Official Journal of the World Water Council
Water Policy



Water Policy Vol 23 No 6, 1388 doi: 10.2166/wp.2021.133

# Water poverty assessment based on the random forest algorithm: application to Gansu, Northwest China

Xiang Gao [a,†], Ke Wang[a,†], Kevin Lo [b,*], Ruiyang Wen[a], Xingxing Huang[a] and Qianwen Dang[a]

[a] College of Earth and Environmental Sciences, Lanzhou University, Lanzhou, China
[b] Department of Geography, Hong Kong Baptist University, Hong Kong, China
*Corresponding author. E-mail: lokevin@hkbu.edu.hk
†These authors contributed equally to this work.

XG, 0000-0001-6812-8156; KL, 0000-0001-7721-4726

## ABSTRACT

This study proposes a random forest algorithm to evaluate water poverty. It shows how the machine learning technique can be used to classify the degree of water poverty into five levels: very severe, severe, moderate, mild, and very mild. The strengths of the proposed random forest method include a high classification accuracy, good operational efficiency, and the ability to handle high-dimensional datasets. The success of the proposed method is empirically illustrated through a case study in Gansu, Northwest China. The analysis shows that from 2000 to 2017, the severity of water poverty in the study area declined. In 2000, most municipalities were classified as level 1 (very severe) or level 2 (severe). In 2017, level 1 water poverty disappeared, with most municipalities classified in as level 3 (moderate) and level 4 (mild). Spatially, there is a significant difference between the water poverty levels of the western, central, and eastern parts of Gansu, and the eastern part is affected by serious water poverty problems.

**Key words:** China, Gansu, Random forest algorithm, Water poverty, Water poverty index

## HIGHLIGHTS

- This study proposes a random forest algorithm to classify the level of water poverty.
- The proposed method is empirically illustrated through a case study in Gansu, Northwest China.
- The strengths of the proposed method include a high classification accuracy, good operational efficiency, and the ability to handle high-dimensional datasets.

## 1. INTRODUCTION

Water poverty can be defined as a situation in which a country or a region cannot provide sustainable, clean, and affordable water continuously for all people (Feitelson & Chenoweth, 2002). Water poverty is highly relevant to the United Nations 2030 Agenda for Sustainable Development, which sets clean water and sanitation as a sustainable development goal (Cetrulo *et al.*, 2020; Koirala *et al.*, 2020; Ladi *et al.*, 2021; Liu & Liu, 2021). Therefore, alleviating water poverty is a key objective of the water and development policy (Hope, 2015).

The water poverty index (WPI) has been widely adopted as a tool to indicate the extent to which societies are impacted by water poverty. The WPI is a multi-disciplinary index that considers physical and socioeconomical factors associated with water accessibility and affordability, thereby inviting policymakers to monitor the

availability of water resources and track the socioeconomical factors that shape water accessibility and affordability (Sullivan, 2002; Ladi *et al*., 2021). The influence of the WPI is palpable in the water poverty literature, with numerous studies applying this tool to examine the water poverty situation at the national (Jemmali, 2017; Pan *et al*., 2017; Goel *et al*., 2020; Ladi *et al*., 2021; Prince *et al*., 2021), regional (Huang *et al*., 2017; Thakur *et al*., 2017; Wurtz *et al*., 2019; Koirala *et al*., 2020), and local scales (Azqueta & Montoya, 2017; Kallio *et al*., 2018).

Whereas the WPI has become one of the most widely adopted assessment tools in the field of water poverty, new methods, models, and technologies are constantly being proposed to improve the WPI. One particularly promising approach is the use of machine learning techniques to aid in the classification of the water poverty levels. The evaluation of the WPI involves complex subsystems with the characteristics of multiple indicators, high dimensionality, and nonlinearity, and is therefore suitable for evaluation with the help of machine learning. Furthermore, machine learning can avoid the subjectivity of determining index weights that results from the rule-based method with human experts – the original WPI weighting system proposed by Sullivan (2002) is often criticized for its arbitrariness (Baquero *et al*., 2017). However, commonly used machine learning methods, such as back propagation (BP) neural networks and support vector machines (SVM), rely heavily on the distribution of training samples and may have problems such as insufficient robustness and over-fitting that, to a certain extent, affect the practicability and accuracy of the model (Jing *et al*., 2012; Wang, *et al*., 2016).

To address these problems, we developed a WPI evaluation method based on the random forest algorithm. Random forest is a non-parametric, supervised learning algorithm based on multiple decision tree classifiers (Breiman, 2001) and has been adopted by researchers to solve classification problems in diverse fields, including water resource management (Naghibi *et al*., 2017; Naghibi *et al*., 2019; Shirzad & Safari, 2019; Pahlavan-Rad *et al*., 2020). A random forest can be understood as a special bagging algorithm and involves the following steps (Ibrahim & Khatib, 2017). First, a bootstrapped dataset is generated from randomly sampling the original data with replacement (i.e., same sample can be selected more than once). Second, a decision tree is created by using the bootstrapped dataset, but only uses a random subset of features. Third, multiple decision trees are created by repeating the first two steps. Fourth, the classification results of each decision tree are integrated, and the most popular category is regarded as the final result. The introduction of randomness prevents the models from overfitting (Pal, 2005). Studies applying the random forest algorithm demonstrate that the method has strong advantages compared to other machine learning classification methods, including a good anti-noise ability, low error risk, and better performance in terms of accuracy and operating efficiency, particularly when there is a large dataset with many input variables (Gao *et al*., 2009; Cui & Bo, 2014; Lai *et al*., 2015).

