Received: 4 May 2022 | Revised: 28 September 2022 | Accepted: 6 November 2022

DOI: 10.1002/ldr.4527

RESEARCH ARTICLE

A high-accuracy vegetation restoration potential mapping model integrating similar habitat and machine learning

Xiaoyu Meng[1] | Huawei Pi[1] | Xin Gao[2] | Panxing He[3] | Jiaqiang Lei[2]

[1]Key Research Institute of Yellow River Civilization and Sustainable Development & Collaborative Innovation Center on Yellow River Civilization of Henan Province, Henan University, Kaifeng, PR China

[2]State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, PR China

[3]Ministry of Education Key Laboratory for Western Arid Region Grassland Resources and Ecology, College of Grassland Science, Xinjiang Agricultural University, Urumqi, PR China

Correspondence
Xin Gao, State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830000, PR China.
Email: gaoxin@ms.xjb.ac.cn

Funding information
The Strategic Priority Research Program of the Chinese Academy of Sciences, Grant/Award Number: XDA20030102; The National Natural Science Foundation of China, Grant/Award Number: 42107367

Abstract

Vegetation restoration potential (VRP) mapping provides important information for ecosystem restoration planning. However, the inappropriate assumption of traditional models that VRPs are identical within an individual similar habitat unit may result in low accuracy of VRP maps. This study proposes an improved data-driven model, namely, the similar habitat and machine learning-based VRP mapping (SHMLVRPM) model. This new model introduces a variety of machine-learning models to mine information on geographical environment heterogeneity in areas of similar habitat, which helps to improve the accuracy of VRP maps. Taking Yan'an City, Shanxi Province, China as our study area, we demonstrate the modelling process and validate the model. Our results show that the SHMLVRPM model can effectively construct high-accuracy VRP maps, and its information entropy is approximately 5.8 greater than that derived from the traditional models. The random forest method has the highest prediction accuracy ($R^2 = 0.8$) among the tested machine-learning methods. The average VRP value of Yan'an is approximately 68%; counties with the low VRP achievement are concentrated in the northern part of Yan'an, only 54%. Our research results can assist policymakers in optimizing vegetation restoration options and promoting the protection and sustainable development of fragile ecosystems.

KEYWORDS
machine learning, potential mapping, similar habitat, spatial analysis, vegetation restoration

1 | INTRODUCTION

Vegetation plays an important role in regulating the carbon cycle, absorbing greenhouse gases, and mitigating climate change (Emamian et al., 2021). Changes in vegetation coverage, especially in arid and semi-arid regions, have particularly important influences on land degradation, wind erosion, and dust emission (Duniway et al., 2019; Pi et al., 2021). As one of the most effective strategies for mitigating climate change and restoring the ecological environment, vegetation restoration is receiving increasing attention. Many countries and international organizations have actively formulated large-scale vegetation restoration programmes, projects and plans, such as the Grain-for-Green (GFG) Program in China (Feng et al., 2016), African Forest Landscape Restoration Initiative (AFR100) (Mansourian, 2021; Messinger & Winterbottom, 2016), and New York Declaration on Forests (NYDF) (Summit, 2021). The Intergovernmental Panel on Climate Change (IPCC) has stated that an additional 1 billion hectares of forest are needed to limit global warming to 1.5°C by 2050 (IPCC, 2018). However, it remains unclear whether these revegetation goals can be achieved without evaluating revegetation potential. Therefore, regional vegetation restoration potential (VRP) must be accurately assessed to formulate scientific restoration plans and improve the efficiency of ecosystem restoration (Bastin et al., 2019).

Vegetation type, scale, and structure are determined primarily by the geographical environment (Xu et al., 2020). Therefore, VRP is usually predicted by constructing a model that relates the geographical environment to vegetation information, based on a comprehensive understanding of regional geographical environment information. VRP mapping (VRPM) models fall into three categories, depending on how they are constructed: multi-factor comprehensive prediction models, similar habitat prediction models, and machine-learning models.

Multi-factor comprehensive prediction models calculate the comprehensive VRP by weighting the geographical environment factors related to vegetation growth (Arianoutsou et al., 2011; Bisson et al., 2008; Yan et al., 2014). Commonly used multi-factor comprehensive evaluation models include post-fire vegetation resilience index model (Bisson et al., 2008), multi-standard evaluation model (Arianoutsou et al., 2011), and analytic hierarchy process (AHP) model (Yan et al., 2014). The main deficiency of multi-factor comprehensive prediction models lies in the determination of the weights of the geographical environment factors. Factor weights are typically determined on the basis of expert knowledge, which is often highly subjective and considerably influenced by the research scale. Therefore, the prediction variability of multi-factor comprehensive prediction models is large (Zhang, Xu, et al., 2020).

A similar habitat prediction model predicts the VRP by determining the optimal vegetation status in areas of similar habitat (Bastin et al., 2019; Nauman et al., 2017; Xu et al., 2020; Zhang, Xu, et al., 2020). A massive amount of geographical information provides essential data for geographical modeling and lays the foundation for building high-quality prediction models. For example, remote sensing can quickly obtain large-scale vegetation status information (Ma et al., 2019; Shen et al., 2018), coupled with a large quantity of geographical environment element data sets (Abatzoglou et al., 2018), which provides a strong data basis for the construction of VRP prediction models (Bastin et al., 2019). Based on the geographical law whereby areas with similar geographical environments have similar geographical phenomena (Zhu, Liu, et al., 2015), some researchers have proposed a VRPM approach that combines multi-source remote sensing monitoring data with GIS spatial statistics methods to form a similar habitat-based VRP mapping (SHVRPM) model (Gao, Pang, et al., 2017). Because of its clear mechanism and ease of operation, the SHVRPM model has been widely used in VRPM in recent years (Lv et al., 2021; Xu et al., 2020; Zhang, Xu, et al., 2020). However, the SHVRPM model is a global model, and adapting the model to the global situation in areas with marked spatial heterogeneity is difficult. There are also limitations to researchers' understanding of the relationships between geographical variables, and bias in the model results is often increased by incomplete selection of geographical environment factors (Zhang, Jia, et al., 2019). To address these problems, Zhang, Xu, et al. (2020) and Xu et al. (2020) proposed a sliding-window-based similar habitat potential mapping model as an improvement on the SHVRPM model. Although the VRPM model based on similar habitats has been continuously improved, the assumption that the restoration potential of vegetation within areas of similar habitat is the same has not changed, which means that the accuracy of the similar habitat assumption determines the accuracy of the potential map. However, similar habitat areas are constructed by discretizing and superimposing geographical environmental factors, and similar habitat areas are usually presented in patches, resulting in the low accuracy of the potential map.

Machine learning has powerful nonlinear fitting capabilities (Meng, Gao, Li, et al., 2021) which can be used as a statistical method to describe the relationship between the geographical environment and vegetation information. Rich computing power and geographical big data make it possible to couple data-driven and machine-learning methods to considerably improve the accuracy of geographical prediction models. Machine-learning methods are often used to construct the relationship between biome distribution information and environmental factors and to predict potential natural vegetation research (Gutierres et al., 2018; Hengl et al., 2018; Raja et al., 2019). Hengl et al. (2018) showed that the prediction accuracy of the random forest method was higher than that using other methods, such as neural networks, gradient boosting, and k-nearest neighbours. Machine-learning methods are currently less applied in VRP prediction studies, probably because vegetation cover is strongly influenced by human activities, and there are no undisturbed vegetation cover data that can be used for modeling, especially when human activities are difficult to quantify. To solve this problem, Bastin et al. (2019) collected vegetation coverage data from global nature reserves and combined machine-learning methods to predict global forest restoration potential. This method cleverly excludes areas disturbed by human activities, obtaining more objective vegetation cover data and improving the prediction accuracy of the machine-learning model. However, the approach of Bastin et al. is not suitable for small-scale studies. This is because nature reserves tend to be limited in number and are nonuniformly distributed in small-scale areas; hence, sufficient sample data are often unavailable.

To solve the aforementioned problems of traditional models, the objective of this paper is to integrate similar habitat and machine-learning models to propose a new model for VRP prediction. Our model has at least two advantages: (1) by abandoning the assumption that VRP is identical in similar habitat areas (as per the traditional similar habitat model), we overcome the difficulty in obtaining undisturbed vegetation data when there is no nature reserve in the study area by separating areas of similar habitat and determining the best vegetation growth information within each habitat area; (2) based on the best vegetation information and corresponding environmental factors within each habitat area, the machine-learning method is used to predict the VRP within similar habitat areas, which achieves the purpose of improving the accuracy of the VRP map.

In the remainder of this paper, we first introduce the principle and modeling processes of the traditional similar habitat model and our model, respectively. We then take Yan'an City, northern China, as a case-study area to compare the two models. Finally, we evaluate the accuracy of the new model and discuss its advantages, disadvantages, and reliability. The main objectives of the present study are: (i) to construct a high-accuracy VRP mapping model, integrating similar habitat and machine learning; (ii) to evaluate the reliability of the new model by taking Yan'an City, northern China, as a case study; and (iii) to accurately assess the potential of regional vegetation restoration during the formulation of restoration plans.

2 | MODEL DESCRIPTION

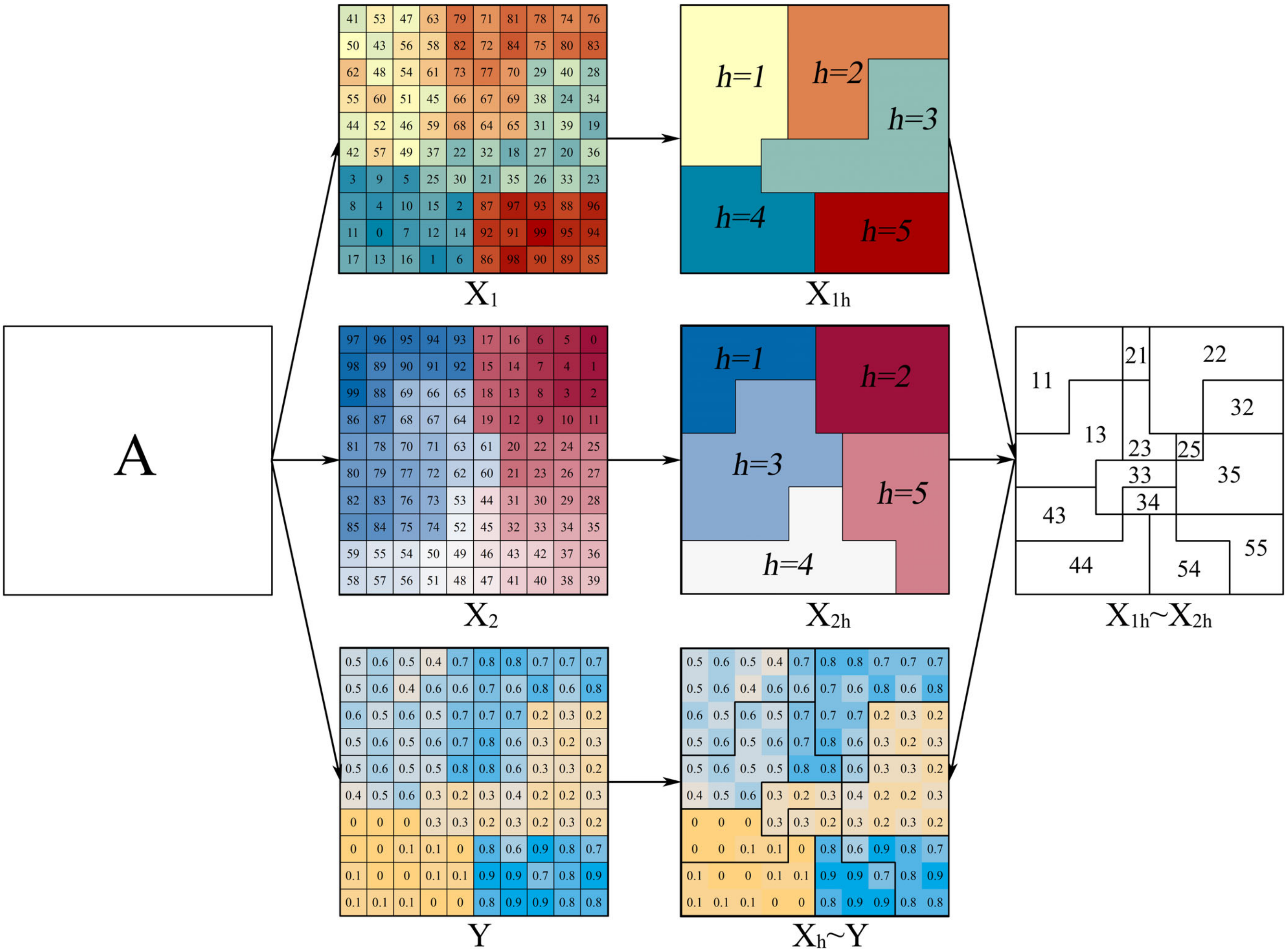
2.1 | Traditional similar habitat model

Similar geographical environments should possess similar geographical phenomena (Zhu, Liu, et al., 2015). If there is a significant difference in vegetation growth states between similar geographical environment areas, it can be considered that there is space for vegetation restoration. Therefore, the best state of vegetation in a similar geographical environment area can be defined as the VRP of the area. The VRP is generally expressed by numerical indicators, such as the resilience index (Bisson et al., 2008), vegetation index (Zhang, Xu, et al., 2020), and vegetation coverage (Bastin et al., 2019). In this study, vegetation coverage was used to characterize the VRP. The traditional SHVRPM model is shown in Figure 1. We set X_1 and X_2 as two layers of data representing environmental factors of study area A, and Y as the vegetation coverage data of A. The model first discretizes X_1 and X_2 into discrete data values X_{1h} and X_{2h}, respectively, via a discretization method. Second, the model superimposes the discrete results for the geographical environment factors to construct similar geographical environment areas $X_{1h} \sim X_{2h}$. Finally, the model superimposes the similar geographical environment areas and vegetation coverage Y, recorded as $X_h \sim Y$, and then traverses the similar geographical environment areas to determine the VRP of each area based on its maximum vegetation coverage value. In this paper, VRP refers to the possibility of future vegetation development, which is the level to which an area can be developed, rather than a spatial representation of potential (Xu et al., 2020; Zhang, Xu, et al., 2020). The VRP is expressed as follows:

