remote sensing based model for prediction of Oncomelania hupensis habitats in the Poyang Lake area, China. Acta Trop 96:213-22.

Global Land Information System (GLIS). Available from: https://pubs.er.usgs.gov/publication/78

Hernandez PA, Graham CH, Master LL, Albert DL, 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29:773-85.

Liu CR, White M, Newell G, 2011. Measuring and comparing the accuracy of species distribution models with presence-absence data. Ecography 34:232-43.

Pedersen UB, Stendel M, Midzi N, Mduluza T, Soko W, Stensgaard AS, Vennervald BJ, Mukaratirwa S, Kristensen TK, 2014. Modelling climate change impact on the spatial distribution of fresh water snails hosting trematodes in Zimbabwe. Parasites Vectors 7:536.

Peterson AT, Papes M, Eaton M, 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30:550-60.

Sage KM, Johnson TL, Teglas MB, Nieto NC, Schwan TG, 2017. Ecological niche modeling and distribution of Ornithodoros hermsi associated with tick-borne relapsing fever in western North America. PLoS Negl Trop Dis 11:e0006047.

Scholte RG, Carvalho OS, Malone JB, Utzinger J, Vounatsou P, 2012. Spatial distribution of Biomphalaria spp., the intermediate host snails of Schistosoma mansoni, in Brazil. Geospat Health 6:S95-101.

Stockwell DRB, Peterson AT, 2002. Effects of sample size on accuracy of species distribution models. Ecol Modell 148:1-13.

Thuiller W, Lafourcade B, Engler R, Araujo MB, 2009. BIOMOD - a platform for ensemble forecasting of species distributions. Ecography 32:369-73.

Warren DL, Seifert SN, 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. Ecol Appl 21:335-42.

WorldClim. Available from: https://www.worldclim.org/

World Wildlife Fund. Available from: https://www.worldwildlife.org/pages/conservation-science-data-and-tools

Xia C, Hu Y, Ward MP, Lynn H, Li S, Zhang J, Hu J, Xiao S, Lu C, Li S, Liu Y, Zhang Z, 2019. 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 13:e0007386.

Zhang J, Yue M, Hu Y, Bergquist R, Su C, Gao F, Cao ZG, Zhang Z, 2020. Risk prediction of two types of potential snail habitats in Anhui Province of China: Model-based approaches. PLoS Negl Trop Dis 14:e0008178.

Zhang Z, Jiang Q, 2011. Schistosomiasis elimination. Lancet Infect Dis 11:345-47.

Zhu G, Fan J, Peterson AT, 2017. Schistosoma japonicum transmission risk maps at present and under climate change in mainland China. PLoS Negl Trop Dis 11:e0006021.

Zhu HR, Liu L, Zhou XN, Yang GJ, 2015. Ecological Model to Predict Potential Habitats of Oncomelania hupensis, the Intermediate Host of Schistosoma japonicum in the Mountainous Regions, China. PLoS Negl Trop Dis 9:e000 4028.

Zou L, Ruan S, 2015. Schistosomiasis transmission and control in China. Acta Trop 143:51-7.

Online supplementary material:
Detailed environmental data collection and processing
Table S1. Summary of all 36 Environment Variables used in Study before Screening.
Table S2. Pair-wise comparisons of AUC for models built by different kinds of sample ratio using t tests.
Table S3. Quantiles of AUC for models built by different kinds of sample ratio in different levels of sample size.
Figure S1. PCC and Kappa.