land

MDPI

Article

Variables and a Validation Data Analysis to Improve the Prehistoric Cultivated Land Predictive Precision of Yulin, Northern Shaanxi, China

Xinyuan Kong, Jianxin Cui * and Yikai Li

Northwest Institute of Historical Environment and Socio-Economic Development, Shaanxi Normal University, Xi'an 710119, China; kongxy@snnu.edu.cn (X.K.); lyk2020@snnu.edu.cn (Y.L.)
* Correspondence: cuijx@snnu.edu.cn

Abstract: The distribution of cultivated land in prehistoric times was primarily influenced by natural conditions and population density. This article presents a case study on modern cultivated land simulation to analyze the potential impact of variable selection and validation data accuracy on model precision. Additionally, methods were explored to enhance the accuracy of prehistoric cultivated land simulation. Seven natural variables and one settlement density variable were selected to simulate the distribution of cultivated land based on a Binary Logistic Regression model. The simulated results were then compared with real land use data from 1985, which are commonly used as validation data for prehistoric farmland reconstruction. The findings revealed that all eight selected parameters could explain the distribution of cultivated land in the research area, with annual precipitation being the most influential factor. The initial prediction accuracy was relatively low at 65.8%, with a Kappa coefficient of 0.316. Several factors were identified as affecting the prediction accuracy. Firstly, the scale effect diminished the impact of the slope and elevation on cultivated land distribution, and errors were introduced in the method used to calculate the distance from residential areas. Secondly, the loess hilly area in the southeastern part of the research area overpredicted cultivated land due to insufficient data on actual residential land demand. Lastly, strong human activity since the 1950s has altered the natural distribution of cultivated land, resulting in poor consistency ratings. To address these issues, a batch modification method was employed to correct the 1985 data. The validation of the prediction model using the corrected data demonstrated a significant improvement in accuracy. Therefore, it is recommended to use the revised 1985 land use data for verifying prehistoric cultivated land simulation in the region. However, further research is required to mitigate the impact of the first two errors.

Keywords: model accuracy; land use simulation; Binary Logistic Regression; human activity; farming–pastoral transitional zone

Citation: Kong, X.; Cui, J.; Li, Y. Variables and a Validation Data Analysis to Improve the Prehistoric Cultivated Land Predictive Precision of Yulin, Northern Shaanxi, China. *Land* **2024**, *13*, 153. https://doi.org/10.3390/land13020153

Academic Editors: Guanghui Dong, Shanjia Zhang and Zhiping Zhang

Received: 17 December 2023
Revised: 24 January 2024
Accepted: 26 January 2024
Published: 28 January 2024

1. Introduction

As a crucial factor contributing to global change, researchers have extensively studied Land Use and Land Cover Change (LUCC) in terms of its development and evolution, driving mechanisms, distribution patterns, and simulation and reconstruction [1–3]. Recent advancements in remote sensing satellites have yielded a plethora of high-precision data for LUCC research [4–7]. However, these satellite observations are limited in terms of the monitoring period. To overcome the scarcity of remote sensing data for long-term land use studies, the model simulation method emerges as a crucial approach [8]. For instance, scholars have utilized various models, including PBM, CLUE, and REVEALS, to reconstruct historical land use patterns [9–11]. Additionally, researchers have been adapting classical models by adjusting parameters and variables to develop region-specific land use models [12–14]. Recently, some scholars have incorporated rules that capture human land

use behavior into cellular automata models, resulting in more precise reconstructions of past land use distributions [15,16].

Compared to the models mentioned above, the Binary Logistic Regression (BLR) model, known for its simplicity and fewer parameter constraints, has been extensively employed in land use reconstruction. Peppler et al. employed the BLR model to simulate land use in northern Hesse, Germany, incorporating physical factors and highlighting their role in determining the distribution pattern of land use in this region [17]. Similarly, Matasov et al. utilized socioeconomic statistics, historical maps, satellite images, and the BLR model to reconstruct the land use cover of the province of Ryazen, Russia, from 1770 to 2010 [18]. In China, Bai et al. utilized a BLR model to generate a probability map depicting the distribution of various land types in Dulbert Autonomous County of Inner Mongolia over the past century [19]. Their study considered both natural and socio-economic factors. In a similar manner, Chen et al. employed the same model to reconstruct the distribution of cultivated land during the late Neolithic period in the North China Plain using factors such as elevation, slope, soil, rivers, and proximity to residential areas [20]. Moreover, Yang et al. achieved a more precise reconstruction of the distribution of cultivated land in the Lower Mississippi Alluvial Valley from 1850 to 2018 [21]. This study showed that using machine learning algorithms and county-level census data has higher accuracy than relying solely on state-level population data.

Most prehistoric or historic cultivated reconstructions obeyed the following steps [22,23]: Firstly, a model is used to establish a quantitative relationship between modern populations and land use, ensuring high prediction accuracy. Next, the model parameters are also applied in conjunction with spatial analysis tools such as GIS to recreate the spatial and temporal distribution patterns of cultivated land in previous periods. In China, most studies on this topic have validated the model using land use data from the 1980s. The accuracy of predicting the distribution of cultivated land in the 1980s is frequently regarded as a significant criterion for assessing their suitability in predicting prehistoric or historic periods. However, when we attempted to apply the same method to simulate the year 1985′s cultivated land in the Yulin region of northern China, we encountered a low accuracy rate with the model. Peppler argued that data quality, parameter selection, and random errors can affect the predictive accuracy of a logistic regression model. Additionally, the absence of human factors may contribute to the poor predictive accuracy of land use simulations over the past 2000 years. However, there are limited studies analyzing the factors influencing the prediction accuracy of the model.

Additionally, previous research on cultivated land reconstruction in China has mainly focused on traditional farming areas in the east, with less attention given to the farming–pastoral transitional zone. Archaeological evidence suggests that advanced agriculture was already present in the local area during the Yangshao period (ca. 5000-4900 BCE) [24]. The Longshan culture (ca. 3000-2000/1900 BCE) developed a diversified subsistence strategy, with agriculture as the main focus and animal husbandry as a supplement [25]. Farmland is the foundation and product of agricultural activities; however, there is currently no clear understanding of the distribution pattern of cultivated land in the region. This ecotone exhibits stronger heterogeneity in various natural factors compared to the plain area in eastern China. At the same time, the landscape pattern of the farming–pastoral ecotone is highly sensitive to both natural and human-induced changes. Some scholars argued that climatic fluctuations, such as precipitation, directly impact the development of arable land in the region [26]. However, other studies suggested that population pressure, policies, and technological advancements are the main drivers behind the recurrent expansion and contraction of arable land [27]. In reality, the spatial and temporal changes in the agricultural and pastoral landscape pattern in the region were a combined response to climate change and human activities [28–30]. Therefore, this study aims to collect relevant physical and social factors as independent variables to build a BLR model. And the potential distribution of arable land in the Yulin area for the time period of 1985 was reconstructed. This reconstructed result was compared with the actual land use data to explore the factors that

may affect the prediction accuracy of the model at a certain scale. In order to minimize the influence of human activity on validating farmland data, this paper presents a method for batch-modifying raster data values to improve the accuracy of prediction evaluation. The findings of this research will serve as a reference for accurately simulating and predicting the spatial distribution patterns of cultivated land in this region, whether historical or prehistoric, in future studies.

