Engenharia Agrícola

ISSN: 1809-4430 (on-line)

www.engenhariaagricola.org.br

Doi: http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v41n6p619-633/2021

APPLICATION OF RANDOM FOREST IN IDENTIFYING WINTER WHEAT USING LANDSAT8 IMAGERY

Xu Li[1], Xifeng Lv[2], Yufeng He[1*], Baoping Zhou[1], Jinmei Deng[1], Anzhen Qin[3]

[1*]Corresponding author. E-mail: heyufeng@hnu.edu.cn | ORCID ID: https://orcid.org/0000-0002-3325-7756

KEYWORDS

remote sensing image, crop classification; planting area extraction; winter wheat identification

ABSTRACT

Mastering accurate spatial planting and distribution status of the crops is significantly important for the nation to guide the agricultural production and formulate agricultural policies from a macro perspective. In this paper, the Landsat-8 OLI satellite images were taken as the data sources. And as for the nine crop types within the study area, such as the wheat, rice, and other crops, three classification methods of the random forest classification (RFC), the support vector machine (SVM), and the maximum likelihood classification (MLC) were applied in extracting the planting area of winter wheat in Wushi County of Xinjiang Uygur Autonomous Region. It can be seen from the results that, general classification accuracy of MLC, SVM, and RFC are respectively 80.58%, 87.95%, and 95.96%, while their Kappa coefficients are respectively 0.61, 0.76, and 0.86. The RFC method shows higher classification accuracy that those of MLC and SVM methods. The principal component analysis (PCA) was carried out on the original 7-band image to extract the first 4 principal components and calculate the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), wide dynamic range vegetation index (WDRVI), and normalized difference water index (NDWI). Meanwhile, the 6 additional auxiliary feature bands were superimposed on the original 7-band images to carry out reclassification, through which, the general accuracy of MLC increased by 3 percent while its Kappa coefficient increased by 0.06; the SVM general accuracy increased by 3.02 percent while its Kappa coefficient increased by 0.13; and the general accuracy of the RFC increased by 0.85 percent while its Kappa coefficient increased by 0.02. This indicates that, the adding of auxiliary information can improve the crop classification and identification ability and accuracy. Based on the comprehensive evaluation, the classification method of random forest is proved to have better performance in winter wheat identification.

INTRODUCTION

The method to identify crops and extract their respective areas accurately by remote sensing images is featured in real-time performance, reliability, and low cost. Besides, it can also be used to carry out spatial mapping of crop distribution Singla et al. (2019). The crop identification and classification technology is vital for monitoring the crop areas by agricultural condition remote sensing. Since the crops are in many varieties, which are all plants with no significant spectral differences but more serious phenomenon that “different objects have the same spectrum and the same objects have different spectrum”, the general requirements on crops classification are comparatively higher and stricter. Traditional classification methods, including supervised classification Khamparia et al. (2020), Murmu & Biswas (2015), Guermazi et al. (2016), unsupervised classification Gašparović et al. (2020), Cong et al. (2018), object-oriented classification Cong et al. (2019), Zhou et al. (2015), and decision tree classification Parida & Ranjan (2019), Muhammad et al.

[1] College of Information Engineering, Tarim University/ Aral City, China.
[2] College of Life Sciences, Tarim University/ Aral City, China.
[3] Institute of Farmland Irrigation, Chinese Academy of Agricultural Science / Xinxiang, China.

Area Editor: Teresa Cristina Tarlé Pissarra
Received in: 6-17-2020
Accepted in: 11-1-2021

(2016), have their respective advantages and disadvantages. Nowadays, as for the agricultural conditions, the main commercialized methods to extract crop areas by remote sensing monitoring include: maximum likelihood classification, support vector machine, and decision tree classification, among which, the decision tree classification method is featured in advantages of rapid classification and capable applicability, and has been widely used in crop area extraction. The main approaches included in decision tree classification methods are: expert knowledge decision tree Xu et al. (2018), ID3 algorithm Wang et al. (2019), C4.5 algorithm Li-ping & Yu-jun (2018), classification and regression tree(CART) algorithm Berhane et al. (2018), and random forest classification (RFC) algorithm Wu et al. (2019), etc.