$$P_{ij} = \max VC(E_1, E_2, E_3, \ldots, E_N), \qquad (1)$$

Where: P_{ij} is the VRP of the pixel in row i and column j. E_1, E_2, E_3, …, E_N are the discrete classes of each environmental factor, and N is the number of environmental factors. $VC(E_1, E_2, E_3, \ldots, E_N)$ is the vegetation coverage data in the same discrete classes as the pixel of row i and column j, and max VC is the maximum value of VC.

Spatial superposition of discrete factors can form different patchy areas, that is, similar habitat areas. The SHVRPM model assumes that the VRP is the same within similar habitat areas, which leads to the VRP map also being in the form of patches. The greatest flaw of the model is that it cannot identify the VRP differences for



X1

41	53	47	63	79	71	81	78	74	76
50	43	56	58	82	72	84	75	80	83
62	48	54	61	73	77	70	29	40	28
55	60	51	45	66	67	69	38	24	34
44	52	46	59	68	64	65	31	39	19
42	57	49	37	22	32	18	27	20	36
3	9	5	25	30	21	35	26	33	23
8	4	10	15	2	87	97	93	88	96
11	0	7	12	14	92	91	99	95	94
17	13	16	1	6	86	98	90	89	85

X2

97	96	95	94	93	17	16	6	5	0
98	89	90	91	92	15	14	7	4	1
99	88	69	66	65	18	13	8	3	2
86	87	68	67	64	19	12	9	10	11
81	78	70	71	63	61	20	22	24	25
80	79	77	72	62	60	21	23	26	27
82	83	76	73	53	44	31	30	29	28
85	84	75	74	52	45	32	33	34	35
59	55	54	50	49	46	43	42	37	36
58	57	56	51	48	47	41	40	38	39

Y

0.5	0.6	0.5	0.4	0.7	0.8	0.8	0.7	0.7	0.7
0.5	0.6	0.4	0.6	0.6	0.7	0.6	0.8	0.6	0.8
0.6	0.5	0.6	0.5	0.7	0.7	0.7	0.2	0.3	0.2
0.5	0.6	0.5	0.6	0.7	0.8	0.6	0.3	0.2	0.3
0.5	0.6	0.5	0.5	0.8	0.8	0.6	0.3	0.3	0.2
0.4	0.5	0.6	0.3	0.2	0.3	0.4	0.2	0.2	0.3
0	0	0	0.3	0.3	0.2	0.3	0.2	0.3	0.2
0	0	0.1	0.1	0	0.8	0.6	0.9	0.8	0.7
0.1	0	0	0.1	0.1	0.9	0.9	0.7	0.8	0.9
0.1	0.1	0.1	0	0	0.8	0.9	0.9	0.8	0.8

Xh~Y

0.5	0.6	0.5	0.4	0.7	0.8	0.8	0.7	0.7	0.7
0.5	0.6	0.4	0.6	0.6	0.7	0.6	0.8	0.6	0.8
0.6	0.5	0.6	0.5	0.7	0.7	0.7	0.2	0.3	0.2
0.5	0.6	0.5	0.6	0.7	0.8	0.6	0.3	0.2	0.3
0.5	0.6	0.5	0.5	0.8	0.8	0.6	0.3	0.3	0.2
0.4	0.5	0.6	0.3	0.2	0.3	0.4	0.2	0.2	0.3
0	0	0	0.3	0.3	0.2	0.3	0.2	0.3	0.2
0	0	0.1	0.1	0	0.8	0.6	0.9	0.8	0.7
0.1	0	0	0.1	0.1	0.9	0.9	0.7	0.8	0.9
0.1	0.1	0.1	0	0	0.8	0.9	0.9	0.8	0.8

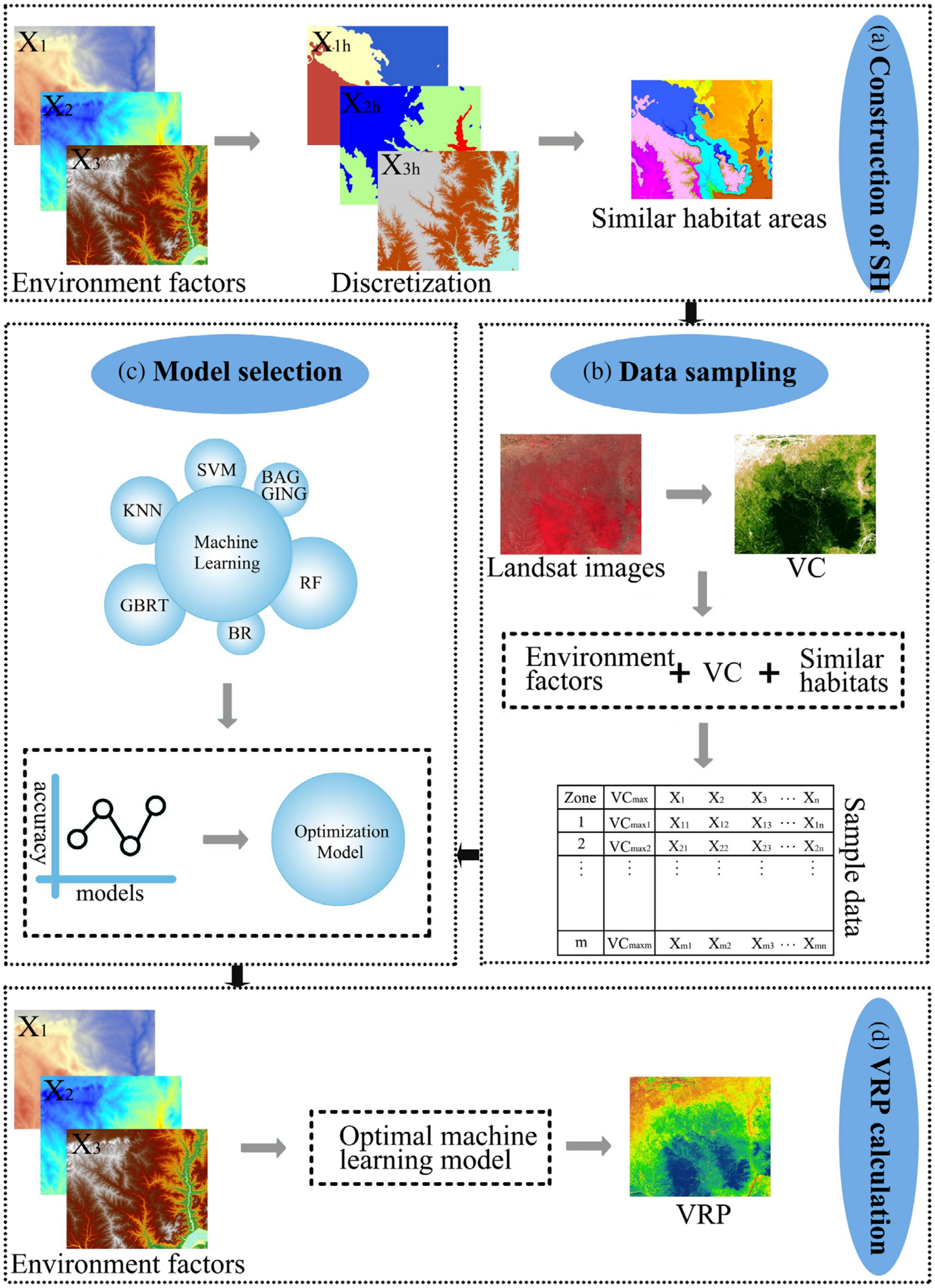
FIGURE 1 Schematic diagram of the vegetation restoration potential model based on similar habitat. [Colour figure can be viewed at wileyonlinelibrary.com]

decision-makers involved in vegetation restoration planning within similar habitat areas. The fundamental reason for this problem is that the discretization of geographical environmental factors is a process of information loss.

2.2 | VRPM model based on similar habitat and machine learning

The similar habitat and machine learning based on VRP mapping (SHMLVRPM) model proposed in this paper are based on habitat similarity and machine-learning methods. This new model discards the illogical assumption that VRP is the same within similar habitat areas and improves the accuracy of VRP prediction by mining the relationship between the heterogeneity of environmental factors and maximum vegetation coverage. The flowchart of the SHMLVRPM model is illustrated in Figure 2 and the modeling steps are as follows:

Step 1: Construction of similar habitat areas. First, data relating to the natural environment factors affecting vegetation growth, including meteorology, topography and soil, and so on, were collected (e.g., X_1, X_2, X_3 in Figure 2). Second, based on the principle of maximum similarity within class and minimum similarity between classes, the natural break method was used to discretize the continuous factor data (Jenks & Caspall, 1971; Meng, Gao, Lei, & Li, 2021) (e.g., X_{1h}, X_{2h}, X_{3h}



Zone	VC_{max}	X_1	X_2	X_3	⋯	X_n
1	VC_{max1}	X_{11}	X_{12}	X_{13}	⋯	X_{1n}
2	VC_{max2}	X_{21}	X_{22}	X_{23}	⋯	X_{2n}
⋮	⋮	⋮	⋮	⋮		⋮
m	VC_{maxm}	X_{m1}	X_{m2}	X_{m3}	⋯	X_{mn}

FIGURE 2 Flow chart of the SHMLVRPM model. (a) Overlaying the 'discretized' environmental factors to form similar habitat areas. (b) Extracting the maximum VC and environmental factors from each similar habitat area to form the sample data. (c) Training models and evaluating them to determine the optimal model. (d) Applying the optimal machine-learning model to calculate the VRP of the study area. SH, similar habitats; VC, vegetation coverage; VRP, vegetation restoration potential. [Colour figure can be viewed at wileyonlinelibrary.com]

in Figure 2). Finally, discrete factors were spatially overlain to obtain similar habitat areas, providing the basic statistical unit for the SHMLVRPM model.

Step 2: Data sampling. First, the modified soil-adjusted vegetation index (MSAVI) was calculated based on LANDSAT satellite images. Second, the vegetation cover was calculated using the pixel decomposition model. Finally, the maximum vegetation cover and the corresponding environmental data were extracted from each similar habitat area to form the sample data.

Step 3: Model selection. Sample data were trained using machine-learning methods, and the accuracy of the trained models was evaluated to determine the optimal model. To assess the accuracy of different machine-learning methods used in the SHMLVRPM model, six commonly used machine-learning methods were compared: random forest (RF), support vector machine (SVM), k-nearest neighbors (KNN), gradient boosting regressor tree (GBRT), Bayesian ridge (BR), and bootstrap aggregating (BAGGING).

Step 4: VRP calculation. The environmental data were substituted into the optimal machine-learning model to predict the VRP pixel by pixel.