We conducted a case study in Gansu, China, to demonstrate the proposed method. As a water-scarce country, water poverty is a serious issue in China (Pan *et al*., 2017; Liu *et al*., 2018). In particular, Northwest China is one of the most arid areas in East Asia and experiences severe water shortages and economic poverty (Lo *et al*., 2016; Rogers *et al*., 2020; Gao *et al*., 2021). With a population of 26 million, Gansu is the second most populous province in Northwest China. Hence, understanding the situation of water poverty in Gansu provides guidance to alleviate water poverty in China's arid and semi-arid areas.

In this study, 11 municipalities in Gansu were selected as basic evaluation units, including Lanzhou, Baiyin, Jiuquan, Jiayuguan, Jinchang, Wuwei, Pingliang, Qingyang, Tianshui, Dingxi, and Zhangye (Figure 1). These 11 municipalities form the Gansu Section of the Silk Road Economic Belt (hereafter, Gansu Section), which is long and narrow from the east to the west, with a length of approximately 1,555 km, accounting for 39% of the total length of the Silk Road in China. There are four types of geomorphic areas: the Longzhong region of the Loess Plateau, the Hexi Corridor, the Qilian Mountains, and the north of the Hexi Corridor. It has a significant continental temperate monsoon climate, with hydrothermal conditions decreasing from southeast to northwest. The Gansu Section covers three basins: the inland river basin, the Yellow River basin, and the Yangtze

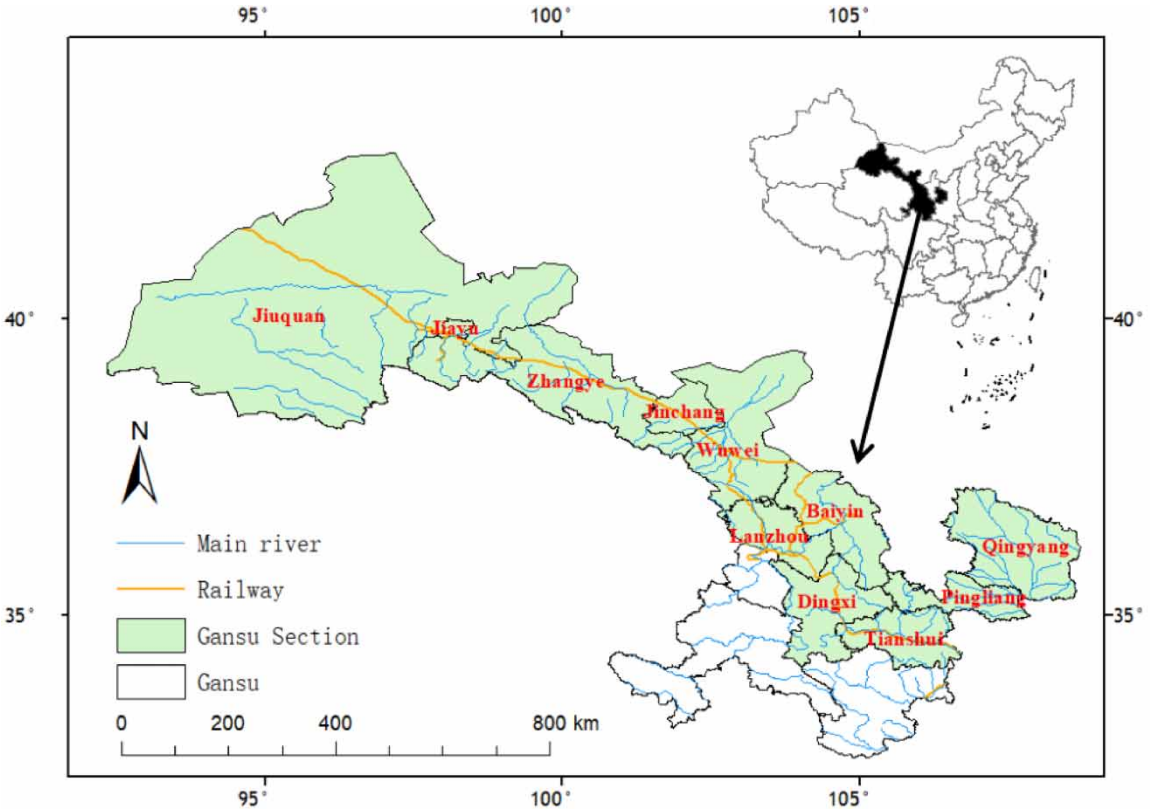


**Fig. 1.** | Location of study area.

River in the Hexi Corridor. Most of these areas are arid. In 2018, the total amount of water resources in the Gansu Section was 13.03 billion $m^3$, accounting for 36.7% of the province's total water resources; of which surface water resources were 12.12 billion $m^3$, accounting for 92.26% of the total water resources (Gansu Provincial Department of Water Resources, 2019).

The remainder of this paper proceeds as follows. First, we examine the steps involved in adopting the random forest algorithm to classify the levels of water poverty. Next, we analyze the results, both spatially and temporally. Finally, we offer certain concluding thoughts regarding the key lessons of this study.

