Applied Geography 135 (2021) 102552

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Combining night time lights in prediction of poverty incidence at the county level

Jianbin Xu [a], Jie Song [b], Baochao Li [b], Dan Liu [b], Xiaoshu Cao [c,d,*]

[a] College of Resources and Environment, Shanxi University of Finance and Economics, Taiyuan, 030006, China
[b] School of Geography and Planning, Sun Yat-sen University, Guangzhou, 510275, China
[c] Shaanxi Normal University Academy of Natural Resources and Territorial Space, Shaanxi Normal University, Xi'an, 710119, China
[d] Yunnan-Guizhou Plateau Observation Station of Coupled Human and Natural System, Shaanxi Normal University, Xi'an, 710119, China

ARTICLE INFO

Keywords:
Night time lights
Machine learning
Poverty estimate
Long sequence
Accuracy
YGGRD

ABSTRACT

Long-term poverty data can support accurate decision-making. This study demonstrates an accurate and reliable method for identifying poverty areas and predicting poverty incidence based on night time light remote-sensing data and machine learning methods. Using data of poverty counties and poverty incidence in Guizhou Province of China as the training dataset, we show how to use machine learning to identify poverty counties and predict poverty incidence in the Yunnan-Guangxi-Guizhou Rocky desertification area. The identification accuracy of poverty-stricken counties was 76.5%. The root mean squared error, mean absolute error, and R^2 values of the poverty incidence rates were 5.01, 4.04, and 0.60, respectively. Using data from 2015 to verify the trained model, the R^2 value of the predicted and actual values of poverty incidence reached 0.95. With the progress in machine learning and night light remote sensing, poverty mapping combined with night time lights and machine learning can compensate for the data gap in deprived areas and provide a decision-making basis for sustainable development in poverty-stricken areas.

1. Introduction

Poverty remains a problem worldwide; eradication of extreme poverty in all its forms is the first goal of the Sustainable Development Goals (Liu, Guo, & Zhou, 2018; Mani, Mullainathan, Shafir, & Zhao, 2013; Subash, Kumar, & Aditya, 2018; Zhou & Liu, 2019). The global distribution of poor people is extremely unbalanced, mainly in the developing countries of Africa and Asia. Accurate poverty data are important for solving the poverty problem. Although the quality of economic development data in developing countries has improved in recent years, reliable and relevant data on poverty are still insufficient (Jean et al., 2016a). In India, the data from sample surveys claiming more accuracy are infrequent, and the census data are decadal and lack validation (Subash et al., 2018). According to a World Bank report, during 2000–2010, only 66% of African countries conducted active surveys related to poverty. Thus, it is difficult to formulate poverty reduction measures based on these surveys (Jean et al., 2016a; Yeh et al., 2020). By the end of 2020, China had eliminated absolute poverty through education, medical care, photovoltaic-based intervention, counterpart-assistance (*Dui kou bang fu*), resettlement, and land consolidation (Guo, Zhou, & Liu, 2019; Liao, Fei, Huang, Jiang, & Shi, 2021; Lo, Xue, & Wang, 2016). Since the implementation of the targeted poverty alleviation policy, China has established archives for poor households. However, these data are confidential and not open to the public; thus, it is difficult for researchers to obtain.

The World Bank has implemented various poverty mapping methods in poverty-stricken regions, such as the small area estimation method, to address the lack of poverty data (Isidro, Haslett, & Jones, 2016). The small area estimation method combines sample survey, census, and administrative data to obtain a small area of poverty estimates. However, a census is only conducted once every ten years in most developing countries (Subash et al., 2018). Moreover, the accessibility of large-scale surveys is limited, and they are expensive. Generally, surveys place enormous financial pressure on deprived areas, and the difficulty of implementation is predictable (Pan, Zhao, & Dong, 2018). Additionally, some scholars use a mixed model and composite index to estimate multidimensional poverty. However, these data are also obtained through sample surveys (Alkire & Foster, 2011; Anderson, Farcomeni,

* Corresponding author. Yunnan-Guizhou Plateau Observation Station of Coupled Human and natural System, Shaanxi Normal University, Xi'an, 710119, China.
E-mail addresses: xujb23@mail2.sysu.edu.cn (J. Xu), songj36@mail2.sysu.edu.cn (J. Song), libch9@mail2.sysu.edu.cn (B. Li), liud69@mail2.sysu.edu.cn (D. Liu), caoxsh@snnu.edu.cn (X. Cao).

https://doi.org/10.1016/j.apgeog.2021.102552
Received 9 August 2020; Received in revised form 15 August 2021; Accepted 21 August 2021
Available online 31 August 2021

Pittau, & Zelli, 2016; Erenstein, Hellin, & Chandna, 2010; Sadath & Acharya, 2017).

Compared with traditional poverty measurement methods, satellite remote-sensing data have a higher temporal and spatial resolution; owing to their relation to high-altitude observations, human subjective factors are limited. Additionally, the observation results are more objective (Wang, Cheng, & Zhang, 2012). Since the 1960s, many studies have shown that satellite remote-sensing data can retrieve the socio-economic characteristics of a specific area (Foster, 1983; Tobler, 1969; Welch, 1980). The rapid development of night light remote sensing technology since the 1990s has promoted the development of research in the field of night time lights. Night time lights directly embody the lights of urban areas at night and indicate the level of regional development. Therefore, many scholars estimate the economic development level of a region using night time lights (Doll, Muller, & Morley, 2006; Forbes, 2013; Kim, 1997), population index (Kim, 1997), built-up area (Shi, Huang, et al., 2014), power consumption index (Propastin & Kappas, 2012), and other socioeconomic indicators (Ebener, Murray, Tandon, & Elvidge, 2005).

Night time light remote sensing data have been used to simulate regional poverty. Most similar studies establish the relationship model between the characteristic variables of night light data and poverty index and then predict poverty (Noor, Alegana, Gething, Tatem, & Snow, 2008). Currently, the simulation of poverty using night light data is mostly at the national and provincial scales, and there are few relevant studies at the county level (Elvidge et al., 2009; Subash et al., 2018). Poverty estimation or measurement and its geographical identification at the county level are necessary and cannot be ignored. Geographical targeting of poverty is a viable way to allocate resources for poverty alleviation, especially for smaller geographic areas such as the county level (Bigman & Fofack, 2000; Zhou & Liu, 2019). Many studies have explored the linkages between night light and poverty using the average light index of night light data and the integrated poverty index at the county level (Pan et al., 2018; Wang et al., 2012; Yu et al., 2015). In addition to the average light index and total light index, diversified night light variables and machine learning can also identify poor counties (Li, Cai, Liu, Liu, & Su, 2019). The most widely used night time light data are DMSP/OLS (1992–2013) and NPP/VIIRS (2012–2020). Although existing studies attempt to solve the problem of inconsistency between the two data resolutions, most of the current poverty simulation studies only use one of the data sources, significantly limiting the long-term simulation and research on regional poverty (Chen & Nordhaus, 2010; Henderson, Storeygard, & Weil, 2012; Mellander, Lobo, Stolarick, & Matheson, 2015). Therefore, this study aims to use DMSP/OLS and NPP/VIIRS night time lights data, combined with machine learning and statistical data (statistical yearbook data), to predict poverty in long-term time series county-level areas lacking data.