The SHMLVRPM model can be expressed by the following equations:

$$M = ML[\max VC\,(E_1,\ E_2,\ E_3,\ \ldots,\ E_N),\ (X_1,\ X_2,\ X_3,\ \ldots,\ X_N)], \quad (2)$$

$$P_{ij} = M_{\mathrm{optimal}}\left(X_{1ij}, X_{2ij}, X_{3ij}, \ldots, X_{Nij}\right), \quad (3)$$

Where: M is the trained machine-learning model; ML denotes model training; max VC (E_1, E_2, E_3, …, E_N) is the maximum vegetation coverage value in a similar habitat area; X_1, X_2, X_3, …, X_N are the geographical environment factor values of the pixels with the same positions as those of the maximum vegetation coverage in each similar habitat area; N is the number of environmental factors; P_{ij} is the VRP value of the pixel in row i and column j. The value of P_{ij} can be obtained by substituting the environmental data X_{1ij}, X_{2ij}, X_{3ij}, …, X_{Nij} into the optimal machine-learning model M_{optimal}.

3 | CASE-STUDY

To further describe the modelling process of the SHMLVRPM model, this study takes Yan'an City, Shaanxi Province, China as a case-study.

3.1 | Study area

Yan'an City is located in the middle reaches of the Yellow River, north of Shanxi (35°21′–37°31′N, 107°41′–110°31′E), with a total area of 37,037 km^2 (Figure 3a). Yan'an is dominated by hills and ravines on the Loess Plateau (Xu et al., 2020), with undulating terrain, an elevation range of approximately 1500 m, and an average elevation of 1200 m (Zhang, Jia, et al., 2020). It has a monsoon climate, with an average annual precipitation (PREC) of approximately 500 mm and an average annual temperature (TEMP) of 9°C (Wen et al., 2021).

The main form of land degradation in Yan'an is soil erosion. The unique geographical location and topography lead to low PREC, high evaporation and low soil moisture, which intensify soil erosion; this soil erosion combines with transient heavy rainfall in summer to form gully landscapes with low vegetation cover (Figure 3b). Hilly/gully areas account for 39% of the land area of the City (Han et al., 2021). In addition, agriculture is the main economic activity of the local population, and soil erosion caused by the natural environment forces the local population to abandon low-fertility land (abandoned land) (Figure 3c) and reclaim new land, thus increasing vegetation destruction and soil erosion (Xu & Zhang, 2021). In order to protect and improve the ecological environment and promote the sustainable use of land, Yan'an launched a pilot large-scale Grain-for-Green (GFG) Program project in 1999. The main measures of the GFG Program were to ban grazing on hillsides and return cultivated land to forest and grass. After more than two decades of the GFG project, the vegetation coverage in Yan'an has increased by 50% (He et al., 2021; Wen et al., 2021). In this paper, VRP denotes the maximum vegetation coverage that can sustainably survive in the natural environment. Although vegetation restoration in Yan'an has been carried out via artificial planting, the restored vegetation can still survive without subsequent artificial irrigation. Therefore, the geographical environment of Yan'an can be considered to have the ability to restore the current vegetation coverage. After more than 20 years of the GFG Program, the current vegetation coverage in Yan'an should be closer to the estimated VRP. Therefore, Yan'an was selected as a typical research area for vegetation restoration, which helped to verify the results of the VRP prediction of our new model.

3.2 | Data sources and preprocessing

The VRP is the predicted maximum vegetation coverage that a natural geographical environment can support. Based on previous studies and the local natural geographical environment of Yan'an, vegetation coverage was selected to represent the VRP. We collected a set of nine geographical environment factors that may conceivably correlate with VRP. The factors can be approximately grouped into three categories: meteorology (PREC, TEMP, potential evapotranspiration (PET), and vapor pressure difference (VPD)), topography (elevation, slope, and aspect), and soil (soil type and soil moisture). Brief descriptions and sources of the factors are given in Table 1.

3.2.1 | Vegetation coverage

Vegetation coverage is the ratio of the vegetated area to the total land area, and the vegetated area is the projected area of plant stems and leaves on the ground (Carlson & Ripley, 1997; Fang et al., 2016). Vegetation coverage reflects the size of the photosynthetic area of

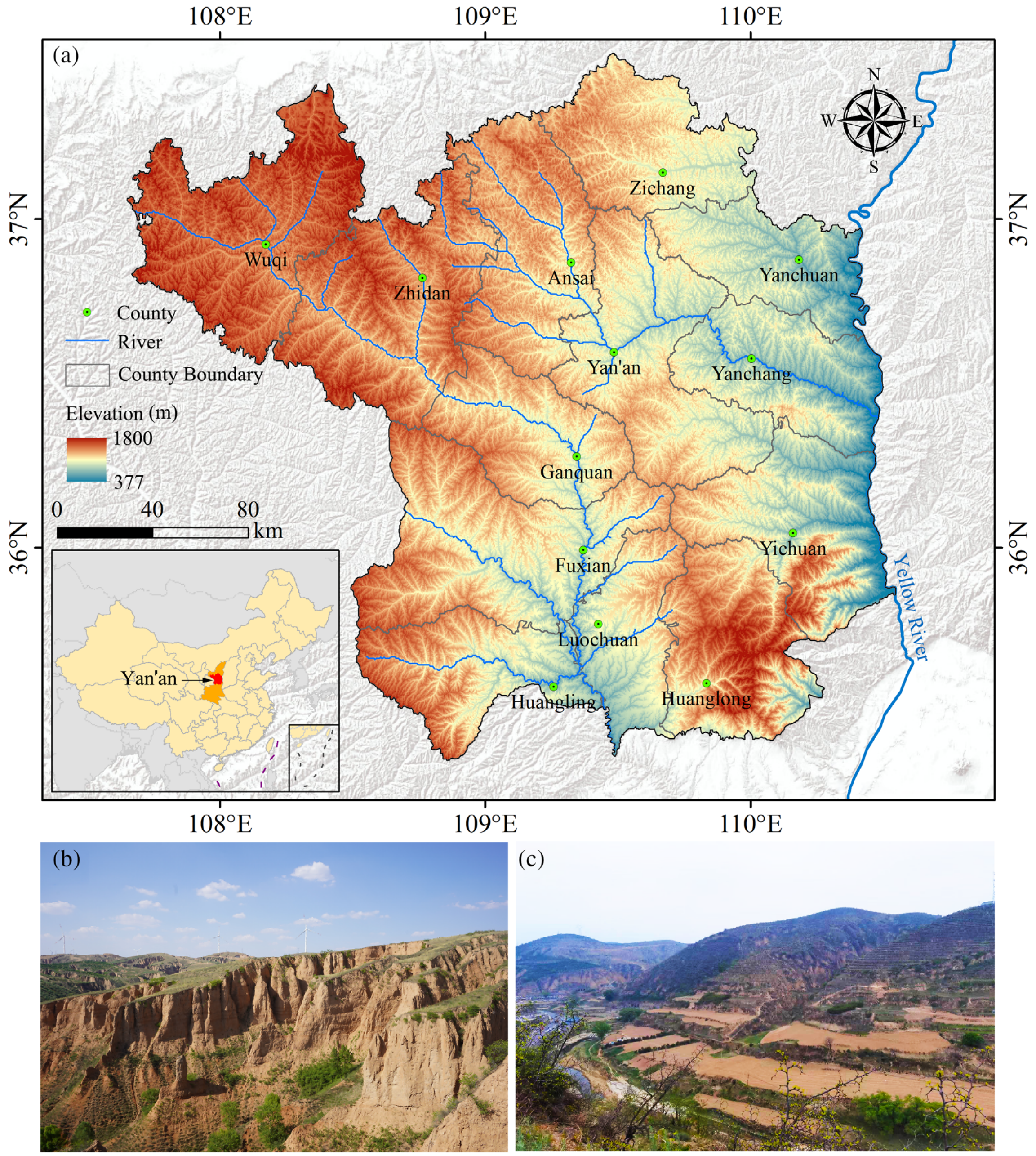


FIGURE 3 Location and landscapes of Yan'an City. (a) Geographical map. (b) Photograph of typical gully landforms. (c) Photograph of typical abandoned land. Elevation data were obtained from the Shuttle Radar Topography Mission V3 (https://www2.jpl.nasa.gov/srtm/). Wiley acknowledges that the borders within the figure are subject to multiple territorial claims. [Colour figure can be viewed at wileyonlinelibrary.com]

vegetation and characterizes the density of vegetation growth (Gao, Gao, et al., 2017; Wen et al., 2013). In this study, vegetation coverage was selected as the characterization index for VRP because it has more definite biophysical characteristics than the vegetation index (Zhang, Xu, et al., 2020). Vegetation coverage is calculated using the pixel decomposition model (Qi et al., 2000; Wittich & Hansing, 1995; Zhang, Chen, et al., 2019). The pixel decomposition model assumes that a mixed pixel consists of two elements, vegetation and bare soil, and that the measured signal for each pixel is a linear combination of the spectral features of ground objects. Therefore, the measured signal can be used to invert vegetation coverage directly on the mixed pixel. The vegetation coverage (VC) can be calculated using the following equation:

$$VC = \frac{VI - VI_s}{VI_v - VI_s}, \tag{4}$$

Where: VI is the vegetation index of each pixel in the study area, VI_s is the vegetation index of a pure bare soil pixel, and VI_v is the vegetation index of a pure vegetation pixel. In this study, the minimum VI value in the study area was used to represent the VI_s of a pure bare soil pixel, and the maximum VI value was used to represent the VI_v of a

TABLE 1 Geographical environment factors selected for modeling.

Category	Factor	Factor code	Resolution	Source/URL
Meteorology	Precipitation	PREC	4 km, interpolated to 30 m	TerraClimate dataset, https://www.climatologylab.org/terraclimate.html
	Temperature	TEMP	4 km, interpolated to 30 m	TerraClimate dataset
	Potential evapotranspiration	PET	4 km, interpolated to 30 m	TerraClimate dataset
	Vapour pressure difference	VPD	4 km, interpolated to 30 m	TerraClimate dataset
Topography	Elevation	ELEV	30 m	Derived from digital elevation model (DEM), Shuttle Radar Topography Mission V3, https://www2.jpl.nasa.gov/srtm/
	Slope	SLOPE	30 m	Derived from DEM
	Aspect	ASPECT	30 m	Derived from DEM
Soil data	Soil type	ST	Vector data, converted into 30 m raster	(Shi et al., 2004)
	Soil moisture	SM	4 km, interpolated to 30 m	TerraClimate dataset

pure vegetation pixel (Anees et al., 2022; Maselli et al., 2014). VI is usually represented by the normalized difference vegetation index (NDVI) (Anees et al., 2022; Gao et al., 2020; Zhang, Chen, et al., 2019), but numerous studies have shown that the modified soil-adjusted vegetation index (MSAVI) is more accurate than NDVI in estimating VC (Fang et al., 2016; Wiesmair et al., 2016; Younes et al., 2019), especially in areas with sparse vegetation. This is because MSAVI considers the soil background and adjusts the spectral influence of the soil to exclude the influence on the vegetation index (Qi et al., 1994), which is beneficial in improving the VC estimation accuracy. The MSAVI is calculated as follows:

$$\mathrm{MSAVI} = \frac{2\mathrm{NIR}+1-\sqrt{(2\mathrm{NIR}+1)^2-8(\mathrm{NIR}-\mathrm{RED})}}{2}, \quad (5)$$

Where: NIR and RED are the reflectance in the near-infrared band and the red band in the satellite image, respectively.

LANDSAT satellite image data from 1998 to 2020 were used to calculate MSAVI. The images were all obtained from the United States Geological Survey (https://www.usgs.gov/) and were acquired and calculated using the Google Earth Engine cloud computing platform. The LANDSAT images used were real surface reflectance data products, meaning that the data were preprocessed by systematic radiation correction and atmospheric correction and were suitable for direct use for surface information extraction (Pekel et al., 2016; Zhou et al., 2019; Zou et al., 2018; Zurqani et al., 2018). At the same time, after correction between different sensors, the LANDSAT reflectivity dataset can be used to analyze ground objects across different sensors (Huang et al., 2021). For remote sensing image preprocessing, we used the CFmask method to remove cloud pollution pixels (Frantz et al., 2018; Zhu, Wang, & Woodcock, 2015). The MSAVI data of all transit satellite images in Yan'an from 1998 to 2020 were calculated using Equation (5). To eliminate the influence of large disturbances and sudden changes in meteorological factors, the vegetation coverage of each year was synthesized by the 95% quantile method, and the vegetation coverage data from 1998 to 2020 were also synthesized using the 95% quantile method. The vegetation coverage of Yan'an was obtained using Equation 4 (Figure 4).