2. Study Area

Yulin is located in the north of Shaanxi Province, China (36°57′~39°34′ N, 107°28′~111° 15′ E), with a total land area of 4.292×10^4 km^2 (Figure 1). In terms of the climate type, the region belongs to the transition zone from a temperate monsoon climate to temperate continental climate, from northwest to southeast, with an annual temperature of 7–9 °C. The frost-free period is short, and the precipitation ranges from 300 to 500 mm, mainly in the summer. Taking the Ming Great Wall as the boundary, the topography and geomorphology of the study area show great differences. Beyond the Great Wall lies an expanse of sandy and grassland terrain, encompassing 42% of the total area. This region features a gentle and undulating topography, characterized by a continuous distribution of sand dunes as well as scattered beaches and lakes. Within the Great Wall, there is a significant expanse of loess hills and gullies, encompassing 58% of the total area. This region can be further classified into the eastern loess hills and gullies, as well as the western low mountain hills area. The eastern loess hilly and gully area is characterized by an alternating distribution of hills and ridges. The surface in this area is fragmented due to the impact of flowing water erosion, resulting in significant fluctuations in the terrain. On the other hand, the western low mountain and hilly area has a higher elevation. The tableland in this area is wide and the slope is relatively gentle. In conclusion, the terrain in this region is higher in the west and lower in the east. Influenced by the topography, most rivers flow from northwest to southeast. Major rivers, including Kuye River, Tuwei River, and Wuding River, all directly discharge into the Yellow River.

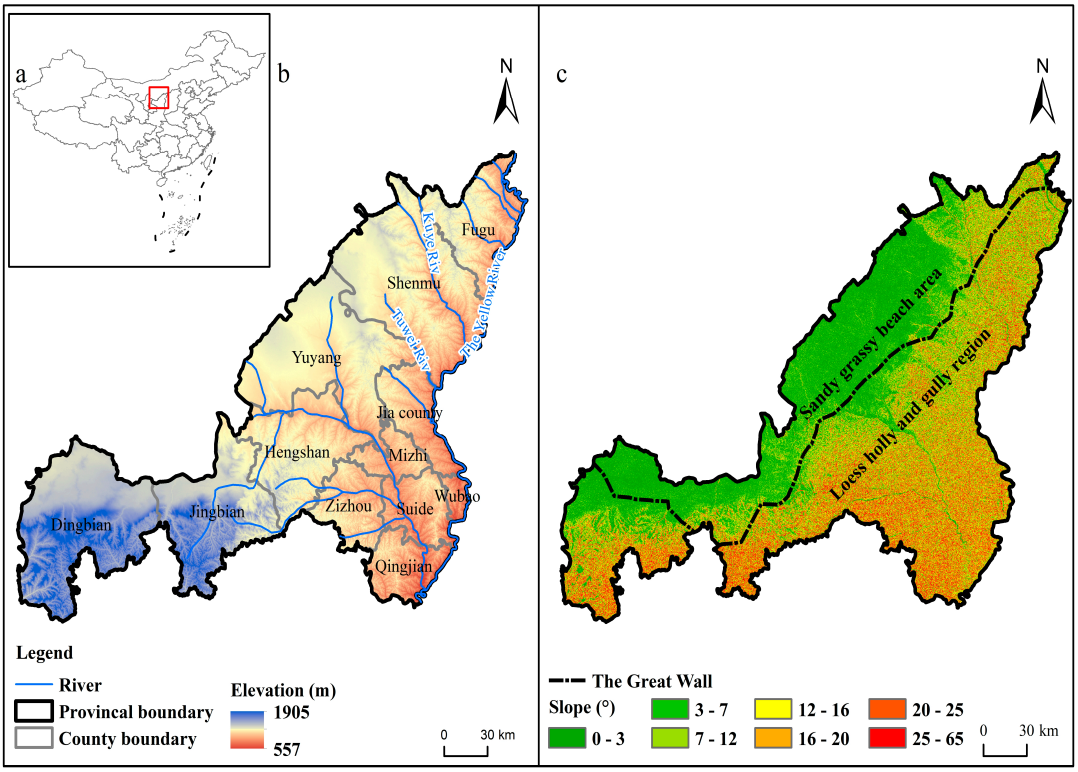

Figure 1. Map of the study area. (**a**) Geographic location of the study area within China; (**b**) administrative divisions, rivers, and terrain in Yulin region; (**c**) spatial distribution of slope in Yulin area.

3. Materials and Methods

3.1. Parameter Factor Selection

Based on previous research and combined with the development process of cultivated land in the region from Ming-Qing (1368-1911CE) to the present [31–34], this article identified the variables that affect the spatial distribution pattern of the cultivated land. Natural variables such as the slope, altitude, soil type, and proximity to rivers have been found to play a crucial role in determining the formation of arable land. Previous studies have also highlighted the significant influence of precipitation and temperature on the development of arable land [30,35,36]. Moreover, the amount of cultivated land is influenced by population density, as it is a result of human activities. Furthermore, the spatial distribution of cultivated land is closely associated with the distance from the residential area [37]. The research area exhibits a rugged and uneven surface, which means that the visible distance (Euclidean distance) may not accurately reflect the actual distance required to reach a specific location. In order to address this problem, this study focused on calculating the time it takes to walk 1 m under different slope conditions and applied the results for a Cost–Distance Analysis. Additionally, eight factors were selected as driving factors that influence cultivated land development: elevation, slope, soil type (ST), annual average temperature (AAT), annual average precipitation (AAP), distance from rivers (DR), distance from settlements (DS), and population density (PD).

3.2. Data Sources

The land use data for Yulin in 1985 were extracted from the dataset by Yang et al. [38], with a spatial resolution of 30 m. Digital Elevation Model (DEM) data were downloaded from the Geographic Spatial Data Cloud (http://www.gscloud.cn, accessed on 5 July 2023) with a spatial resolution of 30 m, and slope data were calculated from the DEM data using ArcGIS 10.2 software. Soil data were obtained from the Resource and Environment Science Data Center of the Chinese Academy of Sciences (http://www.resdc.cn, accessed on 5 July 2023). These data were digitized based on the "1:10,000,000 soil map of the People's Republic of China (PRC)", which was compiled and published by the National Soil Census Office in 1995, according to the traditional "Soil Occurrence Classification System" (SOCS). The STs were divided into 12 soil orders, 61 soil classes, and 227 subclasses, with a spatial resolution of 1 km. AAT and AP data were derived from the WorldClim project (http://worldclim.org, accessed on 6 July 2023). The temporal coverage spans from 1970 to 2000, with a spatial resolution of 1 km. PD data are derived from the 1 Kilometer Grid Population Spatial Distribution Dataset of China [39]. This study utilized population density data from 1990, which is also the year closest to the distribution of cultivated land to be reconstructed. The vector boundaries, river, and settlement point data for the study area were downloaded from the National Geographical Information Resource Catalog Service System (http://www.webmap.cn, accessed on 6 July 2023). The scale of the latter two datasets is 1:250,000.