Kandrika & Roy (2008) used the multi-time-phase IRS-P6 satellite AWi FS (advanced wide field sensor) data to perform the land use and land cover classifications in Orissa region based on See-5 decision tree method, through which, they obtained comparatively high Kappa coefficient. Peña et al. (2014) compared and analyzed the effects of multiple machine learning classification methods, such as C4.5 method and support vector machine (SVM) method, in classifying and identifying summer crops in their study area on the precondition of performing object-oriented segmentation to images. The results showed that the general accuracy of SVM is higher than that of the C4.5 method.

The random forest is a new and efficient combined method for decision tree classification, which has a series of advantages over traditional construction methods of decision tree, including rapid training speed, simple implementation, high accuracy, easy parallelization, and capable anti-noise performance. It has been widely applied in various fields in overseas regions. Jamali (2019) used Landsat images and random forest method to classify the land cover, and made comparison with iterative algorithm, integrated learning method, and support vector machine method. It can be seen from the results that the random forest method is better than all other methods in both efficiency and accuracy; Gu et al. (2019) took use of a series of auxiliary data like multi-spectral data, DEM (digital elevation model), slope, and aspect in their study, made comparison on classification between random forest and CART decision tree, which finally proved that the random forest method has better classification accuracy than CART algorithm; Zhan et al. (2018) used the random forest method and the maximum likelihood method to identify and classify the crops, through which, the accuracy of the random forest method was proved to reach 85.89%. This was 8% higher than that of the maximum likelihood classification method; Deschamps et al. (2012) conducted crops identifications in Eastern and Western regions of Canada based on Radar data. The results indicated that the crop classification accuracy of the random forest method was 7% higher than that of the traditional decision tree.

It can be seen from the above that, as for image classification, the random forest method is much more capable in both accuracy and efficiency. It is necessary to study its potential of being used in fine identification and classification of crops by agricultural remote sensing. As for the adjustment of crop production structure in China, optimizing the planting structure of winter wheat, reducing planting area of winter wheat in non-competence regions, and encouraging farmers to plant other crops with better comprehensive benefits have been the key points during the recent years. To the distribution of grain policy subsidy, it is quite significant to conduct the study on using proper remote sensing crop classification and identification methods to make accurate statistics about winter wheat planting area within relative region. In this paper, the winter wheat, which is the main crop in Wushi County of Xinjiang Uygur Autonomous Region, was taken as the classification object, and the single Landsat-8 OLI image data were used as the classification data sources, through which, we selected a proper number of sample data, and made comparison on classification accuracy and Kappa coefficients among maximum likelihood classification, support vector machine classification, and random forest classification; Meanwhile, in order to evaluate the impacts of auxiliary information on the classification accuracy of different classification methods, four types of computing, including: NDVI, EVI, WDRVI, and NDWI were carried out on the original images, and 4 types of index images were selected and added to the original image as the additional feature bands, and then, methods of the maximum likelihood classification, support vector machine classification, and random forest were adopted respectively to carry out classification and make comparison on the classification accuracy before and after adding the auxiliary feature bands, thus to provide scientific and rational experimental support and theoretical basis for selecting the method to extract and classify agricultural crop areas, thereby showing the feasibility of using single time phase OLI images to identify winter wheat in Xinjiang Uygur Autonomous Region by random forest method.

MATERIAL AND METHODS

Overview of Study Area

Wushi County is situated in the southwest part of Xinjiang Uygur Autonomous Region, the western part of Aksu region, the northwestern edge of the Tarim Basin, the southern foot of the Tianshan Mountains, and the upper reaches of the Taushgan Darya. Its geographical coordinates are 78°23′41″ - 80°01′09″ E, 40°43′08″-41°51′12″N, covering an area of 9082 ㎡ in total. Wushi County is in a warm continental semiarid climate zone with a mean annual temperature of 9.4℃, an extremely high temperature of 35.5℃, and an extremely low temperature of -26.6℃. The annual time of sunshine reaches 2750～2850 hours. And in the river valley plain area, the annual precipitation is 70～120 mm, the annual frost-free period lasts for 183-206 days, and the ≥10℃ active accumulated temperature reaches 3200℃-3600℃. Meanwhile, the overall landform here is high in the northwest and low in the southeast, surrounded by mountains, with valleys in the middle. 59.9% area here is mountainous region, 27.6% is Gobi Desert, leaving only 12.5% area of valley plains. It is commonly known to have: a majority of mountainous region, large part of desert, and only a few applicable for crop planting. The average elevation here is 1,396 meters. And the county, which is located at 1,400 meters above sea level, relies on agriculture development. Its farming area all through the year is maintained at around 460,000 mu, planting wheat, corn, cotton, and rice as the main crops.