3.2.2 | Meteorological data

Meteorological data, including PREC, TEMP, PET, and VPD, were obtained from the TerraClimate dataset (Abatzoglou et al., 2018). TerraClimate is a global climate dataset with high spatial resolution, providing climate data on a monthly scale from 1958 to 2019 with a pixel size of approximately 4 km (1/24°). To facilitate the calculation, the spline function interpolation method was used to downscale the meteorological data, so that the spatial resolutions of the meteorological data and vegetation coverage data were the same. The spatial distributions of the mean annual meteorological factors from 1998 to 2020 are shown in Figure 5a–d.

3.2.3 | Topography and soil data

We selected elevation (ELEV) (Figure 3a), slope (SLOPE) (Figure 5e), and aspect (ASPECT) (Figure 5f) as topographic factors that may affect vegetation growth. All topographic data were derived from a digital elevation model (DEM). The DEM data in this study were obtained from the Shuttle Radar Topography Mission V3 (SRTM Plus) at a resolution of 1 arc-second (approximately 30 m) (Farr et al., 2007). The SRTM dataset is of high quality, with a horizontal and vertical standard error of approximately 1 and 0 m, respectively (Jarvis et al., 2008).

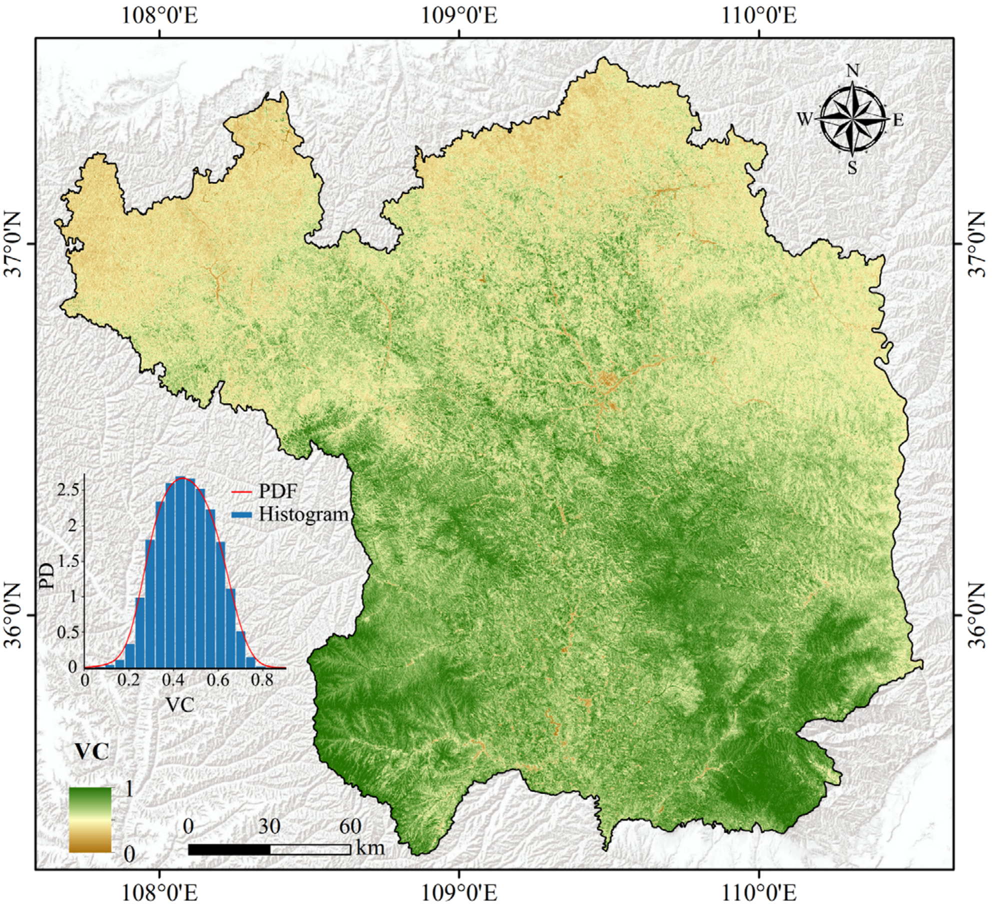


FIGURE 4 Distribution of vegetation coverage in Yan'an. PD, probability density; PDF, probability density function; VC, vegetation coverage. [Colour figure can be viewed at wileyonlinelibrary.com]

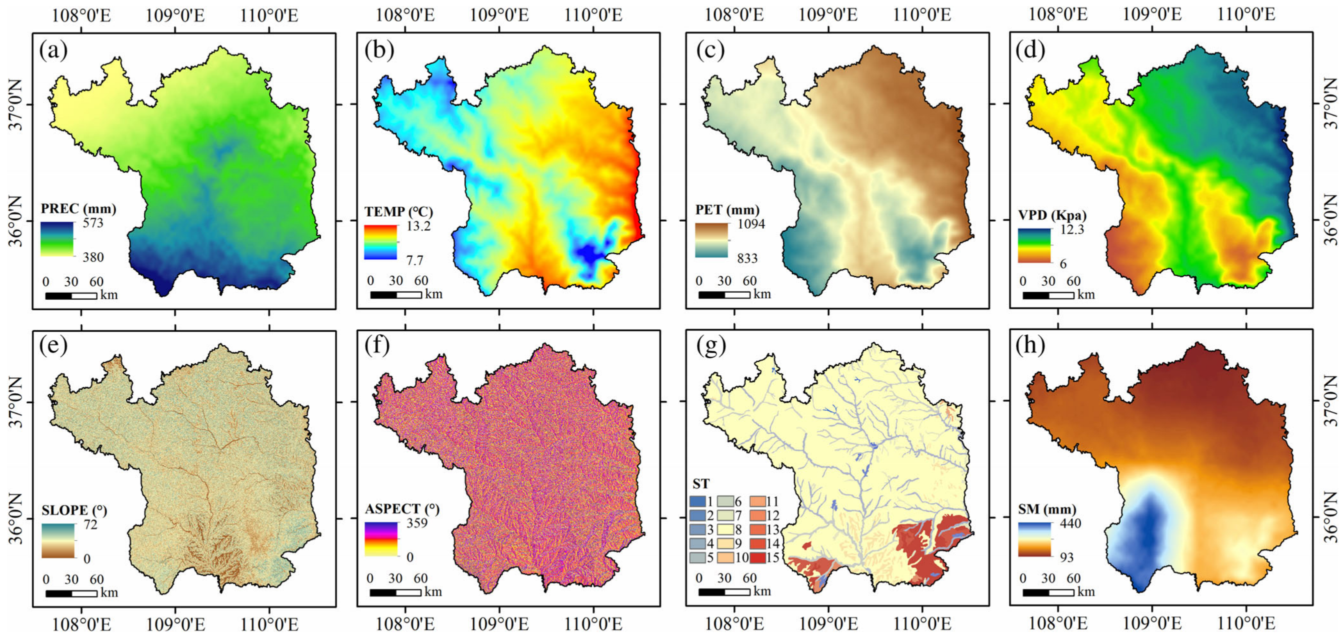


FIGURE 5 Spatial distribution maps of meteorological, topographic, and soil data across Yan'an. (a) Precipitation (PREC). (b) Air temperature (TEMP). (c) Potential evapotranspiration (PET). (d) Vapour pressure difference (VPD). (e) Slope. (f) Aspect. (g) Soil type (ST). (h) Soil moisture (SM). [Colour figure can be viewed at wileyonlinelibrary.com]

The soil type data (ST) were derived from the 1:1,000,000 scale vector soil type dataset of the Institute of Soil Sciences of the Chinese Academy of Sciences (Shi et al., 2004). The data were converted into a raster dataset with a spatial resolution of 30 m. Fifteen soil types have been identified in Yan'an: (1) cinnamon soil, (2) calcareous cinnamon soil, (3) lou soil, (4) developed cinnamon soil, (5) gray cinnamonic

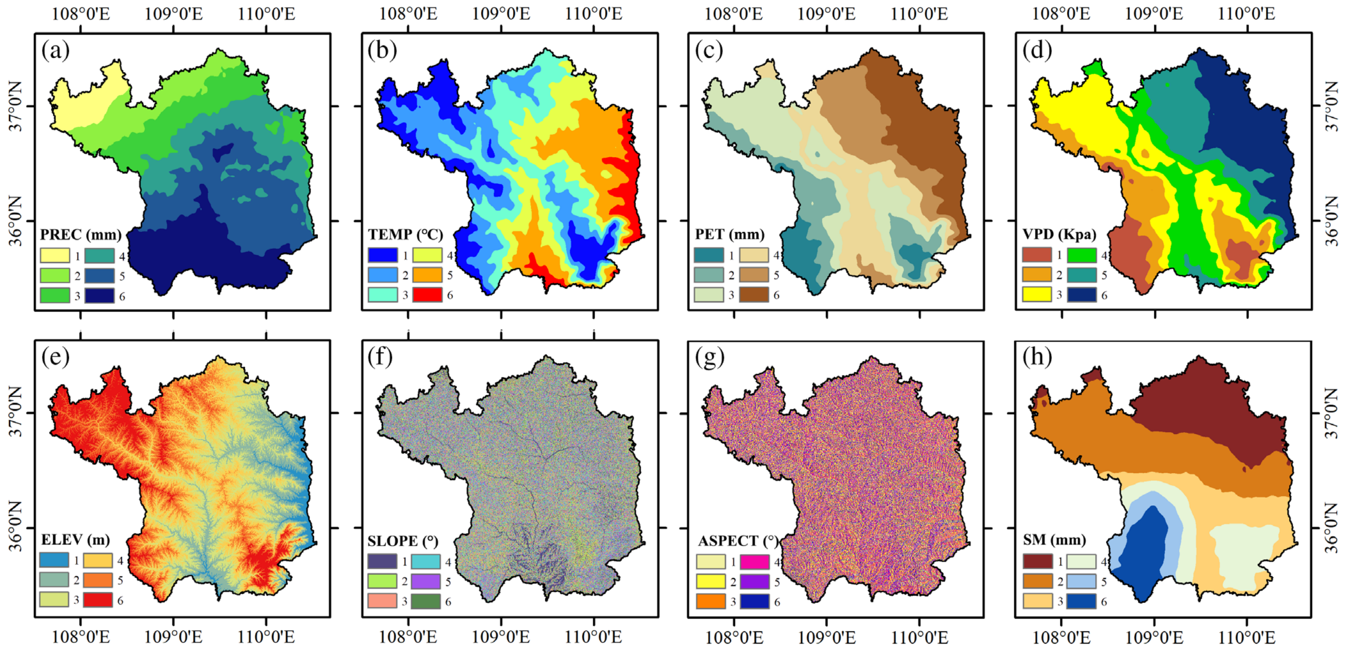


FIGURE 6 Spatial distribution diagrams of the discretization of continuous geographical environment factors. (a) Precipitation (PREC). (b) Air temperature (TEMP). (c) Potential evapotranspiration (PET). (d) Vapour pressure difference (VPD). (e) Elevation (ELEV). (f) Slope. (g) Aspect. (h) Soil moisture (SM). [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Cut point results of the discretization of continuous geographical environment factors.

	'Discretization' results					
Factors	1	2	3	4	5	6
Precipitation (mm)	<442	442–469	469–488	488–506	506–525	>525
Temperature (°C)	<9.0	9.0–9.6	9.6–10.2	10.2–10.8	10.8–11.5	>11.5
Potential evapotranspiration (mm)	<895	895–932	932–965	965–1003	1003–1038	>1038
Vapour pressure difference (mm)	<7.3	7.3–8.2	8.2–9.0	9.0–9.9	9.9–10.8	>10.8
Elevation (m)	<845	845–1027	1027–1168	1168–1310	1310–1458	>1458
Slope (°)	<8	8–15	15–21	21–26	26–33	>33
Aspect (°)	<59	59–118	118–180	180–240	240–259	>259
Soil moisture (mm)	<127	127–170	170–218	218–277	277–349	>349

soil, (6) black soil, (7) meadow chernozemic soil, (8) loessal soil, (9) red clay soil, (10) alluvial soil, (11) atteration soil, (12) chisley soil, (13) skeleton soil, (14) calcicregosols soil, and (15) rock. The soil type raster layer in Yan'an is shown in Figure 5g. Soil moisture data (SM) were also obtained from the TerraClimate dataset. The mean annual soil moisture spatial distribution from 1998 to 2020 is shown in Figure 5h.