3.3. Data Processing

The land use data of 1985 were reclassified in ArcGIS. Cultivated land was separated into one category, while all other land use types were grouped into another category. The distance to rivers raster data were created using the Euclidean Distance tool in ArcGIS 10.2, based on river vector data. Similarly, the distance to settlement points raster data were generated using the Cost–Distance tool. A previous study identified a relationship between walking time and slope, which is as follows [40]:

$$Y = 0.0002x^2 + 0.002x + 0.6086 \quad (1)$$

In the equation, Y represents the time (second) required to walk 1 m, and x represents the slope.

Using the Raster Calculator tool in ArcGIS 10.2, the slope–time raster was derived and used as the cost raster data. This raster, combined with settlement point data, generated the distance to settlement points raster. Combining the soil organic matter content and the "barren land" type in the land use data of 1985, the ST data were reclassified into 5 levels using ArcGIS 10.2. They included the most unsuitable type and the other four types based on organic matter content from low to high. To ensure comparability of contribution rates for each variable, it is essential to standardize the variables due to their non-uniform units and significant differences in numerical values. The processing formula is as follows [41]:

$$GY_i = \frac{MaxY_i - Y_i}{MaxY_i} \tag{2}$$

i represents the grid number, $MaxY_i$ is the maximum value of each independent variable, Y_i is the value of each independent variable, GY_i is the standardized value of the independent variable for grid i, with a range of [0, 1].

The spatial scope of the research area determined that the simulated reconstruction of cultivated land in this study has a resolution of 250 m. Additionally, resampling was conducted on each grid dataset to ensure that the number and resolution of each variable grid can be overlaid and analyzed.

3.4. Model Configuration

The BLR model is an equation model that predicts a binary dependent variable (0 or 1, yes or no) based on continuous or categorical independent variables. It calculates separate coefficients for each explanatory variable and determines the probability of the dependent variable occurring through weighted calculations. In this study, we considered the distribution of cultivated land in 1985 as the dependent variable and took eight natural and social factors that influence cultivated land distribution as independent variables. A BLR model was built to calculate the weights of each independent variable. Finally, the Raster Calculator tool in ArcGIS 10.2 was utilized to derive the spatial probability distribution of cultivated land in 1985. The formula of the BLR model is the following [42]:

$$log\left(\frac{P_i}{1 - P_i}\right) = \alpha + \sum_{k-1}^{k} \beta_k X_{k_i} \tag{3}$$

i represents the raster number, P_i is the probability of raster i becoming cultivated land or non-cultivated land, α is the constant term, X is the variable value, β is the regression coefficient, and k is the number of independent variables. For the analysis of the model's independent variables, a p-value of less than 0.05 indicates that the selected factors have reached a significant level and are variables that influence the probability of cultivated land distribution. This article assumed $p \leq 0.5$ for arable land and $p > 0.5$ for non-arable land. Comparing the predicted results with the actual land use data, the higher the prediction accuracy, the more reasonable the model construction is considered as.

To determine the parameters of the BLR model, 50,000 grids were randomly selected from a total of 686,541 grids in the study area for cultivated land and non-cultivated land, respectively. Label points were then used to extract the corresponding values of each independent variable. The extracted data were exported and loaded into Excel. The calculation of the model's paraments was conducted using IBM SPSS Statistics 22.

3.5. Precision Evaluation

This paper uses the Kappa coefficient to evaluate the model's prediction accuracy. The formula for calculating the Kappa coefficient is as follows:

$$Kappa = \frac{P_0 - P_c}{1 - P_c} \tag{4}$$

P_0 is the proportion of correctly simulated data and P_c is the proportion of expected correct simulated data under a random situation. When the P_0 is greater than the P_c, the Kappa value is positive, and a larger Kappa value indicates better consistency. When the predicted results are entirely consistent with the actual data, the Kappa coefficient equals 1. Detailed classifying criteria can be found in Table 1 [43].

Table 1. The classified Kappa coefficient indicates the predictive power of the model.

Kappa Index	Degree of Agreement
<0.05	None
0.05–0.20	Very poor
0.20–0.40	Poor
0.40–0.55	Fair
0.55–0.70	Good
0.70–0.85	Very good
0.85–0.99	Excellent
0.99–1.00	Perfect

4. Results

4.1. Model Parameter Analysis

The *p*-values of the selected independent variables for model construction are all below 0.05, indicating that all eight factors significantly influence the probability of cultivated land distribution and effectively explain its spatial patterns. Among these factors (Table 2), precipitation has the strongest impact on the distribution of cultivated land. The likelihood of the grid becoming farmland increases by 52.329 times for every doubling of precipitation, which is in line with previous research emphasizing the significance of water resources in constraining the expansion of cultivated land in the region [30,32–36]. Additionally, PD and ST also have a significant impact on the distribution of cultivated land. Each increase in their values by one unit increases the likelihood of the grid becoming cultivated land by 15.398, and 12.366 times, respectively.

Table 2. The estimation of the significant parameter of independent variables.

Parameter	β	Standard Deviation	Exp (B)
Elevation	−0.547 **	0.093	0.579
Slope	−0.949 **	0.059	0.387
ST	2.515 **	0.043	12.366
AAT	0.851 **	0.130	2.342
AAP	3.958 **	0.148	52.329
DR	2.271 **	0.076	9.689
DS	−5.460 **	0.129	0.004
PD	2.734 **	0.104	15.398
Constant	−0.156	0.155	0.856

Only standardized variables with $p < 0.05$ value were used in the model; "**" correspond to $p < 0.01$.

4.2. Model Accuracy

By substituting the regression coefficients of each variable and the constant into the BLR, the model is ultimately determined as

$$P = \frac{\exp[A]}{1 + \exp[A]} \quad (5)$$

$$A = -0.547 \times (\text{Elevation}) - 0.949 \times (\text{Slope}) + 2.515 \times (\text{ST}) + 0.851 \times (\text{AAT}) + 3.958 \times (\text{AAP}) + 2.271 \times (\text{DR}) - 5.460 \times (\text{DS}) + 2.734 \times (\text{PD}) - 0.156$$

The predicted results using the model are presented in Table 3. In this table, a value of 1 denotes cultivated land and a value of 2 represents non-cultivated land. We randomly selected 50,000 cultivated grids and 50,000 non-cultivated grids from the actual land use data of 1985. In the model, a total of 33,292 cultivated grids and 32,451 non-cultivated grids were accurately predicted. The overall classification accuracy is 65.8%, with a Kappa coefficient of 0.316 indicating limited consistency in the model's predictions.

Table 3. BLR predictive classification results.