## 2. ADOPTING THE RANDOM FOREST METHOD TO CLASSIFY LEVEL OF WATER POVERTY

### 2.1. Establishing the water poverty index system

The WPI typically includes five main components: resources, access, capacity, use, and environment (Ladi *et al.*, 2021). 'Resources' measure the availability of ground and surface water. 'Access' indicates the public availability of water resources. 'Capacity' measures a set of socioeconomical and institutional factors that influence water accessibility and affordability. 'Use' evaluates the amount of water use and water consumption efficiency in different sectors (for example, domestic, agricultural, industrial). Finally, 'environment' measures the environmental indicators related to water supply and management. Each of the five components of the WPI contains a set of criteria that can be used to calculate the composite index. Following this five-component approach, we developed a set of indicators suitable for China's local context. The evaluation index system comprised 17 positive and 8

negative indicators – a positive indicator implies that the larger the original data value, the better is the water poverty condition, whereas a negative indicator implies that the larger the original data value, the worse is the water poverty condition. The 25 indicators are shown in Table 1.

Based on the evaluation index system, the WPI can be calculated using the method developed by Sullivan (2002):

$$WPI = \frac{\omega_r R + \omega_a A + \omega_c C + \omega_u U + \omega_e E}{\omega_r + \omega_a + \omega_c + \omega_u + \omega_e} \quad (1)$$

where $R$, $A$, $C$, $U$, and $E$ *represent the five sub-components (resources, access, capability, use, and environment)*, and $\omega_r$, $\omega_a$, $\omega_c$, $\omega_u$, and $\omega_e$ represent the weights of each sub-component. Figure 2 shows the WPIs of 11 municipalities from 2000 to 2017.

**Table 1.** | Evaluation index system of water poverty.

| Component | Sub-component | Indicator | Code | Positive or negative |
|---|---|---|---|---|
| Resources | Utilizability | Percentage of water supply from other sources (%) | R1 | Positive |
| | | Per capita surface water resources ($m^3$/person) | R2 | Positive |
| | | Per capita groundwater resources ($m^3$/person) | R3 | Positive |
| | Variability | Water production modulus (%) | R4 | Positive |
| | | Coefficient of variation of precipitation (%) | R5 | Negative |
| Access | Water facility | Daily comprehensive urban water supply capacity ($m^3$/person/day) | A1 | Positive |
| | | Per capita water supply from water conservancy projects ($m^3$/person) | A2 | Positive |
| | | Density of urban water supply and drainage pipes (km/$km^2$) | A3 | Positive |
| | Approach to use water | Tap water penetration rate (%) | A4 | Positive |
| | | Water saving irrigation rate (%) | A5 | Positive |
| Capability | Economic foundation | Per capita GDP (yuan) | C1 | Positive |
| | | Urban household disposable income (yuan) | C2 | Positive |
| | Social welfare | Practitioners and assistants per thousand people | C3 | Positive |
| | | Number of middle school students among ten thousand residents | C4 | Positive |
| | Management of water resources | Fiscal revenue and expenditure ratio | C5 | Positive |
| | | Ratio of R&D expenditure | C6 | Positive |
| Use | Utilization efficiency | Water consumption per 10,000 yuan of industrial added value ($m^3$) | U1 | Negative |
| | | Water consumption per 10,000 yuan of GDP ($m^3$) | U2 | Negative |
| | | Water usage per unit of food production ($m^3$/t) | U3 | Negative |
| | | Percentage of water used in agriculture (%) | U4 | Negative |
| Environment | Ecological pressure | Percentage of drought-affected area | E1 | Negative |
| | | Fertilizer application intensity (kg/$hm^2$) | E2 | Negative |
| | Intensity of water utilization | Per capita domestic water consumption ($m^3$/person) | E3 | Negative |
| | Environmental governance | Green coverage rate in built-up areas (%) | E4 | Positive |
| | | Daily treatment capacity of urban sewage treatment plants ($m^3$/day) | E5 | Positive |