4 | RESULTS

4.1 | Construction of similar habitat areas

Discretization of environmental factors is the first step in data preprocessing for constructing similar habitat areas. Figure 6 and Table 2 show the results of the discretization of the continuous geographical environment factors for Yan'an. All environmental factors are discretized into six classes. Here, we take the meteorological factors as an example for a brief description and analysis. The PREC discretization results vary along the northwest–southeast direction, where the zones with PREC <488 mm are mainly located in the northwest, with an area of approximately 12,324 km^2. The zones with PREC ≥488 mm have an area of about 24,676 km^2 and are concentrated in the south of Yan'an. From the results of TEMP discretization, medium–high TEMP zones (≥10.2°C) are mainly distributed along the Yellow River in the east and the valley area in the south, and medium–low TEMP zones (<10.2°C) are mainly distributed across the western and southern mountainous areas, with areas of 21,796 and 15,204 km^2, respectively. The spatial distribution patterns of PET and VPD have high a similarity with TEMP. In general, Yan'an has more areas with medium–high PREC and TEMP, covering 62% of the city. However, notably, the spatial distributions of PREC and TEMP in Yan'an City show obvious inconsistencies. Particularly in the eastern part of Yan'an, the temperature, PET and VPD are high, while PREC is only moderate, which indicates that the climatic conditions in the east are not favorable for vegetation growth.

Discrete factors were used to construct similar habitat areas. Any pixel in the habitat area layer contains nine attributes, which correspond to the discrete numbers of environmental factors. Here, we used a 10-digit code to represent similar habitat areas. Digits 1–8 represent the class numbers of the continuous factors, and the 9th and 10th digits represented the soil type code. The entire study area could theoretically form up to $6^8 \times 15 = 25{,}194{,}240$ similar habitat areas. However, because of the strong collinearity between environmental factors (e.g., TEMP, ET, and VPD), the actual number of different similar habitat areas in Yan'an is only 76,307.

4.2 | Accuracy of different machine-learning models

The sample dataset was divided into five equal parts, and a five-fold cross-validation method was used to establish the training and testing data sets for use in testing the prediction accuracy of the six machine-learning models. The R^2 and root mean square error (RMSE) were used to characterize the model accuracy. The results show that the RF method yields the highest accuracy ($R^2 = 0.80$, RMSE = 0.036), followed by the BAGGING ($R^2 = 0.75$, RMSE = 0.042), GBRT ($R^2 = 0.62$, RMSE = 0.051), KNN ($R^2 = 0.62$, RMSE = 0.051), BR ($R^2 = 0.39$, RMSE = 0.065), and SVM ($R^2 = 0.35$, RMSE = 0.067). Figure 7 shows the consistency between the true and predicted values of the testing data for the different machine learning methods. The red points in the figure are the true values, and the colored pixels are the two-dimensional (2D) histogram. The pixel value refers to the number of predicted values in pixels. It can be seen from the figure that the range of values predicted by the RF (Figure 7a) is narrow and concentrated near the true values, whereas the ranges of values predicted by the KNN and BR methods are more dispersed. In general, the RF and BAGGING methods perform best among all methods, whereas the accuracies of the GBRT, KNN, BR, and SVM methods are relatively low. However, even the RF method with the highest prediction accuracy has relatively limited performance. This may be caused by the large differences in sample data between similar habitat areas; in other words, large differences in vegetation cover may occur in geographic environments with small differences influenced by human activities or other anomalous disturbances, which can easily lead to lower model accuracy when this phenomenon occurs within different habitat areas. In addition, although machine learning strategies are considered black box solutions and lack interpretability, machine learning can provide more accurate and informative maps of VRP than do the unreasonable simplistic assumptions in traditional models. Therefore, we used the RF model to predict the VRP.

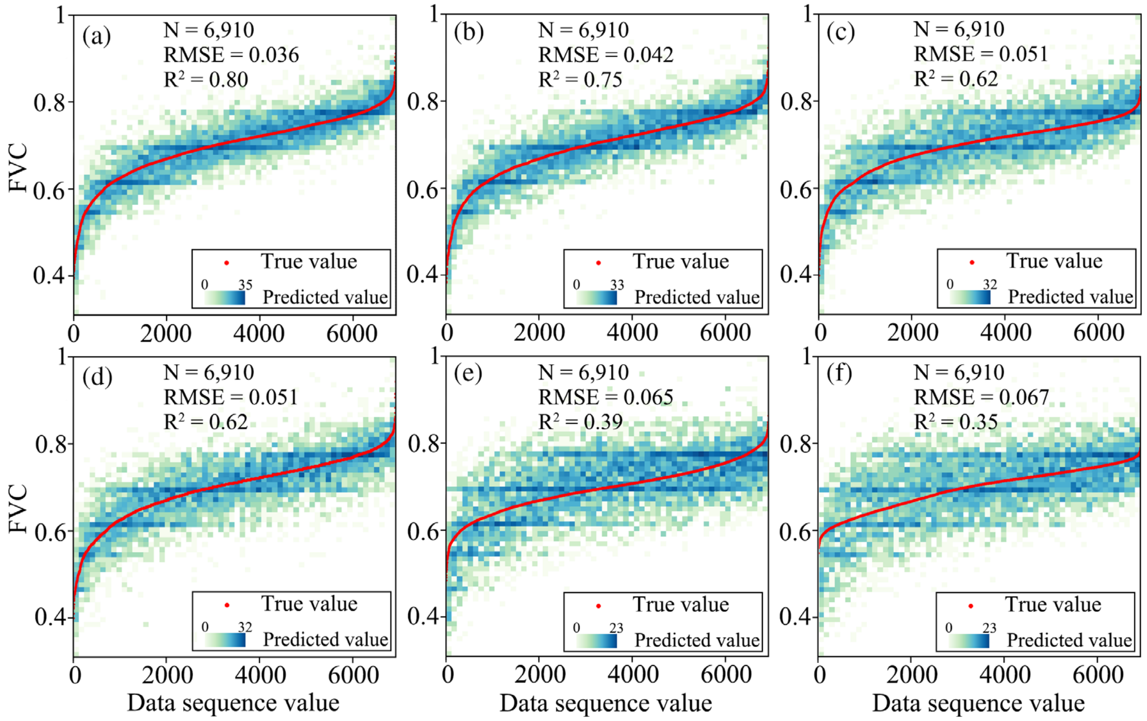


FIGURE 7 Accuracy of different machine-learning models. (a) Random forest. (b) Bootstrap aggregating. (c) Gradient boosting regressor tree. (d) K-nearest neighbors. (e) Bayesian ridge. (f) Support vector machine. N, sample size; RMSE, root mean square error; R^2, correlation coefficient. [Colour figure can be viewed at wileyonlinelibrary.com]

4.3 | VRP comparison of SHMLVRPM and SHVRPM models

Figure 8a, b shows the VRP results for Yan'an obtained with the SHMLVRPM and SHVRPM models, respectively. The VRP map of the SHVRPM model is visibly patchy, whereas the VRP map of the SHMLVRPM model is more spatially continuous. To further express the differences between the VRP maps obtained from the two models, we calculated the VRP information entropy within each similar habitat area. Information entropy is often used as a quantitative indicator of the information content of a system, with higher values of information entropy indicating greater information content; conversely, an information entropy of zero implies complete consistency of information within the statistical aggregate (Gray, 2011; Xia et al., 2021). Our results show that the average information entropy of the VRP map of the SHMLVRPM is 5.8, which is much higher than that of the SHVRPM model (because the SHVRPM model assumes the same VRP within the similarity habitat areas, the information entropy of the VRP map is 0). Therefore, the information content of the VRP map obtained from the SHMLVRPM model is richer. In addition, the VRP of the SHMLVRPM model is more in line with the continuous distribution characteristics of the geographical environment (Figure 8c, d). Here the VRP of the SHMLVRPM model uses the maximum vegetation coverage in the habitat area and the corresponding environmental data as sample data, revealing a nonlinear relationship between the geographical environment and the vegetation coverage. However, owing to inappropriate assumptions, the SHVRPM model fails to fully exploit the subtle relationship between environmental factors and vegetation coverage within habitat areas. Although the discretization process of both models leads to information loss in similar habitat areas, which also exaggerates spatial homogeneity and ignores spatial heterogeneity, the SHMLVRPM model uses a machine learning approach to compensate for the information loss caused by the discretization process.

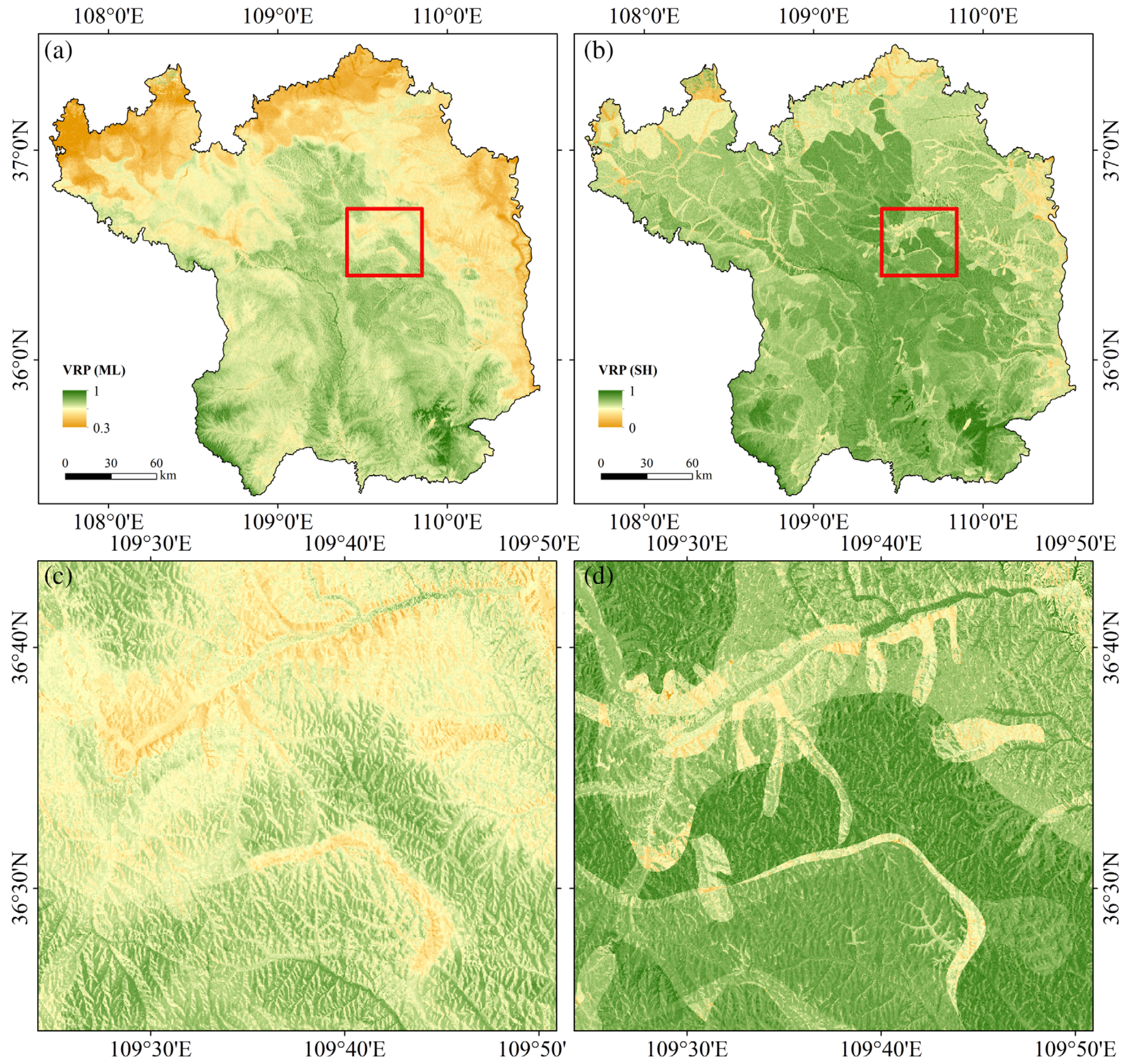


FIGURE 8 VRP of different models. (a) SHMLVRPM. (b) SHVRPM. (c) Local enlarged map of the SHMLVRPM model. (d) Local enlarged map of the SHVRPM model. VRP, vegetation restoration potential; ML: Machine-learning model (our model); SH: Similar habitat model (traditional model). [Colour figure can be viewed at wileyonlinelibrary.com]

4.4 | Spatial distribution characteristics of VRP

The VRP is the performance of the ecological carrying capacity of the natural geographical environment of a region. A small value of VRP means that the area is not suitable for vegetation growth. The VRP results for Yan'an predicted by the SHMLVRPM model (Figure 8a) indicate that regions with higher VRP values are located in the southern mountains, whereas the northern and eastern regions (near to the Yellow River) have lower VRP values. To further analyze the spatial distribution characteristics of VRP, we calculated the mean and standard deviation of VRP (Table 3) and analyzed the reasons for its value in each county of Yan'an. It can be seen from Table 3 that the VRP of most counties in Yan'an is ≥0.6, and half of the values are ≥0.7, which means that Yan'an generally has a high VRP. Huanglong County has the highest VRP (up to 0.75); this is mainly because the average annual PREC is relatively large, the elevation is relatively high, and the slope difference is relatively large, leading to unsuitable conditions for farming (corresponding to less interference from human activities and almost no land degradation). Therefore, the present high vegetation coverage of Huanglong contributes to a high VRP value for the whole county. The VRPs of Wuqi, Zichang, Ansai, and Zhidan Counties are relatively low. This is because these counties are located in the middle of the Loess Plateau and therefore have relatively low PREC and mainly herbaceous land cover (Zhang et al., 2006). In the eastern part of Yan'an, low VRP values are mainly distributed along the Yellow River, where VRP values are affected by the depressed topography. The TEMP, PET and VPD values in this area are high, but the PREC value is low.