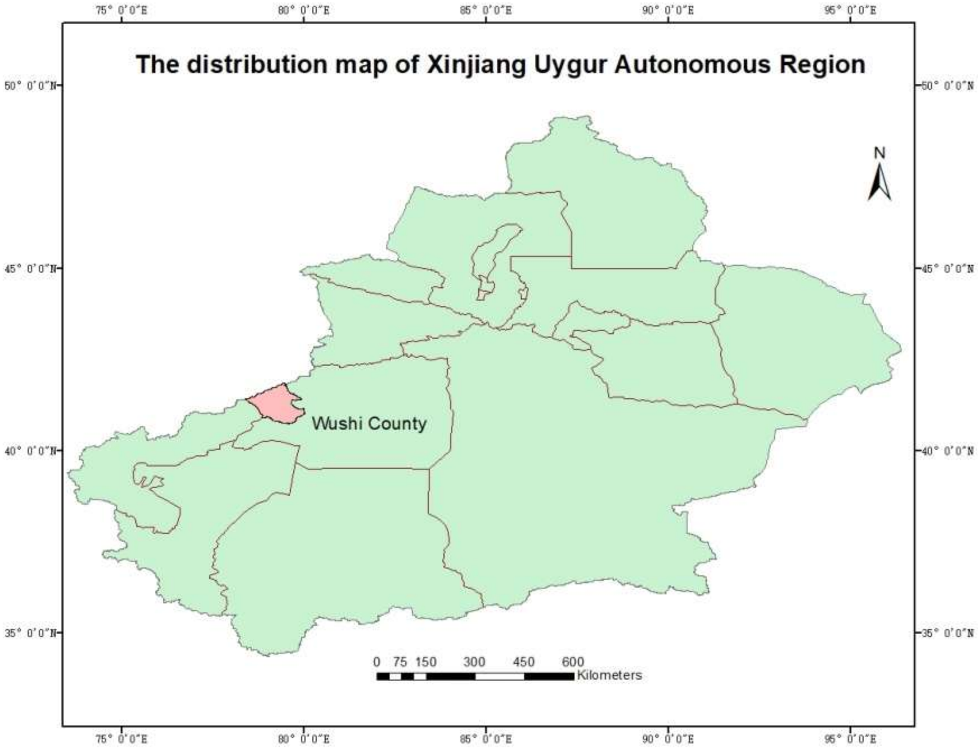


FIGURE 1. The study area.

Acquisition and Processing of Test Data

Processing of Remote Sensing Data

The data of Landsat 8 satellite, which was launched by NASA on February 11, 2013, were mainly used in this study. OLI is the main sensor that the satellite carries and includes 7 bands: coast/aerosol (430~450nm), blue (450~510) nm), green (530~590 nm), red (640~670nm), near infrared (850~880nm), shortwave infrared 1 (1570~1650nm) and shortwave infrared 2 (2110~2290nm). For all bands mentioned above, their spatial resolutions are all 30m. According to the development period characteristics of the main crop --- winter wheat in the study area, the Landsat 8 OLI satellite images covering the entire study area on May 15, 2017, were selected in this paper.

The ENVI 5.0 software was applied to conduct radiometric calibration, atmospheric correction, and geometric correction on the acquired remote sensing images. The radiometric calibration formula is as follow:

$$L_z(\lambda_z) = Gain \times DN + Bias$$

Where,

$L_z(\lambda_z)$ refers to the spectral radiance at the entrance pupil of the sensor;

Gain refers to the calibration slope;

DN refers to the image gray value, and

Bias refers to the calibration intercept. Both Gain and Bias were provided by the satellite data supplier, which could be read from the metadata file of the original Landsat image. The radiation calibration coefficients of each band of Landsat 8 OLI are as shown in Table 1.

