Supplementary materials

Detailed environmental data collection and processing

Thirty-six environmental variables were involved in our study. Most of the environmental variables were remotely sensed from Earth-orbiting satellite sensors. We calculated the normalized difference vegetation index (NDVI) and the land surface temperature (LST) based on the Landsat 7 ETM+ images of our study area and obtained elevation data (DEM) from the Global Land Information System (GLIS) of the United States Geological Survey (USGS), then extracted aspect (Asp) and slope accordingly. The distance to nearest water body (Water) was calculated from water body data that were downloaded from Conservation Science Data Sets of World Wildlife Fund. The climatic variables were Bio3, Bio6, Bio8, Bio9 obtained from WorldClim. The other climatic variables and soil data, geomorphic type (Geo), land use type (Lucc), ecosystem type (Eco) and vegetation type (Veg) all came from the Data Center for Resources and Environmental Sciences of Chinese Academy of Sciences.

The snail habitat's locations were re-projected using a 100×100 m grid of matrix where a habitat was marked as '1' in the grid with the grid centre as its location, otherwise '0'. For the modelling, the re-located locations (grid centres) were used, not the actual location, since the map of the whole area was needed for the prediction. To make sure that the environmental raster data had the same geographical scope and same scale as the study area, we used the polygon of Anhui Province as the mask for all environmental data and then converted them into the raster image with the same scale.

To control the potential multi-collinearity among the environmental variables, correlation analysis was conducted for all climate raster images to gain the correlation coefficients for the matrices. One of the variables was excluded from every pair of variables with a correlation coefficient greater than 0.7. Which variables to be excluded depended on the results of the t test for the pair of variables with respect to the two groups (presence and absence), and the variable with the lower *P*-value would be preserved. We screened all variables given for the habitat region (Table S1)

Table S1 Summary of all 36 Environment Variables used in Study before Screening.

Data description	Label	Variable type
Normalized difference vegetation index	NDVI	Continuous
Land surface temperature	LST	Continuous
Digital elevation module	DEM	Continuous
Aspect	Asp	Continuous
Slope	Slope	Continuous
Distance to nearest water body	Water	Continuous
WorldClim	Bio1~19*	Continuous
Accumulated temperature beyond 0℃	Aat0	Continuous
Accumulated temperature beyond 10℃	Aat10	Continuous
Moisture index	Im	Continuous
Annual average precipitation	Pa	Continuous
Annual average temperature	Tadem	Continuous
Soil type	Soil	Categorical
Soil texture	Clay/sand/silt	Continuous
Geomorphic type	Geo	Categorical
Land use type	Lucc	Categorical
Ecosystem type	Eco	Categorical
Vegetation type	Veg	Categorical

*

Bio1: annual mean temperature;

Bio2: mean diurnal range;

Bio3: isothermality;

Bio4: temperature seasonality;

Bio5: max temperature of warmest month;

Bio6: min temperature of coldest month;

Bio7: temperature annual range;

Bio8: mean temperature of wettest quarter;

Bio9: mean temperature of driest quarter;

Bo10: mean temperature of warmest quarter;

Bio11: mean temperature of coldest quarter;

Bio12: annual precipitation;

Bio13: precipitation of wettest month;

Bio14: precipitation of driest month;

Bio15: precipitation seasonality;

Bio16: precipitation of wettest quarter;

Bio17: precipitation of driest quarter;

Bio18: precipitation of warmest quarter;

Bio19: precipitation of coldest quarter.

Table S2 Pair-wise comparisons of AUC for models built by different kinds of sample ratio using t tests

1. Total sample size of 100

	1:1	1:2	1:3	1:4	2:1	3:1
1:2	>0.999	-	-	-	-	-
1:3	>0.999	>0.999	-	-	-	-
1:4	0.978	0.188	0.925	-	-	-
2:1	0.188	0.010	0.137	>0.999	-	-
3:1	<0.001	<0.001	<0.001	<0.001	0.174	-
4:1	<0.001	<0.001	<0.001	<0.001	<0.001	0.925

2. Total sample size of 500

	1:1	1:2	1:3	1:4	2:1	3:1
1:2	0.233	-	-	-	-	-
1:3	0.162	0.939	-	-	-	-
1:4	0.939	0.939	0.719	-	-	-
2:1	0.162	<0.001	<0.001	<0.001	-	-
3:1	<0.001	<0.001	<0.001	<0.001	0.162	-
4:1	<0.001	<0.001	<0.001	<0.001	<0.001	0.925

3. Total sample size of 1,000

	1:1	1:2	1:3	1:4	2:1	3:1
1:2	>0.999	-	-	-	-	-
1:3	0.070	0.198	-	-	-	-
1:4	0.112	0.268	>0.999	-	-	-
2:1	<0.001	<0.001	<0.001	<0.001	-	-
3:1	<0.001	<0.001	<0.001	<0.001	0.174	-
4:1	<0.001	<0.001	<0.001	<0.001	<0.001	0.925

4. Total sample size of 5,000

	1:1	1:2	1:3	1:4	2:1	3:1
1:2	0.563	-	-	-	-	-
1:3	<0.001	>0.999	-	-	-	-
1:4	<0.001	<0.001	0.141	-	-	-
2:1	<0.001	<0.001	0.897	0.130	-	-
3:1	<0.001	<0.001	<0.001	0.641	0.579	-
4:1	<0.001	<0.001	<0.001	<0.001	<0.001	0.925

Table S3 Quantiles of AUC for models built by different kinds of sample ratio in different levels of sample size

Sample size		100			500			1,000			5,000		
		P25	median	P75	P25	median	P75	P25	median	P75	P25	median	P75
Sample ratio	1:1	0.8835	0.9152	0.9290	0.9485	0.9516	0.9549	0.9598	0.9625	0.9651	0.9834	0.9841	0.9850
	1:2	0.8936	0.9169	0.9281	0.9507	0.9547	0.9580	0.9612	0.9635	0.9657	0.9828	0.9837	0.9846
	1:3	0.8867	0.9139	0.9272	0.9506	0.9549	0.9592	0.9623	0.9649	0.9675	0.9824	0.9831	0.9839
	1:4	0.8754	0.9083	0.9242	0.9478	0.9546	0.9585	0.9627	0.9649	0.9680	0.9819	0.9826	0.9834
	2:1	0.8651	0.8924	0.9215	0.9442	0.9489	0.9530	0.9567	0.9599	0.9626	0.9822	0.9834	0.9843
	3:1	0.8572	0.8807	0.9114	0.9386	0.9461	0.9505	0.9555	0.9590	0.9626	0.9817	0.9831	0.9838
	4:1	0.8457	0.8715	0.9090	0.9332	0.9412	0.9461	0.9515	0.9553	0.9589	0.9790	0.9812	0.9826

Figure S1 PCC and Kappa