5 | DISCUSSION

5.1 | Model reliability verification

As described in Section 4, we analyzed the spatial distributions of VRP obtained using different models. Our results were highly consistent with those of previous studies on the spatial distribution of VRP. For example, the VRP in the north of Yan'an is lower than that in the south (Xu et al., 2020; Zhang, Xu, et al., 2020), which demonstrates the reliability of our new model to a certain extent. Over the course of 20-year GFG Program (1999 to date), the vegetation coverage should have approached that reflected by the VRP values. We compared the consistency between vegetation coverage and VRP in 1998, and in2020, to further evaluate the accuracy of the SHMLVRPM model. Here, we argue that agreement between vegetation restoration rates and VRP is not sufficient to ensure model accuracy (Zhang, Xu, et al., 2020). The reason for this is that vegetation restoration rates are not only related to regional natural resource characteristics but also to the process of the GFG Program.

TABLE 3 Vegetation restoration potential of counties in Yan'an.

County	Mean	SD	Area (km^2)	PREC (mm)
Huanglong	0.749	0.049	2751	529
Luochuan	0.730	0.038	1792	533
Ganquan	0.729	0.040	2276	511
Huangling	0.728	0.051	2291	544
Fuxian	0.723	0.040	4175	522
Baota	0.709	0.046	3536	510
Yichuan	0.691	0.055	2939	511
Ansai	0.677	0.068	2951	481
Zhidan	0.673	0.044	3789	473
Yanchang	0.644	0.050	2362	501
Yanchuan	0.621	0.050	1986	490
Zichang	0.611	0.058	2393	474
Wuqi	0.593	0.075	3789	435

Abbreviations: PREC, precipitation; SD, standard deviation.

We calculated the vegetation coverage of Yan'an in 1998 and 2020, randomly generated a series of verification points in the study area, and extracted the vegetation coverage and VRP values corresponding to each verification point in 1998 and 2020. Figure 9 shows a scatter plot of the vegetation coverage and VRP data. The abscissa is the VRP, and the ordinate is the vegetation coverage. The vegetation coverage and VRP deviate from the 1:1 line, and the R^2 is only 0.37 for 1998 data. This means that before the implementation of the GFG Program, there was a significant difference between the vegetation coverage and VRP, and vegetation coverage was not close to equaling the VRP. In 2020 however, the vegetation coverage and VRP were closer to the 1:1 line, and R^2 reached 0.47. This indicates that after the GFG Program, the vegetation coverage of Yan'an in 2020 was closer to the VRP; the effect of the implementation of an ecological restoration policy has been remarkable. This result has been confirmed in a number of studies of vegetation cover change (Wang et al., 2019; Zhi et al., 2019). Figure 9 shows that the coefficient of the linear fitting formula between vegetation coverage and VRP is 1.1 and the overall scatter plot is at the lower right of the 1:1 trend, which indicates that the areas that do not reach VRP mainly occur in the north central part of Yan'an. Wang et al. (2019) showed that the northern part of Yan'an was the main area targeted for vegetation restoration during completion of the project, although the vegetation cover in north–central Yan'an was still lower than that in the south even after the GFG Program. In the next vegetation restoration plan, the north–central area should remain the priority area for restoration (Zhang, Xu, et al., 2020). This is mainly because the northern area was historically disturbed to a greater extent more by human activities (Wang et al., 2019) and still has a greater recovery potential even after 20 years of the GFG Program.

5.2 | Vegetation restoration potential achievement in Yan'an

The ratio of the actual VC value to the VRP value is defined as the degree of VRP achievement (VRPA) (Xu et al., 2020). This section discusses the spatial distribution characteristics of VRPA and its relation to different environmental factors. A lower VRPA indicates that the vegetation has more scope to recover, and also indicates that the land degradation caused by human activities is more serious

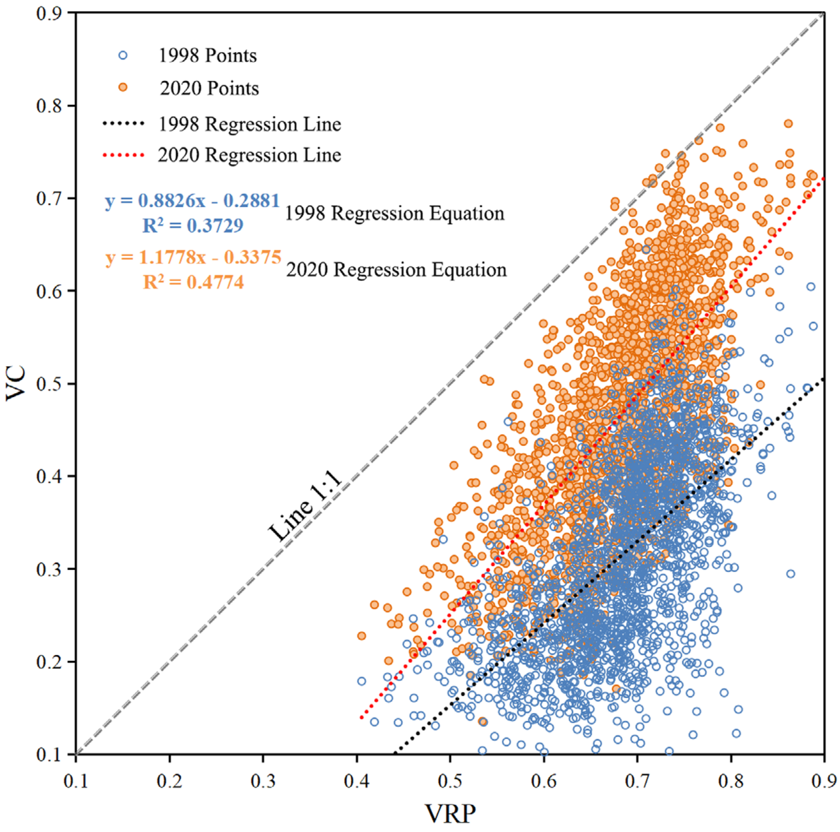


FIGURE 9 Scatter diagram of vegetation coverage (VC) and VRP before (1998) and after (2020) the GFG Program. [Colour figure can be viewed at wileyonlinelibrary.com]

(Yin & Yin, 2010). A higher VRPA means that anthropogenic disturbance is relatively small and the vegetation is close to saturation point under the current resource endowment conditions. Figure 10 shows the VRPA of Yan'an; areas with lower VRPA are mainly distributed in the north, whereas those with highter VRPA are distributed in the south. The counties with the lowest VRPA are Zichang (54%), Wuqi (55%), and Ansai (59%), and the counties with the highest VRPA are Fuxian (73%), Huangling (74%), and Huanglong (76%). Vegetation recovery in Yan'an is closely related to the distribution of PREC (Sun et al., 2015). High PREC in southern Yan'an determines its high vegetation cover (Zhi et al., 2019), which also explains its larger restoration achievement. Based on MODIS land cover data (University of Maryland (UMD) classification), the mean VRPA values corresponding to different land cover types in Yan'an were extracted. These were as follows: permanent wetlands: 26%; urban and built-up lands: 46%; grasslands: 60%; croplands: 68%; woody savannas: 72%; savannas: 72%; cropland/natural vegetation mosaics: 72%; mixed forests: 73%; closed shrublands: 78%; and deciduous broadleaf forests: 80%. It can be seen that the VRPA of permanent wetlands is the lowest, which is because wetlands can provide sufficient water for vegetation to grow and therefore have excellent recovery conditions. The VRPA values of urban and built-up lands rank second lowest because urbanization expansion leads to serious damage to vegetation under favorable geographical conditions. For example, the VRPA values in urban areas of Yan'an are observably low, as shown in Figure 10. In terms of land cover type, broadleaf forests have the highest VRPA values, because there is almost no vegetation destruction in the southern mountains of Yan'an. The VRPA in agricultural areas is at a mid-level. This result is similar to the work of Wang et al. (2019), who found that returning farmland to forest had a low impact on vegetation restoration in Yan'an, accounting for only 15% of the total restoration. Therefore, the focus should be on wetland, urban areas, and grassland in future vegetation restoration schemes. In terms of soil-types, alluvial soils have the lowest VRPA (63%). Alluvial soils are mostly located in valleys with relatively good water and heat conditions, and can be used for afforestation and farmland. It can also be seen from Figure 10 that higher VRPA values are mostly located in the river valley area, which is distributed in a long, narrow belt. The VRPA of red clay soils is also relatively low (65%). Red clay soils evolved under a hot and humid climate where annual PREC was greater than evaporation. It can be seen from Figure 5g that the red clay soils in Yan'an are mostly located in the upper reaches of the river valley.

In summary, an analysis of the differences in VRPA values for different environmental factors types shows that the key areas for future vegetation restoration in Yan'an are mainly concentrated in the northern region, among which low-lying valleys and wetlands are the most suitable targets. Although the VRPA of urban building land is high, restoration work will be limited to further improving the area of urban greening and landscape construction because of the needs of urbanization development.

5.3 | Model advantages

For both the SHVRPM and SHMLVRPM models, the criteria for similar habitats are ambiguous—there is no clear standard to assess

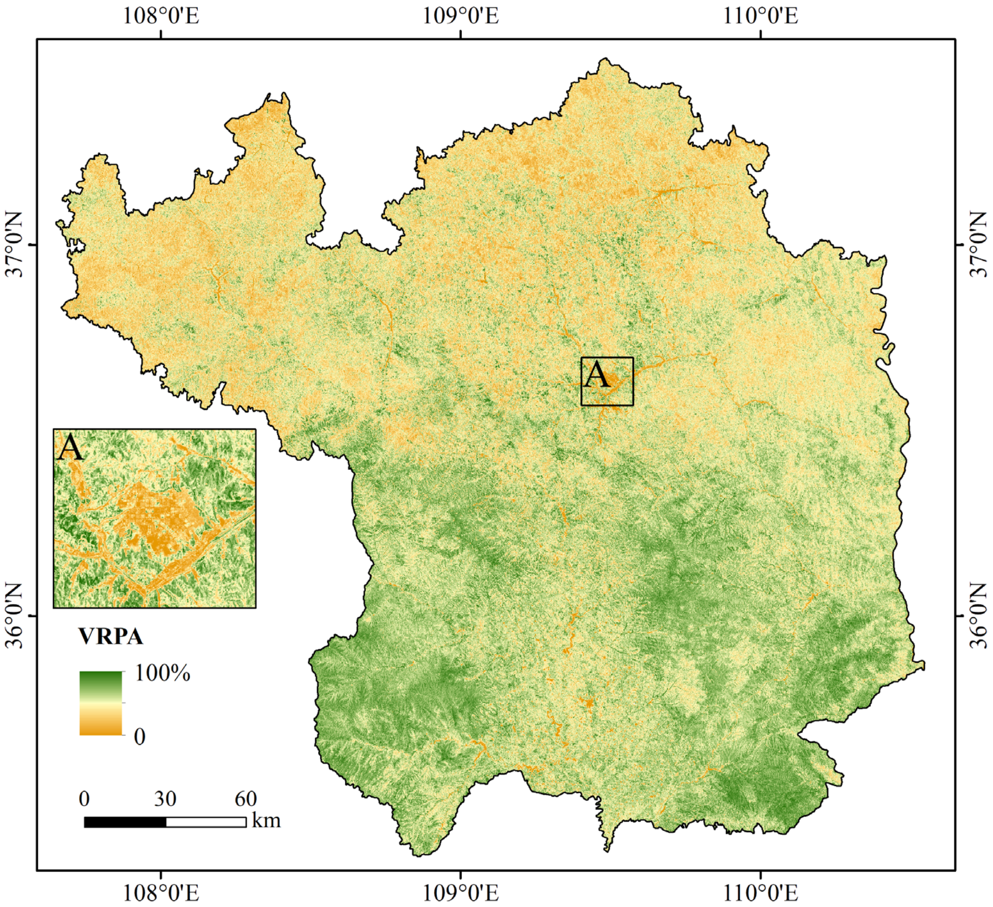


FIGURE 10 Vegetation restoration achievement (VRA) in Yan'an. Region a is Yan'an City. [Colour figure can be viewed at wileyonlinelibrary.com]