		Predicted		
Observed		**Cultivated**	**Non-Cultivated**	**Correct (%)**
	Cultivated	33,292	16,698	66.6
	Non-cultivated	17,533	32,451	64.9
	Total share (%)			65.8

The model was then applied to the whole study area. According to the model results, a total of 304,439 grids were classified as cultivated, while 382,102 grids were categorized as non-cultivated. The predicted number of grids for cultivated land is 105,714 more than the actual number. A further analysis through visual interpretation showed that these additional predicted cultivated areas were mainly located in the southeastern region of the study area. To investigate and quantify these differences between predicted and actual distributions, a comparative analysis using ArcGIS's Raster Calculator function was conducted (Figure 2).

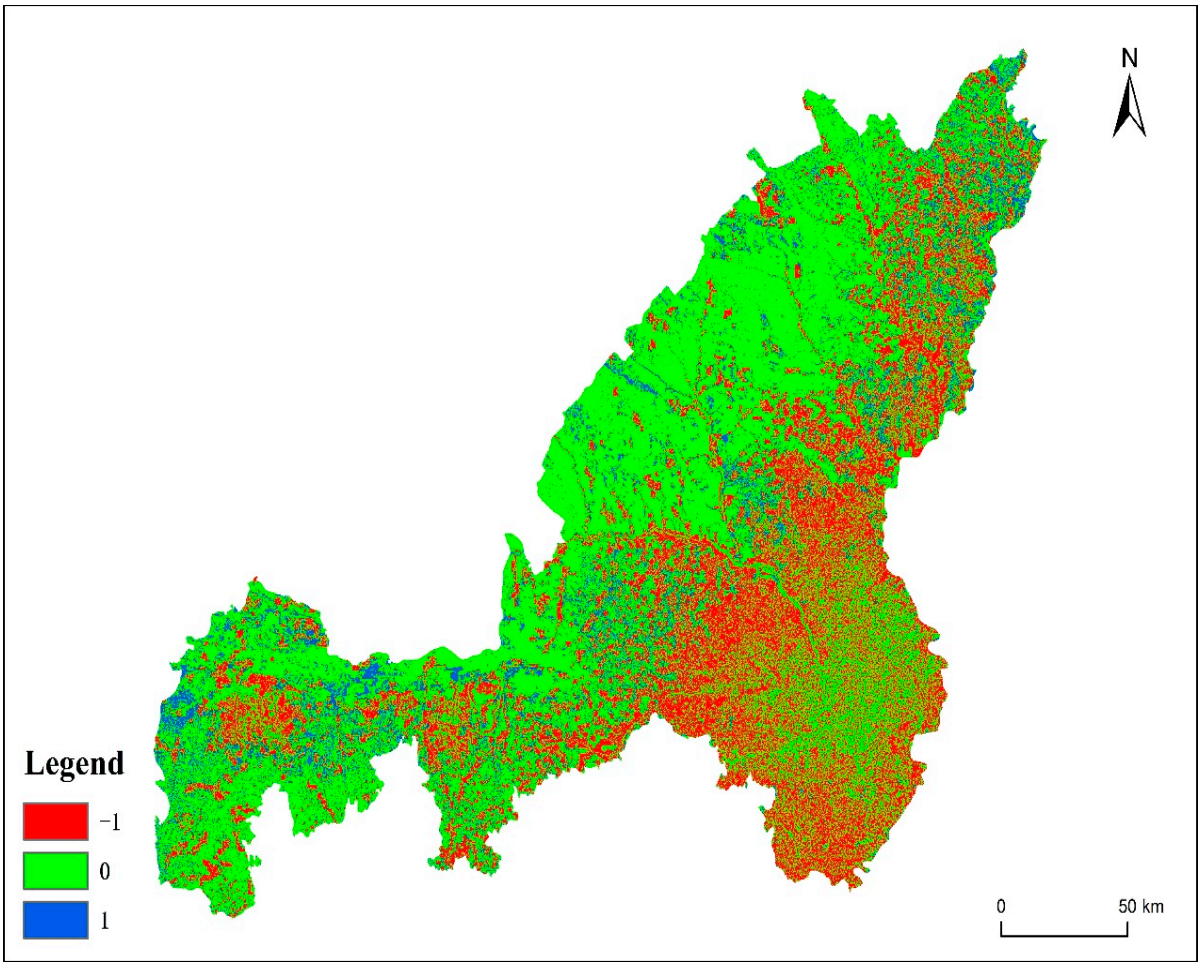

Figure 2. Comparison of first simulated farmland and actual spatial pattern of 1985 in Yulin.

There are 448,063 grids with a value of 0. These grids are accurately predicted using the model, including cultivated and non-cultivated land. Grids with a value of 1 indicated that they were predicted as non-cultivated but were actually cultivated, and these were predominantly found in the northwest sandy grassland area (NSGA). On the other hand, grids with a value of −1 represented grids that were predicted as cultivated but were actually non-cultivated. These grids were mainly distributed in the southeastern loess hilly region (SLHR), as well as in the central zone of Jingbian and Dingbian Counties (Figure 2).

5. Discussion

5.1. Causes Affecting Model Accuracy

5.1.1. Southeastern Loess Hills Region

After comparing the model's predicted results with the actual value, it was found that the cultivated land in the SLHR was overpredicted. The scale effect played a crucial role in the accuracy of land use simulation predictions. Research indicated that as the scale increases and the grid resolution decreases, the influence of elevation and slope on the distribution of cultivated land gradually diminishes [44]. Only excessively high elevations and steep slopes continue to restrict the distribution of the cultivated land. The study suggested that in the Loess Plateau region, as the spatial resolution of DEM data decreases, the overall slope in the study area tends to decrease, with the highest variability observed in the most fragmented loess hilly and gully areas. In the north Loess Plateau, DEM data with a resolution of 5 m provide more accurate slope information [45]. However, this study utilized slope data extracted from the DEM with a resolution of 30 m, which were further resampled to a resolution of 250 m. Consequently, the overall grid slope values may be underestimated due to resampling. As a result, the model may incorrectly classify some grids with steep slopes, which are unsuitable for cultivation, as suitable for agricultural development due to scale-related reasons.

Population density is also an important factor that influences farmland development. There is a positive correlation between population quantity and the extent of cultivated land to some extent. However, to accurately simulate actual farmland distribution, it is necessary to consider the land requirements of residents at each settlement as a constraint. Unfortunately, the data on actual land requirements are currently unavailable. In this study, the distance to residential areas was selected as a factor. The settlement point was taken as the center, and various distances were used as radii to generate cost–distance data. However, this approach did not consider topographical factors. Consequently, grids with shorter distances from the settlement point were more likely to be classified as suitable for cultivated land development, despite the fact that the actual distribution of cultivated land was more fragmented and complex in comparison to the flat areas [46]. Despite the challenges of a large slope and soil erosion resulting from cultivation in the steep slope, the southeastern hilly and loess region generally has a lower elevation and higher AAT and AAP. This makes it more suitable for agricultural cultivation. However, it is important to consider the possibility of overestimating cultivated land in this region, as this could impact the overall accuracy of this study.