Section 2 introduces night time lights data and machine learning methods. Section 3 covers the predicted result using China's Yunnan-Guangxi-Guizhou Rocky desertification (YGGRD) area as the study area, one of the contiguous destitute areas with the highest number of poor people in China. Section 4 discusses data processing, error distribution, and application of model prediction. Finally, Section 5 concludes this paper.

2. Materials and methods

2.1. Study area

The YGGRD area is in southwestern China (Fig. 1), in the transitional zone between the Yunnan-Guizhou Plateau and the Guangxi Basin, bordering Vietnam to the south. The area comprises a typical plateau and mountainous terrain, and the carbonate rocks are widely distributed. The YGGRD area is one of the most typical areas for karst landform development worldwide.

The YGGRD area has the highest number of poor people among the 14 contiguous poverty-stricken regions in China. According to the census, the poverty incidence in this area in 2018 was 5.3%, with 720,000 people living in poverty; therefore, it is a typical poverty-stricken mountainous area. However, the data for each sub-region (county or district) in Yunnan and Guangxi are not available to the public, making it difficult for researchers to explore the associated poverty patterns and formulate corresponding measures. China's population census is conducted every ten years, and the latest was the sixth

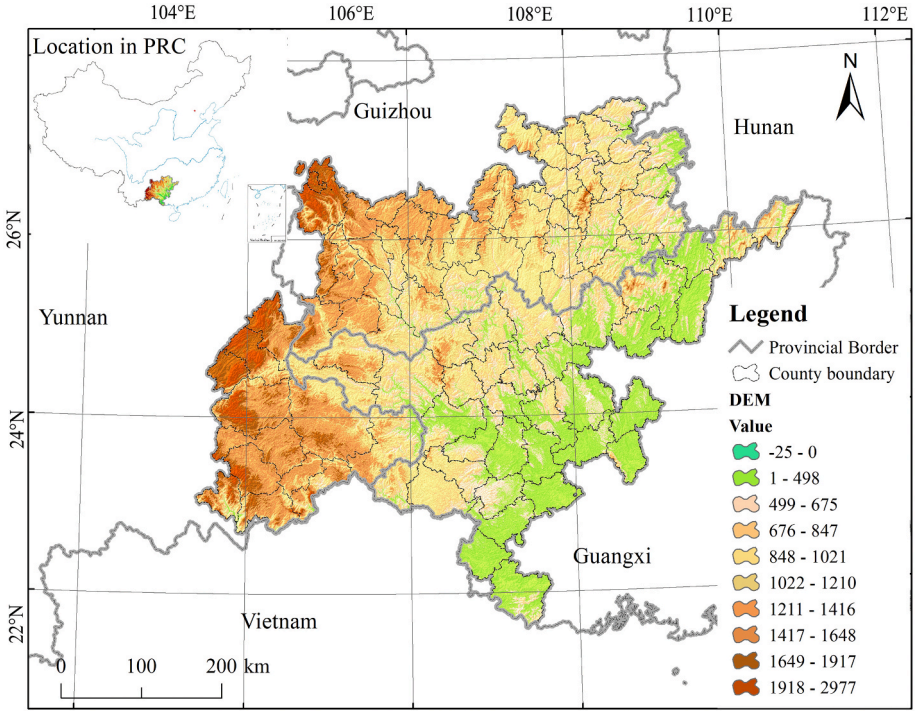

Fig. 1. The study area.

census in 2010. Therefore, there is a problem of timeliness in using the small area estimation method for poverty mapping. Recent publicly available data about poverty are taken from the county poverty incidence rate reported in the *2015 Statistical Yearbook of Guizhou Province*. The poverty incidence is one of the most widely used indicators to measure poverty, which means the proportion of the population whose per capita income or consumption expenditure does not meet the poverty standard in the whole population (Xian, Wang, & Wu, 2016). Other regions do not publish county-level poverty data. Therefore, based on the poverty incidence data of 88 counties in Guizhou Province from 2013 to 2015, combined with night time lights data, the poverty counties in Yunnan, Guangxi, and Guizhou rocky desertification areas were identified using the machine learning method, and the poverty incidence rate was estimated. Socioeconomic statistical data (for 2012, 2013, and 2015) for the 88 selected counties and municipalities were obtained from China's National Bureau of Statistics. The geographic administrative boundaries in a vector format (ESRI shapefiles) were downloaded from the National Fundamental Geographic Information System website (Wang et al., 2012).

2.2. Night time lights

Currently, the main night time lights data include DMSP/OLS data, NPP/VIIRS data, and Luojia1-01 data. Among them, the DMSP/OLS and NPP/VIIRS data were released by the Earth Observation Group at the Colorado School of Mines. Wuhan University released the Luojia1-01 data in China, and the data were made accessible to the public in 2018. However, their spatial resolution reached 130 m, and they were not suitable for time series analysis (Su et al., 2019; Zhang, Li, Jiang, Shen, & Li, 2018). Therefore, this study selects DMSP/OLS and NPP/VIIRS data to predict poverty.

The Defense Meteorological Satellite Program (DMSP) provides night time lights time series from 1992 to 2013 (30-arc-second spatial resolution). The Visible Infrared Imaging Radiometer Suite (VIIRS) provides monthly night time lights time series from 2013 (15-arc-second geographic grids in GeoTIFF format). Version 4 of the 1992–2013 night time light images acquired by the DMSP/OLS satellites and Version 1 of the Nighttime VIIRS Night Band Composites were downloaded from the Earth Observation Group at the Colorado School of Mines (https://eogdata.mines.edu/products/vnl/) (Elvidge, Baugh, Zhizhin, Hsu, & Ghosh, 2017).