TABLE 1. Radiometric calibration coefficient of Landsat 8OLI image.

Band	Gain	Bias
Coastal aerosol	0.01298	-64.89967
Blue	0.013292	-66.45805
Green	0.012248	-61.24053
Red	0.010328	-51.64146
Near infrared	0.0063204	-31.602
SWIR 1	0.0015718	-7.85913
SWIR 2	0.00052979	-2.64895

The module of ENVI/FLAASH atmospheric correction was applied for conducting atmospheric correction, while the ENVI/OLI correction module was taken for conducting geometric correction.

Ground Investigation

To obtain the layout of the main feature types in Wushi County, Xinjiang Uygur Autonomous Region, a ground investigation was conducted in Wushi County, during which, the GPS (global positioning system) handhelds were used to measure the latitude and longitude coordinates of the feature plots and record the vegetation types and photographs. The route of the ground investigation covered most of the group farms in Wushi County, and a small number of sampling points in farms outside the group.

It can be known from the main crops planting status and the ground investigation data that, the study area covers an area of 44040 hectares, including 2370 hectares' rice area, 16330 hectares for wheat planting, and 300 hectares for beans planting. Besides, the ground investigation also showed that vegetables and fruits are also planted in the study area in scattered manner. So, there are 9 classified types within the area, including: wheat, rice, forest land, beans, ice and snow, other vegetation, bare land, and water bodies. As for the original Landsat image and the sample plots layout, please refer to Figure 2.

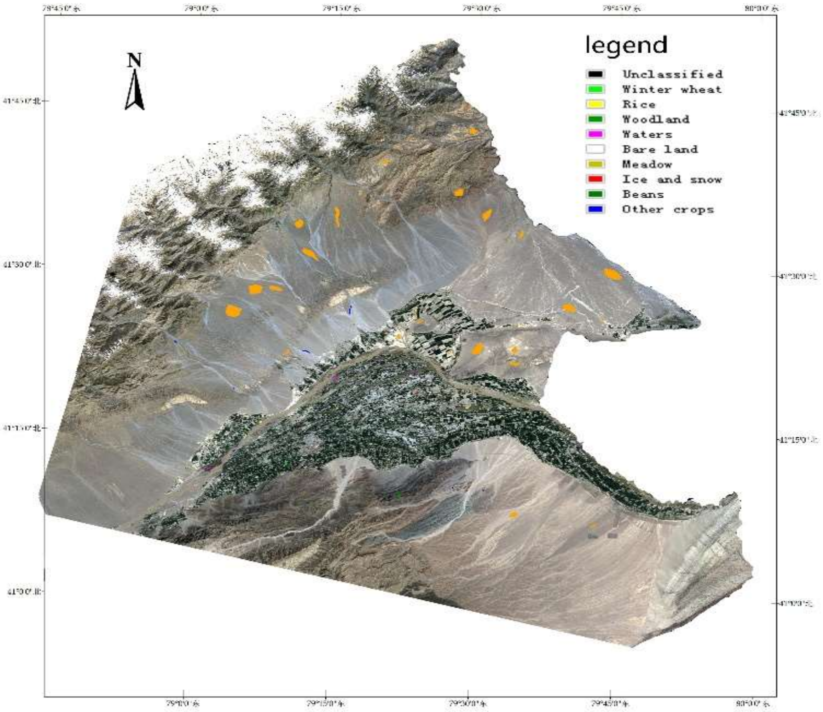


FIGURE 2. Landsat 8 OLI image and distribution of ground sample in study area.

Development Stages of Crops

The lifetime of winter wheat, which starts from being a seed and ends up with reproducing new seed, subjects to the influences from ecological and cultivation conditions greatly. It experiences 4 seasons from the end of autumn, winter, spring, to early summer, amounting to about 230d. Generally, the winter wheat should be planted in late October of the year and harvested in late May of the next year. The entire lifetime of the wheat is divided into 12 stages, which are respectively: seedling stage, three-leaf stage, tillering stage, overwintering stage, re-greening stage, starting stage, jointing stage, booting stage, heading stage, flowering stage, filling stage, and maturity stage.