whether two geographical environments are sufficiently similar to be considered similar habitats. Therefore, the use of different discretization methods can be rather confusing. Discretization refers to dividing continuous numerical variable data into a specified number of intervals using clustering (Tsai & Chen, 2019). Commonly used discretization methods in geography include the equal interval method (Cao et al., 2013), natural breaks method (Shrestha & Luo, 2017), quantile method (Luo et al., 2016), geometric interval method (Tian et al., 2017), and discretization based on expert experience (Du et al., 2017). In previous research applications of the VRPM model, a number of the discretization methods mentioned above have been used to identify similar habitats (Lv et al., 2021; Zhang, Xu, et al., 2020). However, there is still no unified standard for the selection of discretization methods for geographical environmental factors, and different discretization methods have been shown to produce different results for the identification of similar habitat areas, which directly affects the accuracy of the VRP. In this study, the natural breaks method was chosen as the method for discretizing environmental factors. This is mainly because the discretization principle of the natural breaks method best fits the concept of similar habitats (Meng, Gao, Lei, & Li, 2021). In terms of the environmental factors themselves, the natural breaks method is the optimal discretization method to delineate the similar habitat areas. In addition, the SHVRPM model assumes that the VRP values of similar habitat areas are the same, and information mining of geographical environment heterogeneity in similar habitat areas is not sufficient, which also directly reduces the spatial resolution of the potential map. The VRP is determined by topographic, climatic, soil and geological conditions. However, the contribution of each of these factors to vegetation growth may vary with location because of the heterogeneity of geographical environments within similar habitat areas. This suggests that even within an individual similar habitat unit, the potential for vegetation growth may vary (Zhang, Xu, et al., 2020). The SHMLVRPM model developed in this study is based on the similar habitat theory and uses the natural breaks method to discretize geographical environmental factors to identify relatively similar habitat areas. We integrate the maximum vegetation coverage and the corresponding geographical environment factor data in each habitat area, and use machine-learning methods to construct nonlinear relationships between variables. Therefore, the new model can overcome or at least attenuate the adverse effects of spatial heterogeneity, thereby improving the accuracy of VRP measurements.

5.4 | Uncertainties of the model

The vegetation distribution is usually affected by both the natural geographical environment and anthropogenic activities, resulting in a mismatch between the vegetation distribution and the geographical environment carrying capacity. For example, deforestation leads to a reduction in vegetation coverage in high-quality geographical environments (Bastin et al., 2019). Desert agriculture leads to a marked increase in vegetation coverage in harsh geographical environments (Meng, Gao, Li, et al., 2021). The vegetation coverage of the former does not reach the regional environmental carrying capacity, and the vegetation coverage of the latter far exceeds the environmental carrying capacity, which leads to biased VRP according to similar habitat-based models. Therefore, in the actual restoration process, if the estimated potential value is used as the restoration target, the restoration cost is high, and the sustainability is poor. In the future, methods that can quantify the potential biases and uncertainties mentioned above should be designed to provide more comprehensive data support for vegetation restoration work.

In addition, the VRP derived from the SHMLVRPM model was still patchy in some areas, as in the case of Wuqi County, northwest of Yan'an (Figure 8a). This may be caused by the maximum vegetation cover values within the similar habitat areas being significantly different. If the same independent variables data correspond to different dependent variable values, the accuracy of the trained model will be reduced (Maino et al., 2022). Therefore, future research should focus on the delineation of similar habitat areas and the scientific determination of optimal vegetation cover within the habitat areas.

6 | CONCLUSIONS

In this paper, we propose a new high-accuracy VRPM model based on the similar habitat theory. This new model divides the geographical environment into homogeneous regions and integrates the highest value of vegetation coverage and geographical environment data for each region into sample data. It then uses machine learning to construct the relationship between geographical environment and vegetation coverage to improve the accuracy of the potential map. Our case study results show that the potential map obtained using the new model has higher accuracy than that obtained using the traditional models. The potential map is more consistent with the spatial distribution of geographical environment variables, which provide necessary data support for the scientific planning of vegetation restoration projects. Vegetation restoration is an important path to achieving goal 15.3 of the United Nations Sustainable Development Goals and the land degradation neutrality target proposed by the United Nations Convention to Combat Desertification. The ability to forecast VRP with high accuracy will contribute to the scientific formulation of recovery planning.

ACKNOWLEDGMENTS

This research was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA20030102) and the National Natural Science Foundation of China (No. 42107367).

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Xiaoyu Meng https://orcid.org/0000-0001-9372-0899
Huawei Pi https://orcid.org/0000-0002-3877-6213

REFERENCES

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., & Hegewisch, K. C. (2018). TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Scientific Data*, *5*, 1–12. https://doi.org/10.1038/sdata.2017.191

Anees, S. A., Zhang, X., Shakeel, M., Al-Kahtani, M. A., Khan, K. A., Akram, M., & Ghramh, H. A. (2022). Estimation of fractional vegetation cover dynamics based on satellite remote sensing in Pakistan: A comprehensive study on the FVC and its drivers. *Journal of King Saud University—Science*, *34*, 1–7. https://doi.org/10.1016/j.jksus.2022.101848

Arianoutsou, M., Koukoulas, S., & Kazanis, D. (2011). Evaluating post-fire forest resilience using GIS and multi-criteria analysis: An example from Cape Sounion National Park, Greece. *Environmental Management*, *47*, 384–397. https://doi.org/10.1007/s00267-011-9614-7

Bastin, J.-F., Finegold, Y., Garcia, C., Mollicone, D., Rezende, M., Routh, D., Zohner, C. M., & Crowther, T. W. (2019). The global tree restoration potential. *Science*, *365*, 76–79. https://doi.org/10.1126/science.aax0848

Bisson, M., Fornaciai, A., Coli, A., Mazzarini, F., & Pareschi, M. T. (2008). The vegetation resilience after fire (VRAF) index: Development, implementation and an illustration from central Italy. *International Journal of Applied Earth Observation and Geoinformation*, *10*, 312–329. https://doi.org/10.1016/j.jag.2007.12.003

Cao, F., Ge, Y., & Wang, J.-F. (2013). Optimal discretization for geographical detectors-based risk assessment. *GIScience & Remote Sensing*, *50*, 78–92. https://doi.org/10.1080/15481603.2013.778562

Carlson, T. N., & Ripley, D. A. (1997). On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote Sensing of Environment*, *62*, 241–252. https://doi.org/10.1016/S0034-4257(97)00104-1

Du, Z., Zhang, X., Xu, X., Zhang, H., Wu, Z., & Pang, J. (2017). Quantifying influences of physiographic factors on temperate dryland vegetation, Northwest China. *Scientific Reports*, *7*, 1–9. https://doi.org/10.1038/srep40092

Duniway, M. C., Pfennigwerth, A. A., Fick, S. E., Nauman, T. W., Belnap, J., & Barger, N. N. (2019). Wind erosion and dust from US drylands: A review of causes, consequences, and solutions in a changing world. *Ecosphere*, *10*, 1–28. https://doi.org/10.1002/ecs2.2650

Emamian, A., Rashki, A., Kaskaoutis, D. G., Gholami, A., Opp, C., & Middleton, N. (2021). Assessing vegetation restoration potential under different land uses and climatic classes in Northeast Iran. *Ecological Indicators*, *122*, 1–13. https://doi.org/10.1016/j.ecolind.2020.107325

Fang, S., Tang, W., Peng, Y., Gong, Y., Dai, C., Chai, R., & Liu, K. (2016). Remote estimation of vegetation fraction and flower fraction in oilseed rape with unmanned aerial vehicle data. *Remote Sensing*, *8*, 416. https://doi.org/10.3390/rs8050416

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., & Alsdorf, D. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, *45*, 1–33. https://doi.org/10.1029/2005RG000183

Feng, X., Fu, B., Piao, S., Wang, S., Ciais, P., Zeng, Z., Lü, Y., Zeng, Y., Li, Y., Jiang, X., & Wu, B. (2016). Revegetation in China's loess plateau is approaching sustainable water resource limits. *Nature Climate Change*, *6*, 1019–1022. https://doi.org/10.1038/nclimate3092

Frantz, D., Haß, E., Uhl, A., Stoffels, J., & Hill, J. (2018). Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects. *Remote Sensing of Environment*, *215*, 471–481. https://doi.org/10.1016/j.rse.2018.04.046

Gao, Y., Gao, J., Wang, J., Wang, S., Li, Q., Zhai, S., & Zhou, Y. (2017). Estimating the biomass of unevenly distributed aquatic vegetation in a lake using the normalized water-adjusted vegetation index and scale transformation method. *Science of the Total Environment*, *601–602*, 998–1007. https://doi.org/10.1016/j.scitotenv.2017.05.163

Gao, D., Pang, G., Li, Z., & Cheng, S. (2017). Evaluating the potential of vegetation restoration in the Loess Plateau. *Acta Geographica Sinica*, *72*, 863–874. https://doi.org/10.11821/dlxb201705008

Gao, L., Wang, X., Johnson, B. A., Tian, Q., Wang, Y., Verrelst, J., Mu, X., & Gu, X. (2020). Remote sensing algorithms for estimation of fractional vegetation cover using pure vegetation index values: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *159*, 364–377. https://doi.org/10.1016/j.isprsjprs.2019.11.018

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media.

Gutierres, F., Gomes, P., Rocha, J., & Teodoro, A. C. (2018). Spatially explicit models in local dynamics analysis: The potential natural vegetation (PNV) as a tool for beach and coastal management. In C. M. Botero, O. Cervantes, & C. W. Finkl (Eds.), *Beach management tools—concepts, methodologies and case studies* (pp. 159–177). Cham, CH: Springer International Publishing. https://doi.org/10.1007/978-3-319-58304-4_8

Han, L., Huo, H., Liu, Z., Zhao, Y.-H., Zhu, H.-L., Chen, R., & Zhao, Z.-L. (2021). Spatial and temporal variations of vegetation coverage in the middle section of Yellow River basin based on terrain gradient:Taking Yan'an City as an example. *The Journal of Applied Ecology*, *32*, 1581–1592. https://doi.org/10.13287/j.1001-9332.202105.014

He, Z., Shang, X., & Zhang, T. (2021). Spatiotemporal evaluation and driving mechanism of land ecological security in Yan'an, a typical hill-gully region of China's Loess Plateau, from 2000 to 2018. *Forests*, *12*, 1–21. https://doi.org/10.3390/f12121754

Hengl, T., Walsh, M. G., Sanderman, J., Wheeler, I., Harrison, S. P., & Prentice, I. C. (2018). Global mapping of potential natural vegetation: An assessment of machine learning algorithms for estimating land potential. *PeerJ*, *6*, 1–36. https://doi.org/10.7717/peerj.5457

Huang, W., Duan, W., Nover, D., Sahu, N., & Chen, Y. (2021). An integrated assessment of surface water dynamics in the Irtysh River basin during 1990–2019 and exploratory factor analyses. *Journal of Hydrology*, *593*, 1–15. https://doi.org/10.1016/j.jhydrol.2020.125905

IPCC. (2018). *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels andrelated global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* (pp. 1–616). Cambridge, UK and New York, NY: Cambridge University Press.