The resolution of night time lights data is different between DMSP/OLS and NPP/VIIRS and requires correction. In this study, the modified invariant region (MIR) method (Wu, He, Peng, Li, & Zhong, 2013) was used for desaturating image data developed by Shi et al. (2016) to reduce the discrepancies. We chose Hegang city, Heilongjiang province, as the correction area because the economic and social development was stable during 1992–2013 (Liu, He, Zhang, Huang, & Yang, 2012). There were eight images of the DMSP/OLS radiation correction light data. Among them, F162006 has the longest observation time and provided sufficient cloud-free images, which is most suitable for reference images (Cao, Wu, Kuang, & Huang, 2015). The F162006 radiometric correction image was used as the reference image to conduct desaturation correction for DMSP/OLS images in China. The calibration model is linear. Equations (1) and (2) are for intra-annual and inter-annual correction, respectively.

$$DN_{(n,i)} = \begin{cases} 0, \ DN^{c}_{(n,i)} = DN^{d}_{(n,i)} = 0 \\ \left(DN^{c}_{(n,i)} + DN^{d}_{(n,i)}\right)\Big/2, \ otherwise \end{cases} \quad (1)$$

where $DN^{c}_{(n,i)}$ and $DN^{d}_{(n,i)}$ represent the pixel values of the night time lights obtained by sensors c and d of year n, respectively. $DN_{(n,i)}$ is the digital number (DN) value of the ith pixel after the fusion of two images in the same year. N is 1994, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, and 2007.

$$DN_{(n,i)} = \begin{cases} 0, \quad DN_{(n+1,i)} = 0 \\ DN_{(n-1,i)}, DN_{(n+1,i)} > 0 \cap DN_{(n-1,i)} > DN_{(n,i)} \\ DN_{(n,i)}, \quad otherwise \end{cases} \quad (2)$$

Similarly, $DN_{(n-1,i)}$ and $DN_{(n+1,i)}$ are the DN values of the image pixels in years N – 1 and N + 1, respectively.

After correction, the total DN value of the images in the study area from 1992 to 2018 showed excellent continuity, useful for long-term sequence analysis (Fig. 2).

As the DMSP/OLS data are limited to 1992–2013, and the NPP/VIIRS data are monthly images from 2012 to the present, this study is based on the inter-annual synthesis of the NPP/VIIRS data. In this study, we selected the 2013 data for two types of mutual data correction (Yu et al., 2015). First, we selected the highest DN value in the most developed cities in China, Beijing, Shanghai, Guangzhou, and Shenzhen as the upper threshold and filtered the data from 2013 to 2018 for outliers. Second, studies showed that the light distribution patterns of DMSP/OLS data and NPP/VIIRS data were highly consistent; thus, the two data sets are suitable for mutual correction (Li, Li, Xu, & Wu, 2017). Based on the 2013 DMSP/OLS data, following the correction, the region with a DN value greater than 0 is selected, 10,000-point elements are randomly generated throughout China, and the DN values of the two types of image data for the same locations are extracted. Logarithmic, linear, exponential, and power functions and polynomial functions are used to fit and analyse the point element values for the same locations. Finally, the model with the largest R^2 value is selected to correct the NPP/VIIRS and DMSP/OLS data. The correction formula is $Y = -0.0033X^2+1.4806X$, where X is the DN value of the NPP/VIIRS data, and Y is the DN value of the corrected image. After correction, the overall continuity of the image significantly improved, with R^2 reaching 0.98, allowing long-term sequence analysis.

2.3. Methodology

Numerous studies have covered the rapid development of data mining technology based on machine learning, especially in the economic and medical fields. Supervised machine learning techniques have been successfully used in these fields for feature or variable reduction to produce highly predictive models (Dipnall et al., 2016; Zhang & Zhou, 2004). Recently, using machine learning data and night time lights technology, researchers have calculated economic and social indicators on a regional scale (Blumenstock, 2016). This study uses machine learning methods to explore the relationship between night time lights data and poverty to enhance the poverty prediction stability. Compared with traditional data, machine learning methods to extract and analyse remote-sensing data have certain advantages. These methods are generally the most effective and robust concerning supervised learning regimes, especially data mining (Li et al., 2019). Based on the 592 poverty-stricken counties determined by *The State Council Leading Group Office of Poverty Alleviation and Development* in 2012, we randomly selected 100 each of poverty-stricken and non-poverty-stricken counties outside the study area as the dataset for classification training. As only Guizhou Province has publicly released poverty incidence data for each county-level unit, we chose the poverty incidence data of each county in Guizhou Province in 2012 and 2013 as the training set for regression learning.

This study uses DMSP/OLS data from 1992 to 2012 and NPP/VIIRS data from 2013 to 2018. Owing to the difference in the spatial resolution between the data sets, there is a considerable difference between the two types of data when estimating poverty. Therefore, we used statistical data from 2012 to 2013 for bridging the two data sets (Fig. 3), training the classification learning, and regression learning models using the methods listed in Table 1. Using the characteristic variables of night lighting and the poverty incidence in the 88 counties in Guizhou Province in 2012 and 2013 as the training data set, the poverty incidence in counties in the study area from 1992 to 2018 was predicted. The results

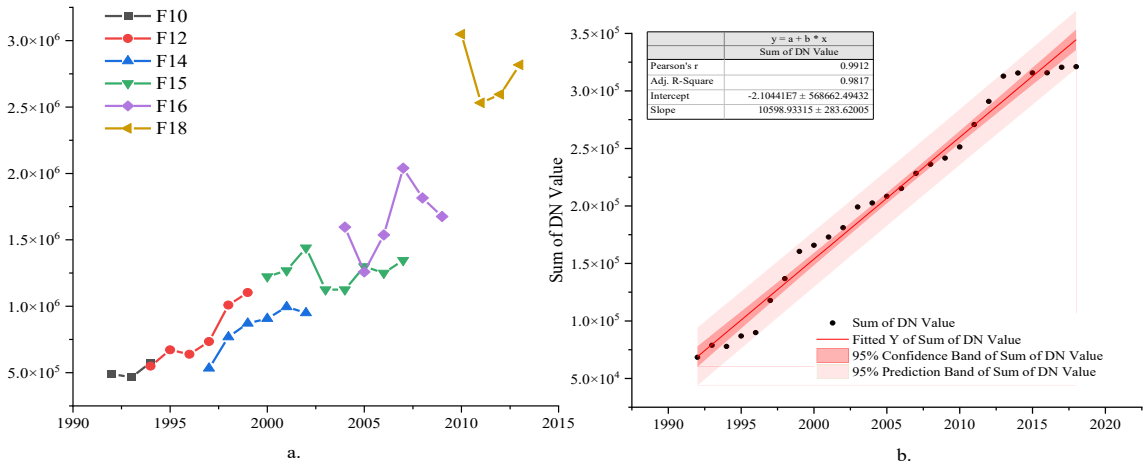

Fig. 2. Total night time lights value of remote-sensing image before(a) and after(b) correction in YGGRD.