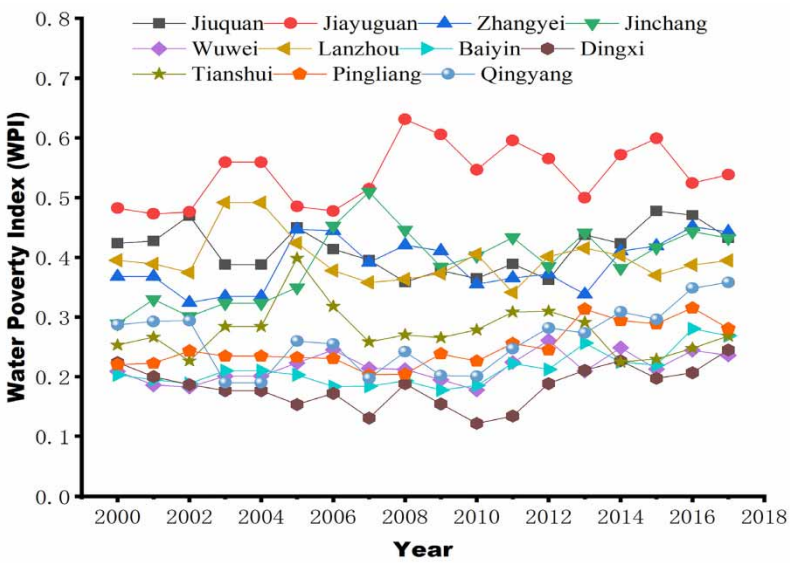


**Fig. 2.** | Changes in Water Poverty Index in 11 municipalities.

### 2.2. Setting classification criteria

The criteria for classification were developed based on the evaluation index system of water poverty. To create the classification criteria, the values of the water poverty indicators were divided into five levels: strongly negative (1), negative (2), neutral (3), positive (4), and strongly positive (5), based on the data from the Gansu Section. We collected data on water poverty from the *Gansu Statistical Yearbook* (2001–2020), *Gansu Provincial Water Conservancy Statistical Yearbook* (2001–2019), *Gansu Water Resources Statistical Bulletin* (2000–2017), and *China City Statistical Yearbook* (2001–2019). Then, we preprocessed the data, where the positive and negative indicators were normalized according to Equations (2) and (3), respectively (Wei *et al.*, 2021).

$$X_{ij} = \frac{X_{ij} - \min\{X_j\}}{\max\{X_j\} - \min\{X_j\}} \tag{2}$$

$$X_{ij} = \frac{\max\{X_j\} - X_{ij}}{\max\{X_j\} - \min\{X_j\}} \tag{3}$$

In Equations (2) and (3), $X_{ij}$ represents the value of the j-th evaluation index of the i-th municipality, min $\{X_i\}$ represents the minimum value of the j-th evaluation index in all years, and max $\{X_j\}$ represents the maximum value. Then, we used Natural Breaks to divide the processed values into five levels, with a total of $25 \times 5$ label information. Table 2 presents the resulting classification criteria for water poverty. The range of the different levels of the indices may be discontinuous because we used the distribution of the original data to determine the range of the different levels. Considering R5 as an example, the original data are typically concentrated below 98, with a few outliers greater than 128, and there are no values between 98 and 128. Therefore, there is a gap between levels 1 and 2.

**Table 2.** | Classification criteria of water poverty in random forest.

| Code | Unit | Level 1 (strongly negative) | Level 2 (negative) | Level 3 (neutral) | Level 4 (positive) | Level 5 (strongly positive) |
|---|---|---|---|---|---|---|
| R1 | % | ≤0.04 | [0.04, 0.08) | [0.08, 0.12) | [0.12, 0.16) | ≥0.16 |
| R2 | $m^3$/person | ≤700 | [700, 1,300) | [1,300, 1,900) | [1,900, 2,500) | ≥2,500 |
| R3 | $m^3$/person | ≤800 | [800, 1,400) | [1,400, 2,200) | [2,200, 2,900) | ≥2,900 |
| R4 | % | ≤4 | [4, 8) | [8, 11) | [11, 15) | ≥15 |
| R5 | % | ≥128 | [68, 98) | [38, 68) | [8, 38) | ≤8 |
| A1 | $m^3$/person/day | ≤1.1 | [1.1, 2) | [2, 2.9) | [2.9, 3.8) | ≥3.8 |
| A2 | $m^3$/person | ≤500 | [500, 1,000) | [1,000, 1,500) | [1,500, 1,900) | ≥1,900 |
| A3 | km/$km^2$ | ≤17 | [17, 31) | [31, 44) | [44, 58) | ≥58 |
| A4 | % | ≤32 | [32, 49) | [49, 66) | [66, 83) | ≥83 |
| A5 | % | ≤0.4 | [0.4, 0.8) | [0.8, 1.2) | [1.2, 1.6) | ≥1.6 |
| C1 | $10^4$ yuan | ≤2.5 | [2.5, 4.7) | [4.7, 7) | [7, 9.3) | ≥9.3 |
| C2 | $10^4$ yuan | ≤0.85 | [0.85, 1.60) | [1.6, 2.4) | [2.4, 3.2) | ≥3.2 |
| C3 | person | ≤240 | [240, 480) | [480, 720) | [720, 950) | ≥950 |
| C4 | person | ≤200 | [200, 330) | [330, 470) | [470, 660) | ≥600 |
| C5 | % | ≤0.3 | [0.3, 0.6) | [0.6, 0.9) | [0.9, 1.1) | ≥1.1 |
| C6 | % | ≤0.5 | [0.5, 1) | [1, 1.4) | [1.4, 1.9) | ≥1.9 |
| U1 | $m^3$ | ≥550 | [340, 440) | [226, 340) | [120, 226) | ≤120 |
| U2 | $m^3$ | ≥3,800 | [2,300, 3,000) | [1,800, 2,300) | [780, 1,800) | ≤780 |
| U3 | $m^3$/t | ≥11,700 | [7,000, 9,300) | [4,700, 7,000) | [2,400, 4,700) | ≤2,400 |
| U4 | % | ≥1.6 | [1.1, 1.3) | [0.8, 1.1) | [0.5, 0.8) | ≤0.5 |
| E1 | % | ≥0.8 | [0.7, 0.8) | [0.5, 0.7) | [0.2, 0.5) | ≤0.2 |
| E2 | kg/$hm^2$ | ≥450 | [270, 360) | [180, 270) | [90, 180) | ≤90 |
| E3 | $m^3$/person | ≥460 | [280, 380) | [200, 280) | [120, 200) | ≤120 |
| E4 | % | ≤14 | [14, 24) | [24, 34) | [34, 44) | ≥44 |
| E5 | $m^3$/day | ≤16 | [16, 30) | [30, 44) | [44, 58) | ≥58 |

*Note*: () denote open interval boundaries, [] denote closed boundaries.