5.1.2. Northwest Sandy Grasslands Area

On a shorter timescale, the impact of physical condition on regional-scale land use changes is limited, with socio-economic factors playing a decisive role [47]. The 1985 land use map indicated a substantial presence of arable land in the NSGA. However, this study predicted a considerable portion of the cultivated land to be non-cultivated land. Historical records from the county annals and 'The Annals of Shaanxi Province: Water Conservancy Chronicles' revealed that this area experienced two waves of farmland development after the establishment of the PRC [48–53]. The first wave occurred during the Great Leap Forward in the 1950s, and the second wave took place during the agricultural development movement from the 1960s to 1970s. During these periods, the communes organized a large workforce to reclaim the sandy and grassland area. The shoal land can be planted directly,

while salinized land was improved using measures such as digging trenches for drainage, adding sand to neutralize alkalinity, and irrigating to flush out alkali. For instance, in Dingbian County, a total of 960 hectares (ha) of salinized land was reclaimed from 1958 to 1985, resulting in a yield increase of over 50% per ha. The local water conservancy construction also progressed during the process of cultivated development [54]. Reservoirs like Jinjisha and Xinqiao were constructed in the Jingbian and Yulin regions for water storage and irrigation. In areas lacking surface runoff, abundant groundwater resources were tapped and utilized through well drilling technology. Concurrently, agricultural supporting facilities such as irrigation canals were constructed. This comprehensive approach facilitated the transformation of extensive sandy grassland into cultivated farmland. Representative examples of this transformation included the town of Ningtiaoliang in Jingbian County and the Balihe Irrigation District in Dingbian County [47,49] (refer to Table 4). The development of these cultivated lands was directly influenced by socio-economic factors such as policies and technologies. However, these difficult-to-quantify factors were not included as independent variables in the predictive model, leading to a decrease in overall predictive accuracy. The failure of prediction for a significant number of cultivated lands in the NSGA played a crucial role in the poor predictive consistency.

Table 4. The number of irrigation channels and wells in the NSGA.

Location	The Length of Irrigation (Unit: km)	The Number of Wells
Dingbian	210	2243
Jingbian	400	1176
Hengshan	233	134
Yuyang	300	1003
Shenmu	167	No data

The length of irrigation canals and the number of wells composed data for around the year 1985. In addition, Dingbian County had 1541 wells equipped with water tanks. By the end of 1993, Yuyang District had a total of 2011 wells, including trough wells and artesian trough wells.

5.2. Land Use Data Revision and Model Reconstruction

In prehistoric and historical periods, the cultivated land distribution was most controlled by the natural conditions and the settlement density. In order to improve the accuracy of prehistoric models, it is crucial to use validation data that closely represent the distribution during that time. The data from 1985, which have been commonly used in previous studies in the region, have already been significantly affected by human activities. Therefore, it cannot be directly used for verifying the accuracy of the model. To obtain a more precise prediction model, this article employs the method involving batch-modifying grid values to eliminate the influence of human activities. In the 1985 land use map, raster values predicted using the model to be non-cultivated that were actually farmland were changed to non-cultivated and a total of 65,910 grids were changed. The modified land-use-type map was then reclassified into binary values, with 132,371 cultivated land grids and 554,170 non-cultivated grids. For the new round of prediction, 50,000 grids representing cultivated land and 50,000 grids representing non-cultivated land were randomly selected. The results showed that out of the 50,000 cultivated grids, 41,266 were accurately predicted, while 8734 were predicted incorrectly. Similarly, for the 50,000 non-cultivated grids, 39,054 were correctly predicted, while 10,946 were wrongly predicted (Table 5). The predictive accuracy of the model was significantly improved compared to the initial predictions, with an overall accuracy of 80.3% and a Kappa coefficient of 0.606.

Table 5. New BLR prediction classification results.

Observed		Predicted		
		Cultivated	Non-Cultivated	Correct (%)
	Cultivated	41,266	8734	82.5
	Non-cultivated	10,946	39,054	78.1
	Total share (%)			80.3

Table 6 demonstrates that in the new model construction, the *p*-value of the eight selected independent variable is less than 0.05. All of these variables reach the significance level and can be used as explanatory variables that affect the distribution of cultivated land. "**" correspond to $p < 0.01$.

Table 6. The new estimation of the significant parameter of independent variables.

Parameter	β	Standard Deviation	Exp (B)
Elevation	−2.305 **	0.119	0.100
Slope	−2.663 **	0.073	0.070
ST	6.665 **	0.068	784.207
AAT	1.673 **	0.162	5.328
AAP	13.859 **	0.197	1,044,307.436
DR	7.364 **	0.102	1577.393
DS	−13.710 **	0.192	0.000
PD	4.784 **	0.119	119.584
Constant	0.500 **	0.196	1.648

Only standardized variables with $p < 0.05$ value were used in the model; "**" correspond to $p < 0.01$.

Among the factors influencing the distribution of cultivated land, AAP remains the most important, followed by DS and ST. The regression coefficients for these three factors are 13.859, 7.364, and 6.665, respectively. In comparison to the initial model, the impact of PD on the distribution of cultivated land has decreased and now ranks fourth. This decrease can be attributed to the batch modification of grids, which eliminated the impact of human activities at a large scale. According to the historical archives [55], by the late 1970s, the PDs in counties (districts) adjacent to the Mu Us Sandy Land, including Hengshan, Yuyang, Jingbian, Shenmu, and Dingbian, had all exceeded 30 persons/km^2. These data surpassed the population thresholds set by the United Nations for semi-arid areas (20 persons/km^2) and arid areas (7 persons/km^2). Alongside population growth, the rising demand for food was a significant factor contributing to the change in land cover in sandy grassland areas. The new BLR model (Formula (6)) is shown below.

$$P = \frac{\exp[A]}{1 + \exp[A]} \tag{6}$$

$$A = -2.305 \times (\text{Elevation}) - 2.663 \times (\text{Slope}) + 6.665 \times (\text{ST}) + 1.673 \times (\text{AAT}) + 13.859 \times (\text{AAP}) + 7.364 \times (\text{DR}) - 13.710 \times (\text{DS}) + 4.784 \times (\text{PD}) + 0.500$$

By utilizing this model to analyze the distribution of cultivated land and comparing it with the modified land use types from 1985, the findings reveal that there are 122,517 grids with a value of −1 (Figure 3). These grids remain concentrated in the southeast loess hilly area as well as the central parts of Dingbian and Jingbian. On the other hand, the number of grids with a value of 1 has notably decreased, with only 23,374 grids remaining. These grids are primarily found in the northern parts of Dingbian and Jingbian, along with other sandy grassland areas.