Considering the development period of crops in the study area, a satellite image located on May 15, 2017, was selected for crop classification and identification. The spectral curves of the main surface feature types (wheat, rice, and forest land, etc.) in the study area during this period are as shown in Figure 3. In the vigorous development period of the plants, namely the 5th bands approaching the near infrared band, the spectral curves of various plants vary significantly and are more favorable for being used in crop classification. In March, the winter wheat is at the re-greening and jointing stage, so that the vegetation index at this time begins to rise; from May to June, winter wheat gradually reaches maturity and grows vigorously, showing strong intra-crop spectrum consistency, but significant inter-crops spectrum differences. So, taking acts in this period could be favorable for identifying crops based on remote sensing because it avoids the spectral differences in the early crop growth period caused by different planting times, and avoids the differences in the late stage caused by different maturity and harvesting times. Therefore, the OLI data on May 15 was taken in this paper to identify various crops like wheat in the study area.

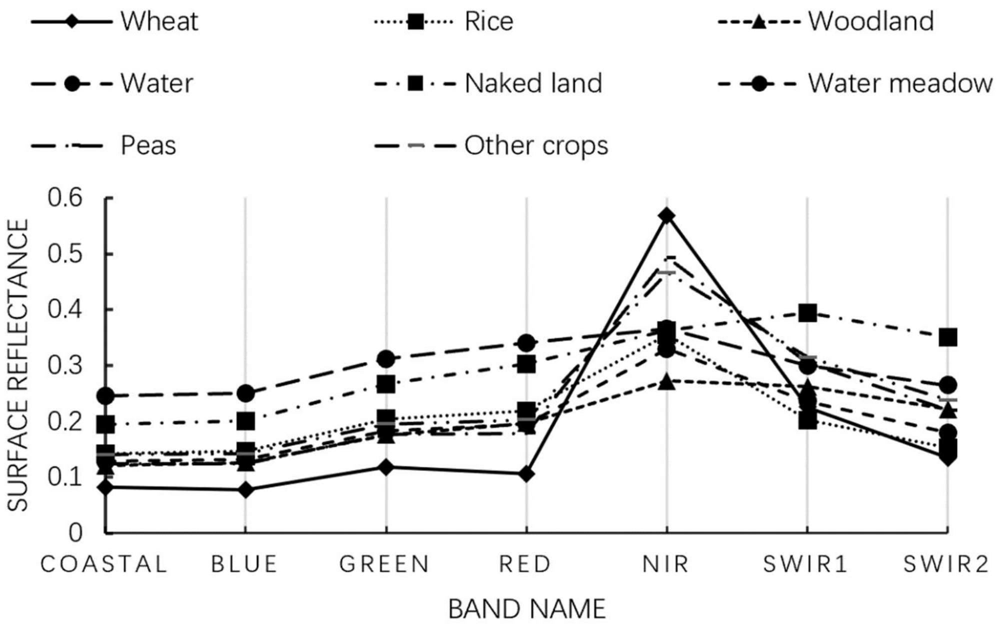


FIGURE 3. Spectral curves of main ground objects in study area.

Description on Research Methods and Algorithms

Technical Thoughts

The overall research thoughts are as shown in Figure 4. According to the crop distribution in the study area, Landsat-8/OLI satellite images on May 15, 2017, were selected. Meanwhile, the appropriate sample data were selected, and 3 classification methods, including: the maximum likelihood classification, support vector machine classification, and random forest classification were used to identify and classify the main crop of the study area --- winter wheat, thus, to evaluate the classification accuracy and applicability of these methods. Based on the local crop planting structure and the existing remote sensing images of Wushi County, the NDVI that can detect vegetation greenness and vegetation coverage, the enhanced vegetation index (EVI) that can correct soil background and aerosol scattering effects, and the normalized difference water index (NDWI), were selected in this paper. As for the calculation Formula (2), the values of coefficients G, C1, C2, and L were taken according to literature Testa et al. (2018). Moreover, a wide dynamic range vegetation index (WDRVI) was selected to narrow the gap between the contributions made by the near-infrared band and the red band on the vegetation index by a specific gravity coefficient. Besides, based on the original images, additional information such as the first 4 bands of the principal component analysis were added, till then, the total number of bands reached 13, which were taken to conduct crop classification again by the 3 classification methods in order to evaluate the influences of additional information on classification accuracy.