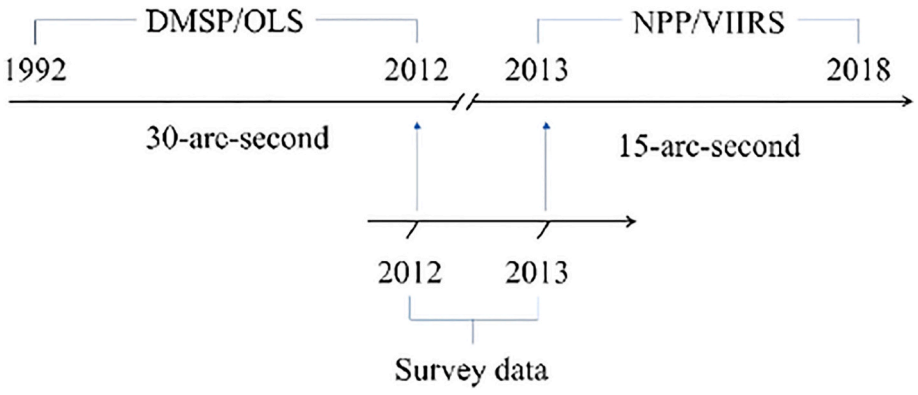

Fig. 3. Survey data as a bridge for night time lights.

Table 1
Accuracy assessment of classification and regression models.

	Classification Models	Accuracy	Regression Models	RMSE	R^2	MAE
1	Decision Trees	72.00%	Linear Regression Models	6.1699	0.38	4.9400
2	Discriminant Analysis	75.00%	**Regression Trees**	**5.0143**	**0.60**	**4.0415**
3	Logistic Regression Classifiers	75.50%	Support Vector Machines	6.2600	0.36	5.0073
4	Naive Bayes Classifiers	71.00%	Gaussian Process Regression Models	6.0588	0.40	4.9100
5	**Support Vector Machines**	**76.50%**	Ensembles of Tress	5.7700	0.46	4.6423
6	Nearest Neighbour Classifiers	68.00%				
7	Ensemble Classifiers	71.5%				

show that the overall trend of poverty incidence predicted after statistical data revision is stable, making the data predicted by the two types of data comparable.

Poverty is predicted using the following: 1) the classification learning method to predict whether the county is poverty-stricken; 2) the regression learning method to predict the poverty incidence in the county (Table 1). We tried training datasets through seven classification models and five regression models and chose two with the highest accuracy. Next, we applied them to the prediction dataset and implemented them in the MATLAB software (R2019a).

The implementation of machine learning determines the indicators of the training dataset. The training dataset is trained to determine the optimum model, and the prediction data set is input into the model to generate the prediction results. This method requires highly representative poverty indicators in the training dataset. Based on a combination of related research (Li et al., 2017, 2019; Zhang, Pandey, & Seto, 2016), five typical indicators are selected as representatives for extracting remote-sensing image information. These indicators include the total value of pixel DN (Sum) corresponding to the night time lights data, the range of pixel DN (Range), the average value of pixel DN (Mean), the standard deviation of pixel DN (Std), and the variability of pixels in the county (variety) areas. The independent variable input to the training dataset is the different feature value of night time lights, and the dependent variable is the poverty incidence obtained through the Guizhou Statistical Yearbook.

The accuracy of the classification model is measured using a receiver operating characteristic (ROC) curve and a confusion matrix (CM) (Li et al., 2019; Machado, Mendoza, & Corbellini, 2015). The accuracy of the regression model is measured using the mean absolute error (MAE), root mean squared error (RMSE), and the R^2 value of the predicted and actual values.

$$MAE = \frac{\sum_{i=1}^{n} |Y_i - X_i|}{n} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - X_i)^2}{n}} \quad (4)$$

In equation (4), Y_i is the predicted value of poverty incidence, X_i is the actual value of poverty incidence, and n is the number of samples. After comparing the machine learning models in Table 1, the training set's average 5-fold cross-validation (CV) accuracy is above 0.65 for all approaches. The model with the highest accuracy is the support vector machine, with an accuracy of 76.50% (Table 1). Therefore, this model can be used to calculate the classification regression for 1992–2018. Among the regression models, the minimum RMSE belongs to the Regression Trees model, at 5.0143, with an MAE of 4.0415 and an R^2 value of 0.60 (Table 1). We used the poverty incidence data of 88 counties in Guizhou Province in 2015 to verify the accuracy of the regression model. According to a scatter plot analysis of the predicted and actual values, the fitted R^2 of the quadratic curve reach 0.95 in 2015. Fig. 4 shows the comparison of the error distribution of the calculation results in 2015. Therefore, the regression model trained by the decision tree model can predict the poverty incidence in 1992–2018.

3. Results

3.1. Results of poverty counties classification

The support vector machine model was used to predict whether the study area counties were poverty counties in 2012 to explore whether our method is feasible and accurate for identifying poverty-stricken counties, with the results shown in Fig. 5.

According to the actual classification of *The State Council Leading Group Office of Poverty Alleviation and Development* in 2012 (Fig. 5), among the 91 counties in the region, 68 were poverty-stricken, and 23 were non-poverty-stricken. From the machine learning results, 73 counties were poverty-stricken, and 18 were non-poverty-stricken. The correct rate for poverty-stricken counties was 82.35%. The overall accuracy was equal to or higher than similar studies (Li et al., 2019; Li, Xu, Chen, & Li, 2013). When using machine learning to predict poverty-stricken counties, the representativeness and balance of the training samples should be the focal point. However, the bias in the prediction results should also be considered to improve the accuracy.

Misclassified counties and districts in the poverty-stricken group are mainly distributed around prefecture-level cities. For example, Longli county is located to the east of Guiyang city (the capital city of Guizhou Province), where the government has focused on developing the port economic zone. Tianyang district is in the eastern part of Baise city, where the tertiary industry has developed rapidly in recent years. The incorrect division of non-poverty-stricken counties is related to the overall spatial structure of these counties. For example, Nandan county in Guangxi has a wide geographical range. However, most of the terrain is mountainous, and areas with lights higher than 0 at night are mainly concentrated in a small part of this county. Therefore, the machine learning model classifies the county as poverty-stricken.

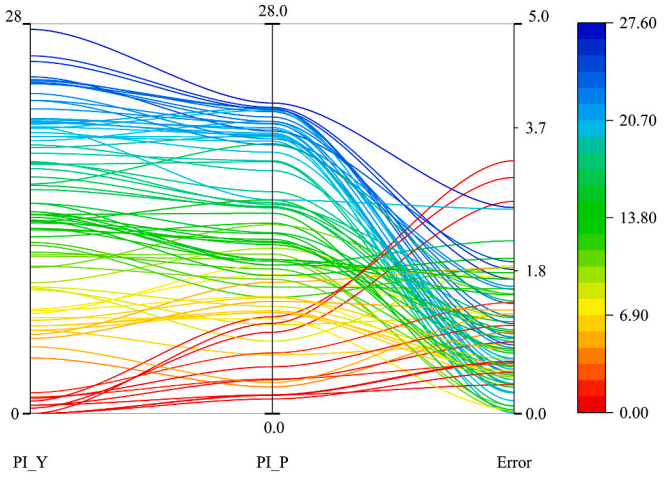

Fig. 4. Error distribution between the PI_Y and PI_P (%) (The lines in Fig. 4 represent the poverty incidence in different counties. PI_Y: Poverty incidence in the statistical yearbook; PI_P: Poverty incidence predicted by the machine learning model; Error: Absolute error of PI_Y and PI_P).