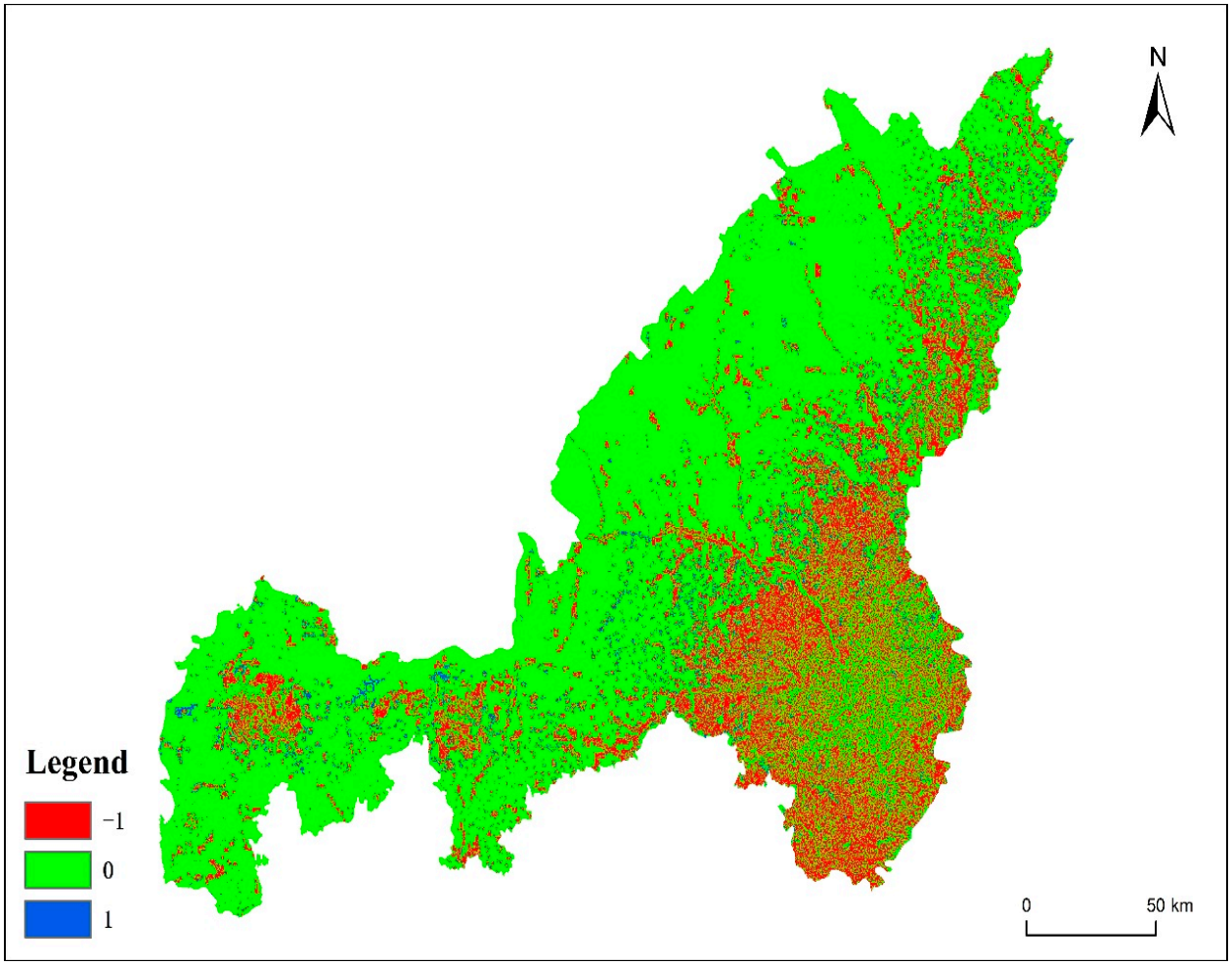

Figure 3. Comparison of second simulated farmland and actual spatial pattern of 1985 in Yulin.

In order to further show the differences between the predicted cultivated land and the 1980s farmland data, we randomly selected five archaeological sites during the Longshan period in both the northwest sandy grassland area and the southeastern loess hilly region. The cultivated land area within the 5 km buffer zone of these points was extracted from the predicted and 1980s datasets, respectively (Table 7). The results reveal that the predicted cultivated land of the five sites in the southeastern loess hilly region is significantly higher than the actual cultivated land, indicating an overestimation in this region. Alternatively, the predicted cultivated area of the five sites in the northwest sandy grassland area was significantly lower than the actual cultivated data, with two sites showing the largest discrepancies. These findings further support our conclusion. Since the farmland data from the 1980s have been strongly disrupted by human activities, it must be corrected before it can be used as validation data for future prehistoric farmland prediction models.

In addition, to further demonstrate the effectiveness of batch-modifying the grid values of the validation dataset in improving prediction accuracy, we also selected land use data from the Yulin region in 2005 for validation. The above experimental process was repeated to obtain results, which showed that the model's prediction accuracy was 64% when cultivated grids in the northern sandy grassland area were not excluded. However, after excluding these grids, the prediction accuracy increased to 77.6%. Therefore, it can be considered that this method presents a novel approach for reconstructing land use in the prehistoric/historical period of the northern agricultural pastoral transitional zone.

Table 7. Site information for verification and the cultivated land area within 5 km buffer zone.

No.	Longitude	Latitude	Location	Site	Prediction Area (Number of Cells)	Actual Area (Number of Cells)
1	109.40	38.57	NSGA	HTGLW	0	131
2	109.62	38.36	NSGA	CHJHSL	0	90
3	110.45	39.22	NSGA	LJSPYW	89	143
4	109.86	39.06	NSGA	ABGD	194	195
5	110.00	38.88	NSGA	TJWDP	49	54
6	110.19	37.64	SLHR	MXZZS	1298	848
7	110.02	37.75	SLHR	ZS	1297	481
8	109.52	37.55	SLHR	WYQZS	993	303
9	110.45	37.81	SLHR	GADMZS	1278	877
10	110.36	37.37	SLHR	JDGD	1285	842

6. Conclusions

This paper examines eight physical and social factors and utilizes the BLR model to simulate the distribution of cultivated land in the study area in 1985. By comparing the simulated results with the actual land use map from 1985, the following preliminary conclusions are drawn:

1. AAP is the major factor that influenced the distribution of cultivated land in the study area. The initial construction of the model yielded an overall predictive accuracy of 65.8%, with a Kappa coefficient of 0.316, indicating poor predictive consistency. The factors that affect accuracy vary in different landforms.
2. The influence of factors such as the elevation and slope on the distribution of cultivated land decreases as the data resolution decreases, mainly due to scale effects. This influence is particularly noticeable in the hilly areas of the southeast. Moreover, the over-prediction of cultivated land distribution in the southeastern loess hilly region can be attributed to the calculation assignment method for distance to residential areas and the absence of actual cultivated land demand data for each residential area.
3. In the northern sandy grassland area, modern humans have constructed agricultural water conservancy projects, which have altered soil moisture conditions and converted some sandy areas into cultivated fields. The presence of these cultivated lands has adversely affected the initial predictions of the model, resulting in lower predictive accuracy.
4. It is important to note that relying on cultivated land distribution maps influenced by modern human activities can lead to significant errors when validating predictive models for prehistoric or historical periods. To address this problem, a batch modification method was used to adjust the values of land use grids, aiming to minimize the impact of human activities and make them more representative of early cultivated use scenarios. The modified land use data will serve as background values for future simulations of prehistoric period farmland, providing parameter references for cultivated land predictive models.