### 2.3. Generating training and testing sets

We randomly selected 100 sets of samples from each classification level (that is, the total number of samples was 500). We randomly extracted 70% of the sample and designated them as the training set, which was used to train the classification tree and generate a label classifier. We used the remaining 30% of the samples as the test set. We imported training and test samples into the R software.

### 2.4. Optimizing model parameters

After importing the training and testing sets, we loaded the random forest package in R and set the ntree and mtry values to run the algorithm. We tested the ntree values of 100, 300, 500, 800, and 1,000. Figure 3 shows the trend of the out-of-bag (OOB) error of the random forest classification model for water poverty in Gansu with different

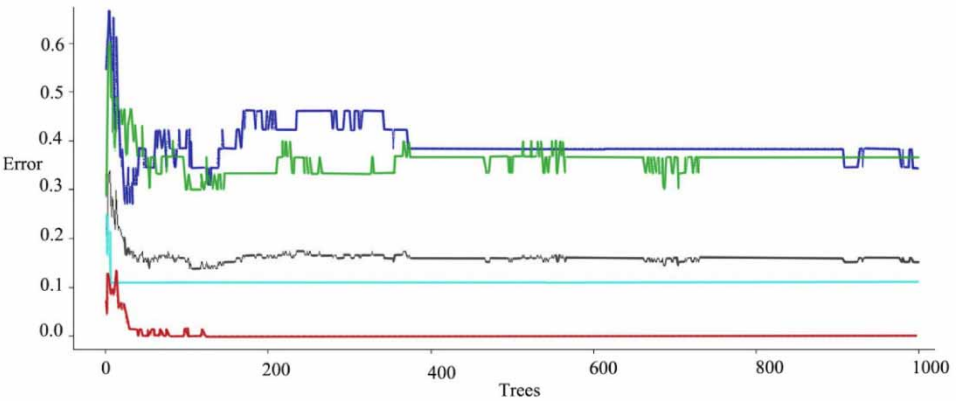


**Fig. 3.** | The OOB error of random forest classification with different ntree values.

ntree values. The error rate value in the range of 0–400 was relatively large, with significant fluctuations. When the ntree value was 800, the error converged. Therefore, we set the ntree value to 800. Subsequently, with a fixed ntree value of 800, the k-fold cross-validation method was used to traverse the mtry parameters to determine the best mtry value. K-fold cross-validation can ensure that when the total number of data samples is small, each subsample participates in training and testing, effectively reducing the generalization error of the model. When mtry is 8, the minimum error value reaches the inflection point; therefore, we set the mtry parameter value to 8.

### 2.5. Training random forests

Using the optimized parameters, a random forest model was trained using the samples. As shown in Table 3, the trained random forest model was the most accurate for classifying 'very severe' water poverty level, and the classification error was 0%. The accuracy was also high for 'mild' water poverty level, with a classification error of 11.11%. The classification error for the 'severe' type was 33.33%, and 1/6-th of the samples were misjudged as 'moderate' and 'very severe'. The classification error for the 'moderate' water poverty level was the highest (38.46%), with nearly 30% of the samples misclassified as 'severe,' and one sample misjudged as 'very severe' and one sample as 'mild'. The OOB error in the confusion matrix was calculated to be 15.2%.

### 2.6. Running the random forest model

Finally, we applied the random forest model to classify the water poverty levels. We input the actual data of each indicator in each municipality in the study area from 2000 to 2017 as new data into the software. The level corresponding to the highest probability in the result was assigned to the actual yearly water poverty level of the

**Table 3.** | Confusion matrix of the running results of the test set model.

| | Very severe | Severe | Moderate | Mild | Classification error |
|---|---|---|---|---|---|
| Very severe | 73 | 0 | 0 | 0 | 0.00% |
| Severe | 5 | 20 | 5 | 0 | 33.33% |
| Moderate | 1 | 8 | 16 | 1 | 38.46% |
| Mild | 0 | 0 | 1 | 8 | 11.11% |

random forest. Because the random forest algorithm is a classification tool, the classification results are integers. Table 4 presents the results of the classification.

## 3. TEMPORAL AND SPATIAL ANALYSIS OF WATER POVERTY LEVEL IN THE GANSU SECTION

The classification results indicated that in the Gansu Section the problem of water poverty was serious, but it gradually improved. In 2000, 9 of the 11 municipalities suffered from extremely severe or severe water poverty; in 2017, there was no more extremely severe water poverty, and only three municipalities suffered from severe water poverty. More specifically, we can distinguish between the three stages from 2000 to 2017. In the first stage (2000–2005), water poverty in the Gansu Section was serious, with level 2 being the most frequent, followed by levels 1, 3, 4, and 5. In the second stage (2006–2011), the situation of water poverty improved, with levels 3 and 2 being the highest frequency of water poverty levels, and the number of municipalities in level 2 significantly decreased. In the third stage (2012–2017), the situation of water poverty continued to improve. Significantly, there were no municipalities in level 1 in this stage.