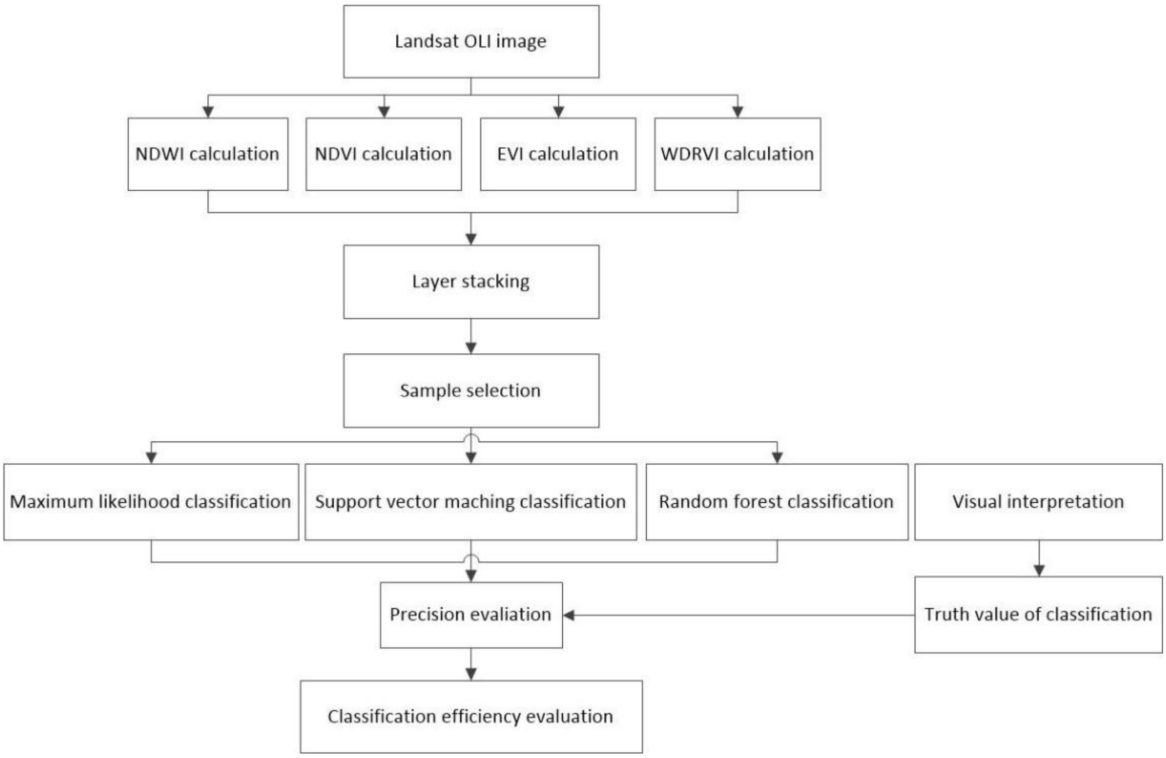

Landsat OLI image
NDWI calculation
NDVI calculation
EVI calculation
WDRVI calculation
Layer stacking
Sample selection
Maximum likelihood classification
Support vector maching classification
Random forest classification
Visual interpretation
Precision evaliation
Truth value of classification
Classification efficiency evaluation

FIGURE 4. Technical flow chart of the study

Four types of indexes were calculated according to the Landsat 8 multi-spectral reflectance data after taking pre-processing and used for carrying out dynamic monitoring on main crops in Wushi County region. The calculation formula is as follows.

$$NDVI = \frac{\rho NIR-\rho RE}{\rho NIR+\rho RED} \tag{1}$$

$$EVI = G\cdot\frac{\rho NIR-\rho RED}{\rho NIR+\ {}_1\cdot\rho RED-C_2\cdot\rho BLUE} \tag{2}$$

$$NDWI = \frac{\rho GREEN-\rho NIR}{\rho GREEN+\rho NIR} \tag{3}$$

$$WDRVI = \alpha\cdot\frac{\rho NIR-\rho RED}{\rho NIR+\rho RE} \tag{4}$$

Where,

ρNIR, ρRED, ρBLUE, and ρGREEN are the reflectivity of the near-infrared, red, blue, and green bands respectively;

G is the gain factor with value of 2.5;

C1 and C2 are the aerosol impedance coefficients with values of 6 and 7.5;

L is the canopy background adjustment factor with value of 1,

α is the weighting coefficient with value of 0.2.