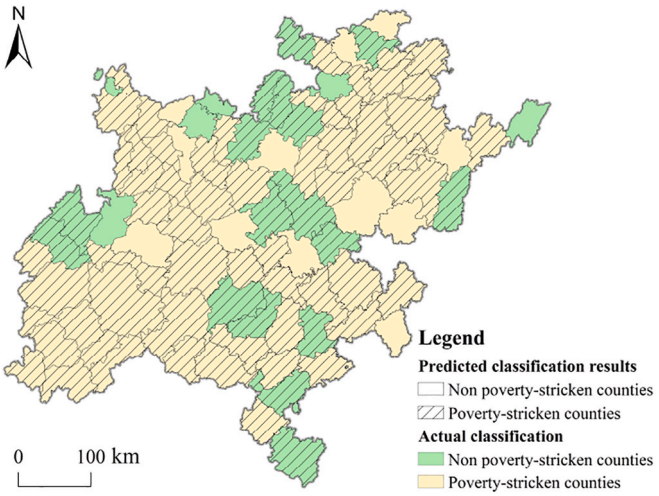

Fig. 5. Classification learning results in 2012.

3.2. Results of poverty incidence regression

The regression tree model is used to predict the poverty incidence in the study area from 1992 to 2018. Fig. 6a shows the average poverty incidence trend (mean PI) in the study area. From the poverty incidence trend, the overall poverty reduction in the region can be divided into three stages: 1992–1996, 1997–2013, and 2014–2018. The pace of poverty reduction in the first (1992–1996) and second (1997–2013) stages is slow and is accelerated in the third stage. This acceleration happened because the Chinese government vigorously promoted a targeted poverty alleviation strategy after 2014. Thus, the overall speed of YGGRD's poverty reduction accelerated, and the regional poverty pattern significantly improved.

We inserted the poverty incidence in 1992, 2013, and 2018 into a histogram to observe the data distribution. Fig. 6b shows that the poverty incidence in the counties of the study area in 1992 was concentrated at around 35%. The centralised distribution range of poverty incidence data from 1992 to 2005 shifted to approximately 25% and 30%, respectively. In the 13 years from 2005 to 2018, the poverty incidence decreased overall, and the concentrated distribution range of the data moved to approximately 8–10%. The distribution of poverty incidence data in 2018 is more scattered, indicating that the implementation of the targeted poverty alleviation policy in the last five years decreased the poverty incidence in various counties in the study area by varying degrees (Fig. 6b).

3.3. Change in the spatial pattern of poverty incidence

We conducted a spatial mapping of poverty incidence in 1992, 2005, and 2018 and conducted a hot spot analysis to explore the spatial pattern of the predicted results (Fig. 7).

Concerning the spatial distribution of poverty incidence in the YGGRD, the overall poverty incidence in 1992 shows a contiguous distribution pattern. In 2005, the patchy area was still prominent, but the poverty incidence decreased. By 2018, there were few poverty-stricken counties with a poverty incidence significantly higher than 30%. From

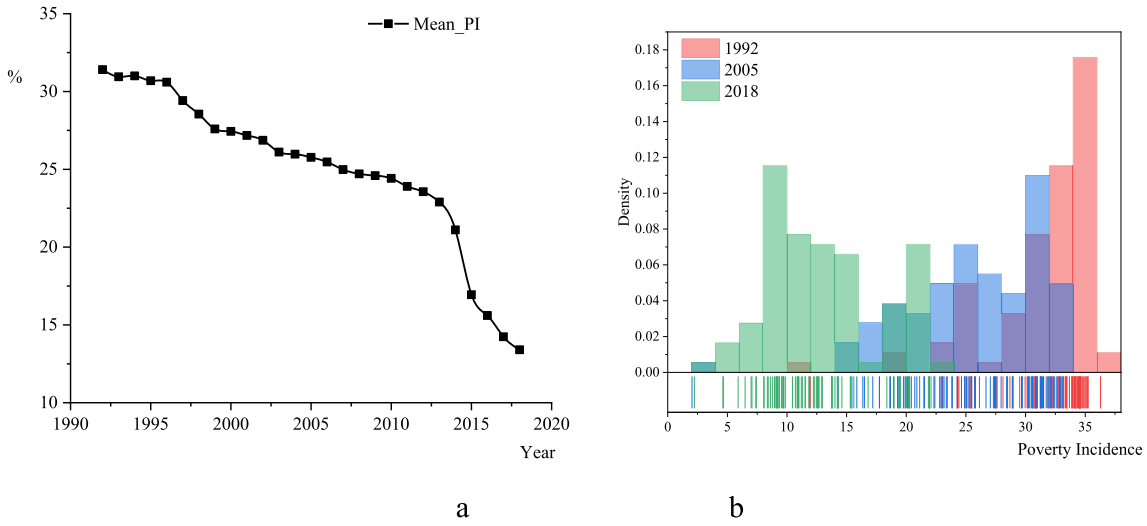

Fig. 6. Distribution of poverty incidence.