Spatially, different municipalities in the Gansu Section exhibit different patterns of water poverty. Water poverty in Jiayuguan is relatively insignificant and has remained at level 5 for several years, and Lanzhou is another mild water poverty area with an index of level 4 for several years. As for the moderate water poverty areas, Jinchang, Jiuquan, and Zhangye have held the level of water poverty mainly at level 3 for several years. Other municipalities, including Baiyin, Tianshui, Qingyang, Dingxi, and Pingliang, are severely water poverty areas. They have experienced water poverty at level 2 or level 1 for several years.

**Table 4.** | Classification results of water poverty levels in each municipality.

| Year | Jiuquan | Jiayuguan | Zhangye | Jinchang | Wuwei | Lanzhou | Baiyin | Dingxi | Tianshui | Pingliang | Qingyang |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 | 2 | 4 | 2 | 2 | 1 | 3 | 2 | 1 | 1 | 1 | 1 |
| 2001 | 2 | 4 | 3 | 1 | 2 | 3 | 2 | 1 | 1 | 1 | 1 |
| 2002 | 2 | 5 | 2 | 2 | 2 | 4 | 2 | 2 | 1 | 1 | 2 |
| 2003 | 3 | 5 | 3 | 5 | 2 | 4 | 2 | 1 | 1 | 1 | 2 |
| 2004 | 3 | 5 | 2 | 5 | 2 | 3 | 2 | 1 | 2 | 1 | 3 |
| 2005 | 3 | 3 | 4 | 3 | 3 | 4 | 2 | 2 | 2 | 2 | 3 |
| 2006 | 2 | 3 | 4 | 3 | 3 | 3 | 2 | 1 | 2 | 1 | 2 |
| 2007 | 3 | 5 | 3 | 4 | 3 | 3 | 3 | 2 | 2 | 1 | 2 |
| 2008 | 3 | 5 | 5 | 3 | 3 | 5 | 3 | 3 | 3 | 1 | 2 |
| 2009 | 3 | 5 | 3 | 4 | 2 | 4 | 2 | 1 | 2 | 2 | 3 |
| 2010 | 3 | 5 | 4 | 3 | 3 | 3 | 2 | 1 | 2 | 1 | 3 |
| 2011 | 3 | 5 | 4 | 4 | 3 | 4 | 3 | 2 | 3 | 1 | 2 |
| 2012 | 3 | 5 | 3 | 4 | 3 | 4 | 3 | 3 | 3 | 5 | 2 |
| 2013 | 3 | 4 | 4 | 5 | 2 | 4 | 2 | 2 | 3 | 2 | 2 |
| 2014 | 3 | 5 | 4 | 3 | 4 | 4 | 2 | 3 | 3 | 4 | 3 |
| 2015 | 3 | 5 | 2 | 5 | 3 | 5 | 3 | 2 | 2 | 5 | 2 |
| 2016 | 4 | 5 | 3 | 3 | 3 | 4 | 3 | 2 | 3 | 2 | 3 |
| 2017 | 3 | 5 | 4 | 4 | 2 | 4 | 3 | 2 | 3 | 2 | 3 |

Figure 4 presents the spatial distribution of the water poverty levels in the research area. In 2000, while most municipalities had severe water poverty issues, western and middle Gansu were better than eastern Gansu. From 2000 to 2017, the water poverty level of most cities steadily improved, particularly for municipalities in eastern and central Gansu. In 2017, western Gansu (Jiuquan) was a medium-value area, most of the municipalities in central Gansu (Jiayuguan, Zhangye, Jinchang, and Lanzhou) had good results in terms of water poverty levels, and the low-value areas were mainly distributed in the eastern section (Pingliang, Dingxi), where water poverty remains serious.

What could the municipalities with serious water poverty problems learn from the better-performing municipalities? To accurately answer this question, it is necessary to determine the constraints of water poverty in each municipality and quantify the degree of its impact. The municipalities with serious water poverty issues typically have a poor water resource availability and low water resource utilization efficiency. The municipalities with mild water poverty levels benefited from sustained and comprehensive water management measures, as evidenced by their high water resource utilization efficiency. This is a successful experience that other municipalities should learn from.

Considering Pingliang, a municipality suffering from serious water poverty, as an example, two indicators, the percentage of water supply from other sources (R1) and percentage of the drought-affected area (E1), significantly