Maximum Likelihood Algorithm

The maximum likelihood classification method, which is also known as the maximum likelihood estimation or Bayes classification method, is a kind of supervised classification method. Being created based on statistics principles, it establishes non-linear discriminant function set by the maximum likelihood ratio Bayes decision criterion and makes assumption that various distribution functions are normally distributed. Then, through sample training, it computes the probability of various pixels belonging to specific classified types. Finally, the pixels will be classified to the type with the highest probability. This is a supervised classification method commonly used in commercialized crop classification extraction process of agricultural condition remote sensing. It is featured in comparatively high classification accuracy, stable and reliable classification results, and rapid classification speed Sanhouse-García et al. (2016).

Support Vector Machine

The Support Vector Machine (SVM), which is a method of machine learning classification firstly put forward by Cortes & Vapnik (1995), is established on the basis of the statistical VC dimension (Vapnik-Chervonenkis Dimension) theory and the principle of minimum structural risk. According to the limited sample information, it tries to seek the optimal compromise between the model complexity and the learning ability, hoping to acquire the best promotion capability. For the image, the gray values of multiple bands of the image are regarded as one vector which should be mapped to a higher-dimensional space to build a hyperplane with the largest interval. This indicates that, 2 parallel hyperplanes should be built up at both sides of the hyperplane which separates data, thus, to achieve the largest interval between the two. It should be known that greater distance or gap should be guaranteed between the parallel hyperplanes, thus to minimize the total error of the classification. Therefore, by this way, we can achieve the optimal classification. On the other hand, this method can also learn sample classification knowledge automatically even in small samples, and acquire comparatively accurate classification results, so that it has been widely applied in multiple fields.

Random Forest Classification

Random Forest Classification (RFC) is a novel multi-decision-tree classification method proposed by Fu et al. (2017). It builds up multiple CART decision trees (no pruning) by random resampling on data and feature variables and determines the specific classified type that the data belongs to by multi-decision-tree voting. Targeting at the remote sensing image classification, the random forest method is featured in good anti-noise performance and high classification accuracy. It is one of the supervised classification methods that can construct the classification decision tree by sample data automatically.

(1) For the random forest algorithm, N training sample sets were extracted from the original sample data sets. And for each sample set, 2/3 was extracted from the original sample set by random sampling with replacement while the rest of 1/3 was taken as the verification sample, which was called the out-of-bag data (OOB) for estimating the internal errors. Besides, the OOB data were also used to study the importance of each feature variable.

(2) When building each tree, the random forest does not take all the features, instead, it adopts the random sampling with replacement to extract k (k⩽K) features from original feature set (suppose there are a total number of K features) as the basis for classification by decision tree, to build up data feature prediction variable sets. In general, the value of k could be set as square root of K.

(3) According to the selected training samples, verification samples, and feature prediction variable sets, a classification binary tree was established by recursion and the CART decision tree construction method. If the sample has k attribute features, for each attribute feature, a best partition value x was selected by referring to the Gini index. The smaller the Gini index, the lower the impurity content and the higher the classification accuracy in the classified types. Assuming that there are m types of a sample, the Gini index of node A of the binary tree was calculated according to the following formula.

$$Gini(A) = 1 - \sum_{i=1}^{m} p_i^2$$

Where,

$p_i$is the probability of belonging to type i. When Gini (A) =0, all samples belong to a same type. The recursive process is to take a try of every attribute feature of the sample for the current node, thus, to calculate and find out the minimum Gini index value in all attribute variables, which shall be taken as the best attribute partition value of the current node, thereby constructing an optimal branching sub-tree. Then according to the above splitting rules, the sufficient binary tree growth should be carried out on samples to build a complete CART tree. But in general, the tree is not pruned.

(4) Repeat step 3 till a number of N classification trees were well built up, thus, to form a forest of random classification trees. Then each pixel of the image was classified by all classification trees, and finally determined the type it belongs to according to the comprehensive classification results obtained by majority voting.