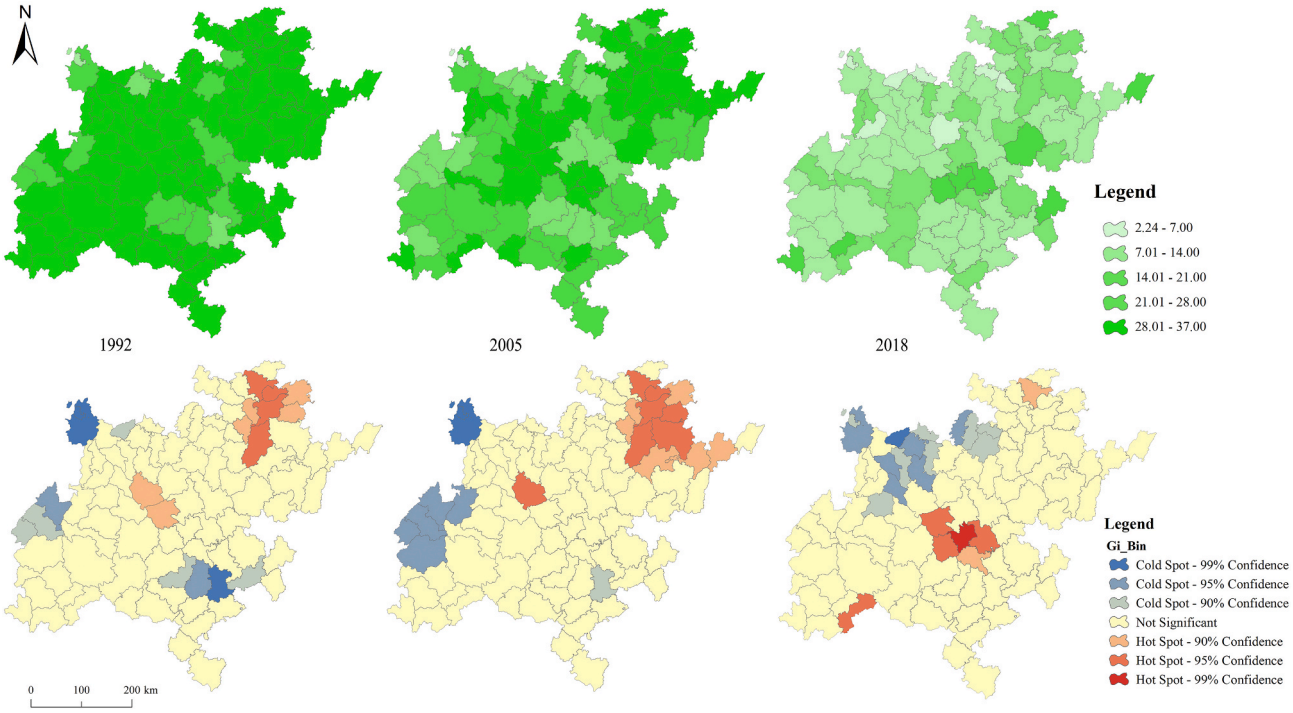

Fig. 7. Analysis of poverty incidence patterns and hot spots in 1992, 2005, and 2018.

1992 to 2018, the spatial evolution of the poverty incidence shows a poverty island effect, consistent with research conducted by scholars, such as Liu, for China (Liu, Liu, & Zhou, 2017). The poverty in their study area transitions from a patchy distribution to point distribution. Deeply poverty-stricken areas are concentrated in individual counties, having absolute spatial dependence. The 'hot spot' in the analysis implies an area that is poorer than its surrounding areas. From the hot spot analysis of poverty incidence, the hot areas of poverty have a certain spatial dependence. Hot spot counties were concentrated in the southeast of Guizhou Province in 1992 and 2005 and the northwest of the Guangxi Autonomous Region in 2018. As this study area is an inter-provincial border area (Cao & Xu, 2018), due to differences in poverty reduction measures in different provinces and the 'targeted poverty alleviation', the 'poorer' areas in different years are unstable in 2005 and 2018.

4. Discussion

Long time-series poverty mapping is data-intensive—extremely challenging in the case of impoverished areas (Erenstein et al., 2010). It is essential to collect as much night time lights data with as long a time series as possible for poverty prediction based on long time series. The resampling method for unifying resolutions has become common because the spatial resolutions of DMSP/OLS and NPP/VIIRS are different (Dong, Li, & Li, 2020). However, the resampling method reduces the accuracy of NPP/VIIRS data and obliterates more data details, unfavourable for predicting the poverty incidence. We found that NPP/VIIRS data can retain more dark details compared with DMSP/OLS data without the influence of image noise; this is extremely useful in estimating poverty (Shi, Yu, et al., 2014; Yu et al., 2015).

The regional differences in poverty reflected in the error distribution

of the poverty index regression model require more attention. The distribution of the error data is interesting because the counties with high errors are those whose poverty incidence is close to 0. This shows that high light intensity values have a more significant impact on the stability of the model for predicting poverty. Conversely, the errors in poverty-stricken counties are smaller than those in non-poverty-stricken counties. Therefore, when using night time lights to predict poverty, the dark details in the images of the night time lights are most important, rarely mentioned in existing studies (Jean et al., 2016b; Li et al., 2019; Yu et al., 2015). Notably, although the poverty incidence data of the non-poverty counties, according to the statistical survey, are close to 0, the poverty incidence of these counties, as judged by machine learning and night time lights, shows significant deviation. This deviation is most likely due to the existence of 'hidden poverty' in non-poverty counties because poor people are not absent in areas judged by the government as non-poverty counties. Consistent with the literature, because the locals have different lighting habits, some errors may occur. However, regarding the overall trend of poor areas, the use of lights for poverty prediction is stable (Elvidge et al., 2009).

It is imperative to select an appropriate training set sample for machine learning. Guizhou Province is one of the poorest mountainous provinces in China (Xu et al., 2020). The 44 counties in Southern Guizhou belong to the YGGRD. Therefore, using poverty-stricken county samples and poverty incidence samples in Guizhou Province as training data sets for predicting poverty in the neighbouring areas has high stability. However, the trained model (the regression trees model) is only suitable for use in China (and countries with the same level of welfare) because the training dataset based on statistical survey data reflects the poverty level under China's unified social welfare level. We need to adopt corresponding levels of welfare to assess poverty in other parts of the world. China's most recent poverty line, set in 2010, refers to the per capita net income of 2300 yuan in rural households, based on the constant price in 2010. The 2010 poverty line is a necessary and stable food and clothing standard formulated in conjunction with 'two no worries and three guarantees'. This means that the basic requirements and key indicators for poverty alleviation are that rural poor people are free from worries about food and clothing and have access to compulsory education, basic medical services, and safe housing by 2020 (Xian et al., 2016).

The unified poverty line makes the poverty incidence predicted by the model comparable in the time series. Judging from the results of the poverty incidence predicted in 2012 and 2013, the overall difference in the data distribution is small. The data distribution has two high-density areas (27.5–30% and 20–25%), small differences in changes, and smooth average transition. Therefore, using statistical survey data to bridge night time lights of different resolutions can avoid the 'roughening' of high-resolution night time lights data and more accurately reflect the characteristics of impoverished areas.

5. Conclusion

The rapid development of night lighting technology provides an opportunity to obtain long-term poverty data. Although different sensors have different resolutions for night light data, long-term poverty prediction can be achieved through methods such as statistical yearbook data combined with machine learning.

To solve the problem of not obtaining poverty time series data, we used machine learning methods and night time lights data to predict the poverty incidence in YGGRD at the county level. The results showed that the prediction of the model was reliable, and the accuracy was acceptable. The long-term predicted poverty data at the county level can be used to understand the temporal changes and spatial patterns of regional poverty and formulate relevant policies better. However, the predicted poverty incidence data are more suitable for macro decision-making and cannot replace actual data. Moreover, as the poverty lines of various countries and regions are different when applying this method in other countries or regions, it is necessary to refer to the local (or World Bank) poverty line to adjust the predictive variables to obtain a more realistic forecast result. Further research can refer to the globally representative poverty line established by the World Bank. It can use machine learning methods to train models suitable for global poverty incidence prediction and predict poverty data in the long term at the county level in the poverty-data gap area.