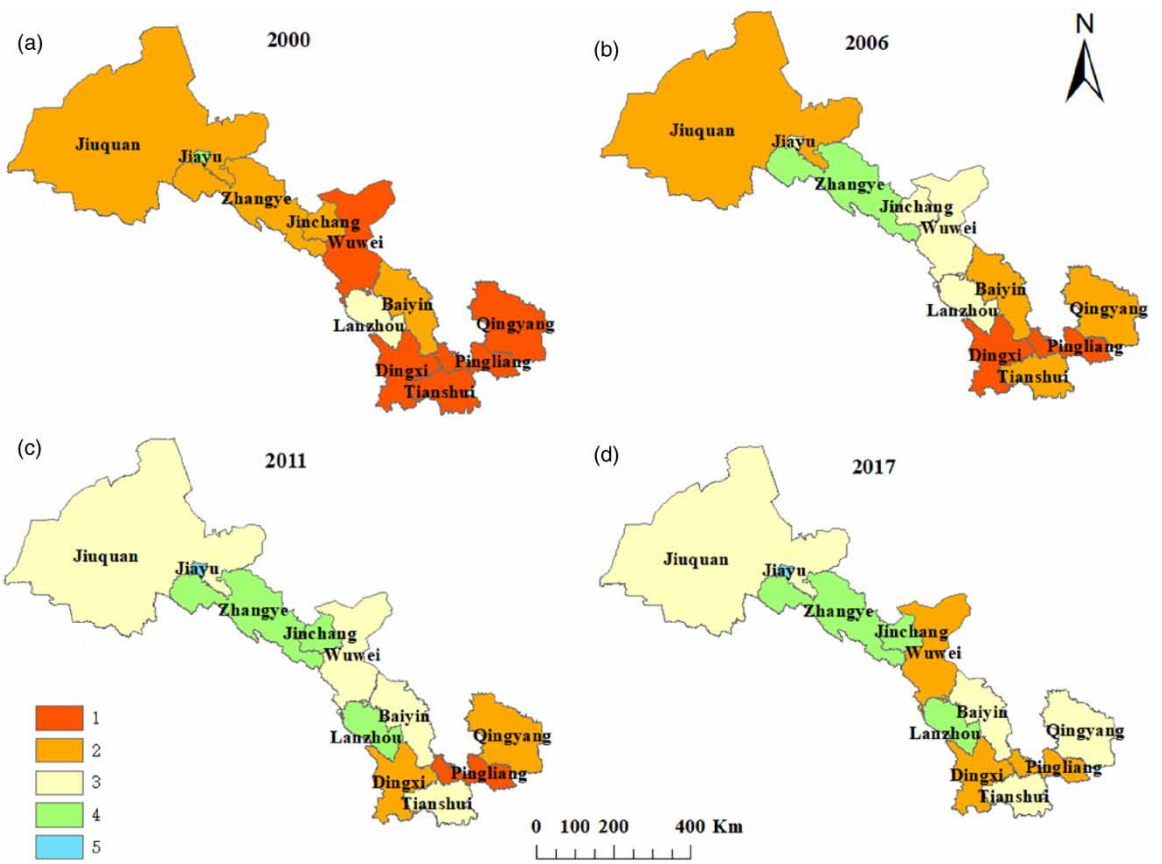


**Fig. 4.** | Spatial distribution of water poverty levels in different years.

impact the classification of water poverty. First, other water sources mainly refer to unconventional water sources such as sewage treated water, which reduce the dependency on surface and underground runoff. Pingliang should take management and technological measures to expand unconventional water sources as an important supplement to conventional water sources. Second, the drought area refers to the sown area in which the actual harvest of crops is reduced by more than 30% compared to the normal annual output in the drought-stricken area. Measures such as improving the monitoring of soil moisture and crop growth and enhancing irrigation infrastructure can reduce the impact of drought.

## 4. CONCLUSION

Over the past two decades, support for addressing water poverty has been increasing worldwide. Quantitative, index-based analyses such as the WPI play an important role in monitoring and understanding water poverty problems. The random forest algorithm offers several advantages for classifying the level of water poverty. The random forest method can process high-dimensional data with less human intervention and faster training. The dimensionless processing of the original data is not required. Only a preliminary training of the model is required to run the evaluation results. The random forest algorithm is suitable for processing datasets with a large number of unknown features, without feature selection. The application to Gansu shows that the random forest method can obtain a comprehensive and highly reliable spatiotemporal evaluation of water poverty.

## CODE AVAILABILITY

Not applicable.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## FUNDING

This research was funded by the National Key Research and Development Program of China (Grant No.2019YFC0507402).

## DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

## REFERENCES

Azqueta, D. & Montoya, Á. (2017). The social benefits of water and sanitation projects in Northern Colombia: cost-benefit analysis, the water poverty index and beyond. *Development Policy Review 35*, O118–O139.

Baquero, O. F., Gallego-Ayala, J., Giné-Garriga, R., de Palencia, A. J. F. & Pérez-Foguet, A. (2017). The influence of the human rights to water and sanitation normative content in measuring the level of service. *Social Indicator Research 133*, 763–786.

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Cetrulo, T. B., Marques, R. C., Malheiros, T. F. & Cetrulo, N. M. (2020). Monitoring inequality in water access: challenges for the 2030 Agenda for Sustainable Development. *Science of the Total Environment 727*, 138746.

Cui, D. & Bo, J. (2014). Comprehensive evaluation of water ecological civilization based on random forest algorithm. *Advance in Science and Technology of Water Resources 34*, 56–60.

Feitelson, E. & Chenoweth, J. (2002). Water poverty: towards a meaningful indicator. *Water Policy 4*, 263–281.

Gansu Provincial Department of Water Resources (2019). *Gansu Province Water Resources*. Available at: http://slt.gansu.gov.cn/xxgk/gkml/nbgb/szygb/201911/t20191111_122952.html.

Gao, D., Zhang, Y. X. & Zhao, Y. H. (2009). Random forest algorithm for classification of multiwavelength data. *Research in Astronomy and Astrophysics 9*, 220.