Author Contributions: Conceptualization, X.K. and J.C.; Methodology, X.K. and J.C.; Software, X.K.; Data Curation, X.K. and Y.L.; Visualization, X.K.; Writing—Original Draft, X.K.; Writing—Review and Editing, X.K., J.C. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Project of the Key Research Base of Humanities and Social Sciences of the Ministry of Education, China (Grant No. 22JJD770053), and the National Natural Science Foundation of China (Grant No. 41571190).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: We sincerely thank the editors and the anonymous reviewers for their insightful comments and suggestions. At the same time, we appreciate the basic data provided by the Resources and Environmental Science and Data Center of the Chinese Academy of Sciences and National Catalogue Service for Geographic Information, China.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Goldewijk, K.K. Estimating global land use change over the past 300 years: The HYDE Data base. *Glob. Biogeochem. Cycles* **2001**, *15*, 417–433. [CrossRef]
2. Lin, Q.; Luo, G.P.; Chen, X. Review of Land-use Model. *Prog. Geogr.* **2005**, *24*, 79–87.
3. Qiao, Z.; Jiang, Y.Y.; He, T.; Lu, Y.S.; Xu, X.L.; Yang, J. Land use change simulation: Progress, challenges, and prospects. *Acta. Ecol. Sin.* **2022**, *42*, 5165–5176.
4. Ulbricht, K.A.; Heckendorff, W.D. Satellite images for recognition of landscape and land use change. *ISPRS J. Photogramm. Remote Sens.* **1998**, *53*, 235–243. [CrossRef]
5. Pan, Y.Z.; Chen, Z.J.; Nie, J.; Wang, X. Research on comprehensive monitoring approach in landuse dynamic change using multisource remote sensing data. *Adv. Earth Sci.* **2002**, *17*, 182–187+303.
6. Dang, A.R.; Shi, H.Z.; He, X.D. Dynamic variation of land use based on 3S technology. *J. Tsinghua Univ. (Sci. Tech.)* **2003**, *43*, 1408–1411.
7. Bai, W.Q.; Zhang, Y.M.; Yan, J.Z.; Zhang, Y.L. Simulation of land use dynamics in the upper reaches of the Dadu river. *Geogr. Res.* **2005**, *24*, 206–212+323.
8. He, F.N.; Li, M.J.; Yang, F. Main progress in historical Land Use and Land Cover Change in China during the past 70 years. *J. Chin. Hist. Geogr.* **2019**, *34*, 5–16.
9. Lemmen, C. World distribution of land cover changes during pre- and protohistoric times and estimation of induced carbon releases. *Geomorphol. Relief Process. Environ.* **2009**, *4*, 303–312. [CrossRef]
10. Fyfe, R.; Roberts, N.; Woodbridge, J. A pollen-based pseudobiomisation approach to anthropogenic land-cover change. *Holocene* **2010**, *7*, 1165–1171. [CrossRef]
11. Trondman, A.K.; Gaillard, M.J.; Mazier, F.; Sugita, S.; Fyfe, R.; Nielsen, A.B.; Twiddle, C.; Barratt, P.; Birks, H.J.; Bjune, A.E.; et al. Pollen-based quantitative reconstructions of Holocene regional vegetation cover (plant-funcetional types and land-cover types) in Europe suitable for climate modelling. *Glob. Chang. Biol.* **2015**, *21*, 676–697. [CrossRef] [PubMed]
12. Yu, Y.Y.; Guo, Z.T.; Wu, H.B.; Finke, A. Reconstructing prehistoric land use change from archeological data: Validation and application of a new model in Yiluo valley, northern China. *Agric. Ecosyst. Environ.* **2012**, *156*, 99–107. [CrossRef]
13. Yang, X.H.; Zhang, S.W.; Liu, Y.S.; Xing, X.; de Sherbinin, A. Analyzing historical land use changes using a Historical Land Use Reconstruction Model: A case study in Zhenlai County, northeastern China. *Sci. Rep.* **2017**, *7*, 41275. [CrossRef] [PubMed]
14. Ustaoglu, E.; Kabadayi, M.E. Reconstruction of Residential Land Cover and Spatial Analysis of Population in Bursa Region (Turkey) in the Mid-Nineteenth Century. *Land* **2021**, *10*, 1077. [CrossRef]
15. Long, Y.; Jin, X.B.; Li, M.Y.; Yang, X.; Cao, X.; Zhou, Y. A constrained cellular automata model for reconstructing historical arable land in Jiangsu province. *Geogr. Res.* **2014**, *33*, 2239–2250.
16. Yang, X.H.; Jin, X.B.; Guo, B.B.; Long, Y.; Zhou, Y. Research on reconstructing spatial distribution of historical cropland over 300 years in traditional cultivated regions of China. *Glob. Planet. Chang.* **2015**, *128*, 90–102. [CrossRef]
17. Peppler-Lisbach, C. Predictive modelling of historical and recent land-use patterns. *Phytocoenologia* **2003**, *33*, 565–590. [CrossRef]
18. Matasov, V.; Prishchepov, A.V.; Jepsen, M.R.; Müller, D. Spatial determinants and underlying drivers of land-use transitions in European Russia from 1770 to 2010. *J. Land Use Sci.* **2019**, *14*, 362–377. [CrossRef]
19. Bai, S.Y.; Zhang, S.W.; Zhang, Y.Z. Digital Rebuilding of LUCC Spatial-temporal Distribution of the Last 100 Years: Taking Dorbod Mongolian Autonomous County in Daqing City as an Example. *Acta Geogr. Sin.* **2007**, *62*, 427–436.
20. Chen, Q.Q.; Liu, F.G.; Fang, X.Q.; Zhou, Q.; Chen, Q.; Chen, R. Reconstruction of cropland distribution in the Late Neolithic period in Northern China. *Geo. Res.* **2019**, *38*, 2927–2940.
21. Yang, J.; Tao, B.; Shi, H.; Ouyang, Y.; Pan, S.; Ren, W.; Lu, C. Integration of remote sensing, county-level census, and machine learning for century-long regional cropland distribution data reconstruction. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *91*, 102151. [CrossRef]
22. Goldewijk, K.K.; Drecht, G.V.; Bouwman, A.F. Mapping contemporary global cropland and grassland distributions on a 5 × 5 minute resolution. *J. Land Use Sci.* **2007**, *3*, 417–433. [CrossRef]
23. Goldewijk, K.K.; Beusen, A.; Janssen, P. Long-term dynamic modelling of global population and built-up area in a spatially explicit way, HYDE 3.1. *Holocene* **2010**, *20*, 565–573. [CrossRef]
24. Sun, Y.G.; Chang, J.Y. A preliminary study on the means of livelihood from late yangshao period to longshan period in Northern Shaanxi Provinve. *J. Liaoning Norm. Univ. (Soc. Sci. Ed.)* **2018**, *41*, 110–117.