Author statement

Jianbin Xu: Writing- Reviewing and Editing, Conceptualization, Visualization, Methodology, Writing, Software.

Jie Song: Data curation, Original draft preparation.

Baochao Li: Investigation.

Dan Liu: Software, Validation.

Xiaoshu Cao: Supervision., Funding acquisition.

Acknowledgments

The authors highly appreciate the constructive comments and suggestions provided by the anonymous reviewers to improve the quality of this manuscript. This work was supported by the National Natural Science Foundation of China [No. 41831284].

References

Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics, 95*(7–8), 476–487. https://doi.org/10.1016/j.jpubeco.2010.11.006

Anderson, G., Farcomeni, A., Pittau, M. G., & Zelli, R. (2016). A new approach to measuring and studying the characteristics of class membership: Examining poverty, inequality and polarization in urban China. *Journal of Econometrics, 191*(2), 348–359. https://doi.org/10.1016/j.jeconom.2015.12.006

Bigman, D., & Fofack, H. (2000). Geographical targeting for poverty alleviation. In *Geographical targeting for poverty alleviation*. https://doi.org/10.1596/0-8213-4625-3

Blumenstock, J. E. (2016). Fighting poverty with data. *Science, 353*(6301), 753–754. https://doi.org/10.1126/science.aah5217

Cao, Z., Wu, Z., Kuang, Y., & Huang, N. (2015). Correction of DMSP/OLS night-time light images and its application in China. *Journal of Geo-Information Science, 17*(9), 1092–1102. https://doi.org/10.1007/11427469_160

Cao, X., & Xu, J. (2018). Spatial heterogeneity analysis of regional economic development and driving factors in China's provincial border counties. *Acta Geographica Sinica, 73*(6), 1065–1075. https://doi.org/10.11821/dlxb201806006

Chen, X., & Nordhaus, W. (2010). The value of luminosity data as a proxy for economic statistics. https://doi.org/10.3386/w16317.

Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N., et al. (2016). Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. *PloS One, 11*(2), Article e0148195. https://doi.org/10.1371/journal.pone.0148195

Doll, C. N. H., Muller, J.-P., & Morley, J. G. (2006). Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics, 57*(1), 75–92. https://doi.org/10.1016/j.ecolecon.2005.03.007

Dong, H., Li, R., & Li, J. (2020). Study on urban spatiotemporal expansion pattern of three first-class urban agglomerations in China derived from integrated DMSP-OLS and NPP-VIIRS nighttime light data. *Journal of Geo-Information Science, 22*(5), 1161–1174. https://doi.org/10.12082/dqxxkx.2020.190711

Ebener, S., Murray, C., Tandon, A., & Elvidge, C. C. (2005). From wealth to health: Modelling the distribution of income per capita at the sub-national level using night-time light imagery. *International Journal of Health Geographics, 4*(1), 5. https://doi.org/10.1186/1476-072X-4-5

Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C., & Ghosh, T. (2017). VIIRS night-time lights. *International Journal of Remote Sensing, 38*(21), 5860–5879. https://doi.org/10.1080/01431161.2017.1342050

Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B., et al. (2009). A global poverty map derived from satellite data. *Computers & Geosciences, 35*(8), 1652–1660. https://doi.org/10.1016/j.cageo.2009.01.009

Erenstein, O., Hellin, J., & Chandna, P. (2010). Poverty mapping based on livelihood assets: A meso-level application in the indo-gangetic plains, India. *Applied Geography, 30*(1), 112–125. https://doi.org/10.1016/j.apgeog.2009.05.001

Forbes, D. J. (2013). Multi-scale analysis of the relationship between economic statistics and DMSP-OLS night light images. *GIScience and Remote Sensing, 50*(5), 483–499. https://doi.org/10.1080/15481603.2013.823732

Foster, J. L. (1983). Observations of the Earth using nighttime visible imagery. *Proceedings of SPIE - The International Society for Optical Engineering, 414*, 187–193. https://doi.org/10.1117/12.935885

Guo, Y., Zhou, Y., & Liu, Y. (2019). Targeted poverty alleviation and its practices in rural China: A case study of fuping county, hebei province. *Journal of Rural Studies*. https://doi.org/10.1016/j.jrurstud.2019.01.007

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *The American Economic Review, 102*(2), 994–1028. https://doi.org/10.1257/aer.102.2.994

Isidro, M., Haslett, S., & Jones, G. (2016). Extended Structure Preserving Estimation (ESPREE) for updating small area estimates of poverty. *Annals of Applied Statistics, 10*(1), 451–476. https://doi.org/10.1214/15-AOAS900

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016a). Combining satellite imagery and machine learning to predict poverty. *Science, 353*(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D., & Ermon, S. (2016b). Machine learning to predict poverty. *Science, 353*(6301), 790–794.

Kim, M. C. (1997). Theory of satellite ground-track crossovers. *Journal of Geodesy, 71*(12), 749–767. https://doi.org/10.1007/s001900050141

Liao, C., Fei, D., Huang, Q., Jiang, L., & Shi, P. (2021). Targeted poverty alleviation through photovoltaic-based intervention: Rhetoric and reality in Qinghai, China. *World Development, 137*, 105117. https://doi.org/10.1016/j.worlddev.2020.105117

Li, G., Cai, Z., Liu, X., Liu, J., & Su, S. (2019). A comparison of machine learning approaches for identifying high-poverty counties: Robust features of DMSP/OLS night-time light imagery. *International Journal of Remote Sensing, 40*(15), 5716–5736. https://doi.org/10.1080/01431161.2019.1580820

Li, X., Li, D., Xu, H., & Wu, C. (2017). Intercalibration between DMSP/OLS and VIIRS night-time light images to evaluate city light dynamics of Syria's major human settlement during Syrian Civil War. *International Journal of Remote Sensing, 38*(21), 5934–5951. https://doi.org/10.1080/01431161.2017.1331476

Liu, Y., Guo, Y., & Zhou, Y. (2018). Poverty alleviation in rural China: Policy changes, future challenges and policy implications. *China Agricultural Economic Review, 10*(2), 241–259. https://doi.org/10.1108/CAER-10-2017-0192

Liu, Z., He, C., Zhang, Q., Huang, Q., & Yang, Y. (2012). Extracting the dynamics of urban expansion in China using DMSP-OLS nighttime light data from 1992 to 2008. *Landscape and Urban Planning, 106*(1), 62–72. https://doi.org/10.1016/j.landurbplan.2012.02.013