Gao, X., Wang, K., Lo, K., Wen, R., Mi, X., Liu, K. & Huang, X. (2021). An evaluation of coupling coordination between rural development and water environment in Northwestern China. *Land 10*, 405.

Goel, I., Sharma, S. & Kashiramka, S. (2020). The water poverty index: an application in the Indian context. *Natural Resources Forum 44*, 195–218.

Hope, R. (2015). Is community water management the community's choice? Implications for water and development policy in Africa. *Water Policy 17*, 664–678.

Huang, S., Feng, Q., Lu, Z., Wen, X. & Deo, R. C. (2017). Trend analysis of water poverty index for assessment of water stress and water management polices: a case study in the Hexi Corridor, China. *Sustainability 9*, 756.

Ibrahim, I. A. & Khatib, T. (2017). A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. *Energy Conversion and Management 138*, 413–425.

Jemmali, H. (2017). Mapping water poverty in Africa using the improved Multidimensional Index of Water Poverty. *International Journal of Water Resources Development 33*, 649–666.

Jing, G., Du, W. & Guo, Y. (2012). Studies on prediction of separation percent in electrodialysis process via BP neural networks and improved BP algorithms. *Desalination 291*, 78–93.

Kallio, M., Guillaume, J. H., Kummu, M. & Virrantaus, K. (2018). Spatial variation in seasonal water poverty index for Laos: an application of geographically weighted principal component analysis. *Social Indicator Research 140*, 1131–1157.

Koirala, S., Fang, Y., Dahal, N. M., Zhang, C., Pandey, B. & Shrestha, S. (2020). Application of water poverty index (WPI) in spatial analysis of water stress in Koshi River Basin, Nepal. *Sustainability 12*, 727.

Ladi, T., Mahmoudpour, A. & Sharifi, A. (2021). Assessing impacts of the water poverty index components on the human development index in Iran. *Habitat International 113*, 102375.

Lai, C., Chen, X., Zhao, S., Wang, Z. & Wu, X. (2015). A flood risk assessment model based on random forest and its application. *Journal of Hydraulic Engineering 46*, 58–66.

Liu, Z. & Liu, W. (2021). Spatial-temporal relationship between water resources and economic development in rural China from a poverty perspective. *International Journal of Environmental Research and Public Health 18*, 1540.

Liu, W., Zhao, M. & Xu, T. (2018). Water poverty in rural communities of arid areas in China. *Water 10*, 505.

Lo, K., Xue, L. & Wang, M. (2016). Spatial restructuring through poverty alleviation resettlement in rural China. *Journal of Rural Studies 47*, 496–505.

Naghibi, S. A., Ahmadi, K. & Daneshi, A. (2017). Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resources Management 31*, 2761–2775.

Naghibi, S. A., Vafakhah, M., Hashemi, H., Pradhan, B. & Alavi, S. J. (2019). Water resources management through flood spreading project suitability mapping using frequency ratio, k-nearest neighbours, and random forest algorithms. *Natural Resources Research 29*, 1915–1933.

Pahlavan-Rad, M. R., Dahmardeh, K., Hadizadeh, M., Keykha, G., Mohammadnia, N., Gangali, M., Keikha, M., Davatgar, N. & Brungard, C. (2020). Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern Iran. *Catena 194*, 104715.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing 26*, 217–222.

Pan, A., Bosch, D. & Ma, H. (2017). Assessing water poverty in China using holistic and dynamic principal component analysis. *Social Indicator Reserach 130*, 537–561.

Prince, B. C., Juran, L., Sridhar, V., Bukvic, A. & MacDonald, M. C. (2021). A statistical and spatial analysis of water poverty using a modified Water Poverty Index. *International Journal of Water Resources Development 37*, 339–356.

Rogers, S., Li, J., Lo, K., Guo, H. & Li, C. (2020). China's rapidly evolving practice of poverty resettlement: moving millions to eliminate poverty. *Development Policy Review 38*, 541–554.

Shirzad, A. & Safari, M. J. S. (2019). Pipe failure rate prediction in water distribution networks using multivariate adaptive regression splines and random forest techniques. *Urban Water Journal 16*, 653–661.

Sullivan, C. (2002). Calculating a water poverty index. *World Development 30*, 1195–1210.

Thakur, J. K., Neupane, M. & Mohanan, A. A. (2017). Water poverty in upper Bagmati River basin in Nepal. *Water Science 31*, 93–108.

Wang, F., Tian, G., Wang, X., Liu, Y., Deng, S., Wang, H. & Zhang, F. (2016). Application of genetic algorithm-back propagation for prediction of mercury speciation in combustion flue gas. *Clean Technologies and Environmental Policy 18*, 1211–1218.

Wei, D., Liu, B., Duan, Z. & Yang, W. (2021). Measuring local progress of the 2030 Agenda for SDGs in the Yangtze River Economic Zone, China. *Environment, Development and Sustainability*. https://doi.org/10.1007/s10668-021-01743-z.

Wurtz, M., Angeliaume, A., Herrera, M. T. A., Blot, F., Paegelow, M. & Reyes, V. M. (2019). A spatial application of the water poverty index (WPI) in the State of Chihuahua, Mexico. *Water Policy 21*, 147–161.

First received 3 June 2021; accepted in revised form 11 October 2021. Available online 26 October 2021