Jarvis, A., Guevara, E., Reuter, H. I., & Nelson, A. D. (2008). Hole-filled SRTM for the globe: version 4. CGIAR Consortium for Spatial Information SRTM 90m Database. https://srtm.csi.cgiar.org

Jenks, G. F., & Caspall, F. C. (1971). Error on Choroplethic maps: Definition, measurement, reduction. *Annals of the Association of American Geographers*, *61*, 217–244. https://doi.org/10.1111/j.1467-8306.1971.tb00779.x

Luo, W., Jasiewicz, J., Stepinski, T., Wang, J., Xu, C., & Cang, X. (2016). Spatial association between dissection density and environmental factors over the entire conterminous United States. *Geophysical Research Letters*, *43*, 692–700. https://doi.org/10.1002/2015GL066941

Lv, Z., Li, S., Fan, J., Liu, G., Wang, H., & Meng, X. (2021). Natural restoration potential of vegetation in Mongolia. *Journal of Desert Research*, *41*, 192–201. https://doi.org/10.7522/j.issn.1000-694X.2021.00047

Ma, J., Xiao, X., Miao, R., Li, Y., Chen, B., Zhang, Y., & Zhao, B. (2019). Trends and controls of terrestrial gross primary productivity of China during 2000–2016. *Environmental Research Letters*, *14*, 1–14. https://doi.org/10.1088/1748-9326/ab31e4

Maino, A., Alberi, M., Anceschi, E., Chiarelli, E., Cicala, L., Colonna, T., De Cesare, M., Guastaldi, E., Lopane, N., Mantovani, F., Marcialis, M., Martini, N., Montuschi, M., Piccioli, S., Raptis, K. G. C., Russo, A., Semenza, F., & Strati, V. (2022). Airborne radiometric surveys and machine learning algorithms for revealing soil texture. *Remote Sensing*, *14*, 1–16. https://doi.org/10.3390/rs14153814

Mansourian, S. (2021). *Review of forest and landscape restoration in Africa 2021*. FAO and AUDA-NEPAD. https://doi.org/10.4060/cb6111en

Maselli, F., Papale, D., Chiesi, M., Matteucci, G., Angeli, L., Raschi, A., & Seufert, G. (2014). Operational monitoring of daily evapotranspiration by the combination of MODIS NDVI and ground meteorological data: Application and evaluation in Central Italy. *Remote Sensing of Environment*, *152*, 279–290. https://doi.org/10.1016/j.rse.2014.06.021

Meng, X., Gao, X., Lei, J., & Li, S. (2021). Development of a multiscale discretization method for the geographical detector model. *International Journal of Geographical Information Science*, *35*, 1650–1675. https://doi.org/10.1080/13658816.2021.1884686

Meng, X., Gao, X., Li, S., Li, S., & Lei, J. (2021). Monitoring desertification in Mongolia based on Landsat images and GoogleEarth Engine from 1990 to 2020. *Ecological Indicators*, *129*, 1–15. https://doi.org/10.1016/j.ecolind.2021.107908

Messinger, J., & Winterbottom, B. (2016). African forest landscape restoration initiative (AFR100): Restoring 100 million hectares of degraded and deforested land in Africa. *Nature & Fauna*, *30*, 14–17. https://www.fao.org/3/i5992e/i5992e.pdf

Nauman, T. W., Duniway, M. C., Villarreal, M. L., & Poitras, T. B. (2017). Disturbance automated reference toolset (DART): Assessing patterns in ecological recovery from energy development on the Colorado Plateau. *Science of the Total Environment*, *584–585*, 476–488. https://doi.org/10.1016/j.scitotenv.2017.01.034

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, *540*, 418–422. https://doi.org/10.1038/nature20584

Pi, H., Huggins, D. R., & Sharratt, B. (2021). Wind erosion of soil influenced by clay amendment in the inland Pacific northwest, USA. *Land Degradation & Development*, *32*, 241–255. https://doi.org/10.1002/ldr.3709

Qi, J., Chehbouni, A., Huete, A. R., Kerr, Y. H., & Sorooshian, S. (1994). A modified soil adjusted vegetation index. *Remote Sensing of Environment*, *48*, 119–126. https://doi.org/10.1016/0034-4257(94)90134-1

Qi, J., Marsett, R. C., Moran, M. S., Goodrich, D. C., Heilman, P., Kerr, Y. H., Dedieu, G., Chehbouni, A., & Zhang, X. X. (2000). Spatial and temporal dynamics of vegetation in the San Pedro River basin area. *Agricultural and Forest Meteorology*, *105*, 55–68. https://doi.org/10.1016/S0168-1923(00)00195-7

Raja, N. B., Aydin, O., Çiçek, İ., & Türkoğlu, N. (2019). A reconstruction of Turkey's potential natural vegetation using climate indicators. *Journal of Forestry Research*, *30*, 2199–2211. https://doi.org/10.1007/s11676-018-0855-7

Shen, X., An, R., Feng, L., Ye, N., Zhu, L., & Li, M. (2018). Vegetation changes in the Three-River headwaters region of the Tibetan Plateau of China. *Ecological Indicators*, *93*, 804–812. https://doi.org/10.1016/j.ecolind.2018.05.065

Shi, X. Z., Yu, D. S., Warner, E. D., Pan, X. Z., Petersen, G. W., Gong, Z. G., & Weindorf, D. C. (2004). Soil database of 1:1,000,000 digital soil survey and reference system of the Chinese genetic soil classification system. *Soil Survey Horizons*, *45*, 129–136. https://doi.org/10.2136/sh2004.4.0129

Shrestha, A., & Luo, W. (2017). An assessment of groundwater contamination in Central Valley Aquifer, California using geodetector method. *Annals of GIS*, *23*, 149–166. https://doi.org/10.1080/19475683.2017.1346707

Summit, U. C. (2021). *New York declaration on forests*. New York, NY: Forest Declaration Platform. Retrieved from https://forestdeclaration.org/

Sun, W., Song, X., Mu, X., Gao, P., Wang, F., & Zhao, G. (2015). Spatiotemporal vegetation cover variations associated with climate change and ecological restoration in the Loess Plateau. *Agricultural and Forest Meteorology*, *209–210*, 87–99. https://doi.org/10.1016/j.agrformet.2015.05.002

Tian, L., Li, Y., Yan, Y., & Wang, B. (2017). Measuring urban sprawl and exploring the role planning plays: A Shanghai case study. *Land Use Policy*, *67*, 426–435. https://doi.org/10.1016/j.landusepol.2017.06.002

Tsai, C.-F., & Chen, Y.-C. (2019). The optimal combination of feature selection and data discretization: An empirical study. *Information Sciences*, *505*, 282–293. https://doi.org/10.1016/j.ins.2019.07.091

Wang, J., Liu, Y., & Li, Y. (2019). Ecological restoration under rural restructuring: A case study of Yan'an in China's Loess Plateau. *Land Use Policy*, *87*, 1–9. https://doi.org/10.1016/j.landusepol.2019.104087

Wen, J., Hou, K., Li, H., Zhang, Y., He, D., & Mei, R. (2021). Study on the spatial-temporal differences and evolution of ecological security in the typical area of the Loess Plateau. *Environmental Science and Pollution Research*, *28*, 23521–23533. https://doi.org/10.1007/s11356-021-12372-4

Wen, J., Lai, X., Shi, X., & Pan, X. (2013). Numerical simulations of fractional vegetation coverage influences on the convective environment over the source region of the Yellow River. *Meteorology and Atmospheric Physics*, *120*, 1–10. https://doi.org/10.1007/s00703-013-0241-0

Wiesmair, M., Feilhauer, H., Magiera, A., Otte, A., & Waldhardt, R. (2016). Estimating vegetation cover from high-resolution satellite data to assess grassland degradation in the Georgian Caucasus. *Mountain Research and Development*, *36*, 56–65. https://doi.org/10.1659/MRD-JOURNAL-D-15-00064.1

Wittich, K.-P., & Hansing, O. (1995). Area-averaged vegetative cover fraction estimated from satellite data. *International Journal of Biometeorology*, *38*, 209–215. https://doi.org/10.1007/BF01245391

Xia, X., Lin, K., Ding, Y., Dong, X., Sun, H., & Hu, B. (2021). Research on the coupling coordination relationships between urban function mixing degree and urbanization development level based on information entropy. *International Journal of Environmental Research and Public Health*, *18*, 242. https://doi.org/10.3390/ijerph18010242

Xu, X., & Zhang, D. (2021). Evaluating the effect of ecological policies from the pattern change of persistent green patches—A case study of Yan'an in China's Loess Plateau. *Ecological Informatics*, *63*, 1–10. https://doi.org/10.1016/j.ecoinf.2021.101305

Xu, X., Zhang, D., Zhang, Y., Yao, S., & Zhang, J. (2020). Evaluating the vegetation restoration potential achievement of ecological projects: A case study of Yan'an, China. *Land Use Policy*, *90*, 104293. https://doi.org/10.1016/j.landusepol.2019.104293

Yan, H., Zhan, J., Liu, B., Huang, W., & Li, Z. (2014). Spatially explicit assessment of ecosystem resilience: An approach to adapt to climate changes. *Advances in Meteorology*, *2014*, 1–10. https://doi.org/10.1155/2014/798428

Yin, R., & Yin, G. (2010). China's primary programs of terrestrial ecosystem restoration: Initiation, implementation, and challenges. *Environmental Management*, *45*, 429–441. https://doi.org/10.1007/s00267-009-9373-x

Younes, N., Joyce, K. E., Northfield, T. D., & Maier, S. W. (2019). The effects of water depth on estimating fractional vegetation cover in mangrove forests. *International Journal of Applied Earth Observation and Geoinformation*, *83*, 1–18. https://doi.org/10.1016/j.jag.2019.101924

Zhang, S., Chen, H., Fu, Y., Niu, H., Yang, Y., & Zhang, B. (2019). Fractional vegetation cover estimation of different vegetation types in the Qaidam basin. *Sustainability*, *11*, 864. https://doi.org/10.3390/su11030864

Zhang, D., Jia, Q., Wang, P., Zhang, J., Hou, X., Li, X., & Li, W. (2020). Analysis of spatial variability in factors contributing to vegetation restoration in Yan'an, China. *Ecological Indicators*, *113*, 1–13. https://doi.org/10.1016/j.ecolind.2020.106278

Zhang, D., Jia, Q., Xu, X., Yao, S., Chen, H., Hou, X., Zhang, J., & Jin, G. (2019). Assessing the coordination of ecological and agricultural goals during ecological restoration efforts: A case study of Wuqi County, Northwest China. *Land Use Policy*, *82*, 550–562. https://doi.org/10.1016/j.landusepol.2019.01.001

Zhang, J.-T., Ru, W., & Li, B. (2006). Relationships between vegetation and climate on the Loess Plateau in China. *Folia Geobotanica*, *41*, 151–163. https://doi.org/10.1007/BF02806476

Zhang, D., Xu, X., Yao, S., Zhang, J., Hou, X., & Yin, R. (2020). A novel similar habitat potential model based on sliding-window technique for vegetation restoration potential mapping. *Land Degradation & Development*, *31*, 760–772. https://doi.org/10.1002/ldr.3494

Zhi, Z., Yin, H., Lu, N., Zhang, X., Yu, K., Guo, X., & Qi, H. (2019). Spatial-temporal changes of vegetation restoration in Yan'an based on MODIS NDVI and Landsat NDVI. In *2019 IEEE international conference on signal, information and data processing (ICSIDP)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICSIDP47821.2019.9173313

Zhou, Y., Dong, J., Xiao, X., Liu, R., Zou, Z., Zhao, G., & Ge, Q. (2019). Continuous monitoring of lake dynamics on the Mongolian Plateau using all available Landsat imagery and Google Earth Engine. *Science of the Total Environment*, *689*, 366–380. https://doi.org/10.1016/j.scitotenv.2019.06.341

Zhu, A. X., Liu, J., Du, F., Zhang, S. J., Qin, C. Z., Burt, J., Behrens, T., & Scholten, T. (2015). Predictive soil mapping with limited sample data. *European Journal of Soil Science*, *66*, 535–547. https://doi.org/10.1111/ejss.12244

Zhu, Z., Wang, S., & Woodcock, C. E. (2015). Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsat 4–7, 8, and entinel 2 images. *Remote Sensing of Environment*, *159*, 269–277. https://doi.org/10.1016/j.rse.2014.12.014

Zou, Z., Xiao, X., Dong, J., Qin, Y., Doughty, R. B., Menarguez, M. A., Zhang, G., & Wang, J. (2018). Divergent trends of open-surface water body area in the contiguous United States from 1984 to 2016. *Proceedings of the National Academy of Sciences*, *115*, 3810–3815. https://doi.org/10.1073/pnas.1719275115

Zurqani, H. A., Post, C. J., Mikhailova, E. A., Schlautman, M. A., & Sharp, J. L. (2018). Geospatial analysis of land use change in the Savannah River basin using Google Earth Engine. *International Journal of Applied Earth Observation and Geoinformation*, *69*, 175–185. https://doi.org/10.1016/j.jag.2017.12.006

How to cite this article: Meng, X., Pi, H., Gao, X., He, P., & Lei, J. (2022). A high-accuracy vegetation restoration potential mapping model integrating similar habitat and machine learning. *Land Degradation & Development*, 1–17. https://doi.org/10.1002/ldr.4527