Liu, Y., Liu, J., & Zhou, Y. (2017). Spatio-temporal patterns of rural poverty in China and targeted poverty alleviation strategies. *Journal of Rural Studies, 52*, 66–75. https://doi.org/10.1016/j.jrurstud.2017.04.002

Li, X., Xu, H., Chen, X., & Li, C. (2013). Potential of NPP-VIIRS nighttime light imagery for modeling the regional economy of China. *Remote Sensing, 5*(6), 3057–3081. https://doi.org/10.3390/rs5063057

Lo, K., Xue, L., & Wang, M. (2016). Spatial restructuring through poverty alleviation resettlement in rural China. *Journal of Rural Studies, 47*, 496–505. https://doi.org/10.1016/j.jrurstud.2016.06.006

Machado, G., Mendoza, M. R., & Corbellini, L. G. (2015). What variables are important in predicting bovine viral diarrhea virus? A random forest approach. *Veterinary Research, 46*(1), 85. https://doi.org/10.1186/s13567-015-0219-7

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science, 341*(6149), 976–980. https://doi.org/10.1126/science.1238041

Mellander, C., Lobo, J., Stolarick, K., & Matheson, Z. (2015). Night-time light data: A good proxy measure for economic activity? *PloS One, 10*(10), 1–18. https://doi.org/10.1371/journal.pone.0139779

Noor, A. M., Alegana, V. A., Gething, P. W., Tatem, A. J., & Snow, R. W. (2008). Using remotely sensed night-time light as a proxy for poverty in Africa. *Population Health Metrics, 6*(1), 5. https://doi.org/10.1186/1478-7954-6-5

Pan, J., Zhao, H., & Dong, L. (2018). Spatial identification of multi-dimensional poverty in rural China by using nighttime light and sustainable livelihoods. *Acta Ecologica Sinica, 38*(17), 6180–6193. https://doi.org/10.5846/stxb201709101627

Propastin, P., & Kappas, M. (2012). Assessing satellite-observed nighttime lights for monitoring socioeconomic parameters in the republic of Kazakhstan. *GIScience and Remote Sensing, 49*(4), 538–557. https://doi.org/10.2747/1548-1603.49.4.538

Sadath, A. C., & Acharya, R. H. (2017). Assessing the extent and intensity of energy poverty using Multidimensional Energy Poverty Index: Empirical evidence from households in India. *Energy Policy, 102*, 540–550. https://doi.org/10.1016/j.enpol.2016.12.056

Shi, K., Chen, Y., Yu, B., Xu, T., Yang, C., Li, L., et al. (2016). Detecting spatiotemporal dynamics of global electric power consumption using DMSP-OLS nighttime stable light data. *Applied Energy, 184*, 450–463. https://doi.org/10.1016/j.apenergy.2016.10.032

Shi, K., Huang, C., Yu, B., Yin, B., Huang, Y., & Wu, J. (2014). Evaluation of NPP-VIIRS night-time light composite data for extracting built-up urban areas. *Remote Sensing Letters, 5*(4), 358–366. https://doi.org/10.1080/2150704X.2014.905728

Shi, K., Yu, B., Huang, Y., Hu, Y., Yin, B., Chen, Z., et al. (2014). Evaluating the ability of NPP-VIIRS nighttime light data to estimate the gross domestic product and the electric power consumption of China at multiple scales: A comparison with DMSP-OLS data. *Remote Sensing, 6*(2), 1705–1724. https://doi.org/10.3390/rs6021705

Subash, S. P., Kumar, R. R., & Aditya, K. S. (2018). Satellite data and machine learning tools for predicting poverty in rural India. *Agricultural Economics Research Review, 31*(2), 231. https://doi.org/10.5958/0974-0279.2018.00040.X

Su, Z., Zhong, X., Zhang, G., Li, Y., He, X., Wang, Q., et al. (2019). High sensitive night-time light imaging camera design and in-orbit test of luojia1-01 satellite. *Sensors*. https://doi.org/10.3390/s19040797

Tobler, W. R. (1969). *Satellite confirmation of settlement size coefficients*. Area. Retrieved from https://www.jstor.org/stable/20000359.

Wang, W., Cheng, H., & Zhang, L. (2012). Poverty assessment using DMSP/OLS night-time light satellite imagery at a provincial scale in China. *Advances in Space Research, 49*(8), 1253–1264. https://doi.org/10.1016/j.asr.2012.01.025

Welch, R. (1980). Monitoring urban population and energy utilization patterns from satellite Data. *Remote Sensing of Environment, 9*(1), 1–9. https://doi.org/10.1016/0034-4257(80)90043-7

Wu, J., He, S., Peng, J., Li, W., & Zhong, X. (2013). Intercalibration of DMSP-OLS night-time light data by the invariant region method. *International Journal of Remote Sensing, 34*(20), 7356–7368. https://doi.org/10.1080/01431161.2013.820365

Xian, Z., Wang, P., & Wu, W. (2016). Rural poverty lines and poverty monitoring in China. *Statistics Research, 33*(9), 3. https://doi.org/10.19343/j.cnki.11-1302/c.2016.09.001

Xu, J., Song, J., Li, B., Liu, D., Wei, D., & Cao, X. (2020). Do settlements isolation and land use changes affect poverty? Evidence from a mountainous province of China. *Journal of Rural Studies, 76*, 163–172. https://doi.org/10.1016/j.jrurstud.2020.04.018

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., et al. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications, 11*(1), 2583. https://doi.org/10.1038/s41467-020-16185-w

Yu, B., Shi, K., Hu, Y., Huang, C., Chen, Z., & Wu, J. (2015). Poverty evaluation using NPP-VIIRS nighttime light composite data at the county level in China. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8*(3), 1217–1229. https://doi.org/10.1109/JSTARS.2015.2399416

Zhang, G., Li, L., Jiang, Y., Shen, X., & Li, D. (2018). On-Orbit relative radiometric calibration of the night-time sensor of the LuoJia1-01 satellite. *Sensors, 18*(12), 4225. https://doi.org/10.3390/s18124225

Zhang, Q., Pandey, B., & Seto, K. C. (2016). A robust method to generate a consistent time series from DMSP/OLS nighttime light data. *IEEE Transactions on Geoscience and Remote Sensing, 54*(10), 5821–5831. https://doi.org/10.1109/TGRS.2016.2572724

Zhang, D., & Zhou, L. (2004). Discovering golden nuggets: Data mining in financial application. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 34*(4), 513–522. https://doi.org/10.1109/TSMCC.2004.829279

Zhou, Y., & Liu, Y. (2019). The geography of poverty: Review and research prospects. *Journal of Rural Studies*. https://doi.org/10.1016/j.jrurstud.2019.01.008