25. Sun, Z.Y.; Shao, J.; Liu, L.; Cui, J.; Bonomo, M.F.; Guo, Q.; Wu, X.; Wang, J. The first Neolithic urban center on China's north Loess Plateau: The rise and fall of Shimao. *Archaeol. Res. Asia* **2018**, *14*, 33–45. [CrossRef]
26. Zou, Y.L. The transition zone of agriculture-livestock in northern China and the change of cold and warm climate in Ming and Qing Dynasties. *FuDan J. (Soc. Sci.)* **1995**, *1*, 25–33.
27. Guo, L.Y.; Liu, Y.S.; Ren, Z.Y. Analysis of the land landscape changes and its driving mechanism in vulnerable ecological area: A case study of Yulin City. *Resour. Sci.* **2005**, *27*, 128–133.
28. Fang, X.Q.; Ye, Y.; Zeng, Z.Z. Extreme climate events, migration for cultivation and policies: A case study in the early Qing Dynasty of China. *Sci. China Ser. D Earth Sci.* **2007**, *50*, 411–421. [CrossRef]
29. Jia, K.L.; Chang, Q.R.; Zhang, J.H. Analysis of Land Use changes and Driving mechanisms in the Mixed agriculture-livestock production region in Northern Shaanxi. *Resour. Sci.* **2008**, *30*, 1053–1060.
30. Shi, X.L.; Shi, W.J. Review on boundary shift of farming-pastroal ecotone in northern China and its driving forces. *Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 1–11.
31. Yu, Y.Y.; Wu, H.B.; Finke, P.A.; Guo, Z. Spatial and temporal changes of prehistoric human land use in the Wei river valley, northern China. *Holocene* **2016**, *26*, 1788–1801. [CrossRef]
32. Hou, Y.J. The natural environment of the Ordos plateau and land use during the Ming and Qing dynasties. *J. Chin. Hist. Geogr.* **2007**, *22*, 28–39.
33. Qi, W.; Fan, Q.; Zhang, Y.Q.; Wang, J.; Hui, Z. Analysis of evolution and driving force of circular farmland in northern Shaanxi. *Geomat. Spat. Inf. Technol.* **2019**, *42*, 189–192.
34. Liu, X.K.; Dong, Z.B.; Ding, Y.P.; Lu, R.; Liu, L.; Ding, Z.; Li, Y. Development of center pivot irrigation farmlands from 2009 to 2018 in the Mu Us dune field, China: Implication for land use planning. *J. Geogr. Sci.* **2022**, *32*, 1956–1968. [CrossRef]
35. Sun, T.S.; Li, B.; Zhang, X.S. The response of agro-ecosystem productivity to climatic fluctuations in the farming-pastoral ecotone of northern China: A case study in Zhunger County. *Acta. Ecol. Sin.* **2012**, *32*, 6155–6167.
36. Gao, J.; Feng, L.Z.; Zhang, J.K.; Wang, Y.; Cao, M.; Chang, S.G. Effect of climate change on corn yield in Yulin, North Shaanxi province of China. *Chin. J. Agric. Resour. Reg. Plan.* **2016**, *37*, 118–124+135.
37. Wang, X.C.; Huang, Q.H.; Cai, Y.L.; Peng, J. Simulation of farmland spatial pattern in Maotiao river basin, Guizhou province. *Sci. Geogr. Sincia* **2007**, *27*, 188–192.
38. Yang, J.; Huang, X. The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019. *Earth Syst. Sci. Data* **2021**, *13*, 3907–3925. [CrossRef]
39. Xu, X.L. *China Population Spatial Distribution Kilometer Grid Dataset*; Resource and Environmental Science Data Registration and Publication System: Beijing, China, 2017. [CrossRef]
40. Qiao, Y. Development of complex societies in the Yiluo region: A GIS based population and agricultural area analysis. *Acta Archaeol. Sin.* **2010**, *179*, 423–454. [CrossRef]
41. Li, S.C.; He, F.N.; Chen, Y.S. Gridding Reconstruction of cropland spatial patterns in Southwest China in the Qing Dynasty. *Prog. Geogr.* **2012**, *31*, 1196–1203.
42. Wang, J.C.; Guo, Z.G. *Logistic Regression Models: Methods and Applications*; High Education Press: Beijing, China, 2001.
43. Monserud, R.A.; Leemans, R. Comparing global vegetation maps with the Kappa statistic. *Ecol. Model.* **1992**, *62*, 275–293. [CrossRef]
44. Deng, X.Z.; Zhan, J.Y. Scale-effect analysis of LUCC driving forces in the farming-pasturing interlocked area in Northern China. *Geogr. Geo-Inform. Sci.* **2004**, *20*, 64–68.
45. Tang, G.A.; Li, F.Y.; Yang, X.; Xiong, L.Y. *Exploration and Practice of Digital Terrain Analysis on Loess Plateau*; Science Press: Beijing, China, 2015; pp. 101–105.
46. Zhang, Y.; Chang, Q.R.; Zhao, Y.T.; Sun, L.P.; Wei, W. Land Use Structure and Its Spatial Distribution in Northern Shaanxi on the Loess Plateau. *Res. Soil Water Conserv.* **2014**, *21*, 72–78.
47. Li, X.B. Explanation of Land use Changes. *Prog. Geogr.* **2002**, *21*, 195–203.
48. The Local Compilation Committee of Jingbian County. *The Annals of Jingbian County*; Shaanxi People's Publishing House: Xi'an, China, 1993; pp. 43–44+49+55–59+106–118+152–169.
49. The Local Compilation Committee of Yulin County. *The Annals of Yulin County*; Sanqin Publishing House: Xi'an, China, 1996; pp. 76–87+91–103+106–110+164–174+177–178+185–202+204–207+223–227.
50. The Local Compilation Committee of Dingbian County. *The Annals of Dingbian County*; Local Annals Publishing House: Beijing, China, 2003; pp. 92–109+123–124+131–136+142–158+172–181.
51. The Local Compilation Committee of Shenmu County. *The Annals of Shenmu County*; Economic Daily Publishing House: Beijing, China, 1990; pp. 43–48+52–58+111+116–123+142–147.
52. The Local Compilation Committee of Hengshan County. *The Annals of Hengshan County*; Shaanxi People's Publishing House: Xi'an, China, 1993; pp. 67–74+87–92+99–100+145–151+243–250+253+257–260.
53. The Local Compilation Committee of Shaanxi Province. *The Annals of Shaanxi Province: Water Conservancy Chronicles*; Shaanxi People's Publishing House: Xi'an, China, 1999; pp. 96–98+191–192+283–284.

54. Liu, Y.H. *Study on Reservoir Construction and Utilization in Shaanxi Province (1950–2020)*; Shaanxi Normal University: Xi'an, China, 2020.
55. Yan, F.; Wu, B. Desertification progress in Mu Us Sandy land over the past 40 years. *Arid Land Geogr.* **2013**, *36*, 987–996.