As for the random forest method, we adopted a double random sampling of both samples and features to build a decision tree, so even if we do not conduct pruning operation to the decision tree, it won’t result in any overfitting phenomenon as the traditional CART decision tree does.

Accuracy Verification Method

By method of visual interpretation combining with ground surveys, we selected areas of interest and carried out the supervised classification on the wheat, rice, woodland, beans, ice and snow, bare land, water bodies and other types of features in the entire study area. The results obtained by visual interpretation were taken as the data to verify the accuracy of the research results. 5 methods were used to describe and compare classification accuracy, including: confusion matrix, Kappa coefficient, overall accuracy of classification, cartographic accuracy, and user accuracy. For relevant definitions and detailed descriptions, please refer to the literature Piramanayagam et al. (2018), Ojaghi et al. (2016), Khorram et al. (2017). Besides, the statistical yearbook data were also involved in verifying the above accuracy.

RESULTS AND ANALYSIS

Comparison of Crop Classification Results Obtained by 3 Classification Methods

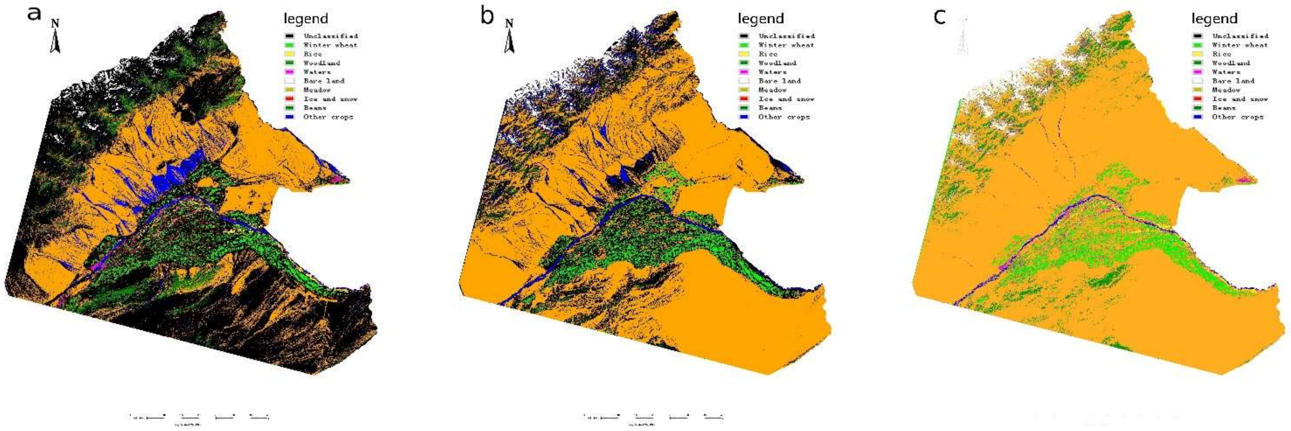
Based on the research technical process and the pre-processing of original images, the sample data were classified respectively by maximum likelihood classification, support vector machine, and random forest classification to acquire the classification results of 9 kinds of ground features, including: wheat, rice, forest land, beans, ice and snow, other vegetation, bare land and water bodies. Then the visual interpretation results were used to evaluate the classification accuracy and analyze the advantages and disadvantages of the three methods. Meanwhile, besides the original 7-band image, the NDVI, EVI, WDRVI, and NDWI were also calculated additionally. And the principal component analysis was carried out on the original image to extract the first 4 bands of the principle component. A total of 6 auxiliary bands were superimposed with the original image, thereby forming the original classification image data containing 13 feature bands. And then, the 3 classification methods were applied for classifying again, aiming to evaluate the influences of the additional feature bands on crop classification accuracy.

The 3 classification methods and a same ground sample were applied to conduct ground features classification, please refer to Figure 5. The number of decision trees in the random forest method was set as 100, and the number of the input feature variables during node splitting process was set as the square root of the total number of all features; as for the support vector machine method, the radial basis function (RBF)was taken as the kernel function while its Gamma value was set to be 0.071; for the maximum likelihood classification method, the segmentation probability threshold was set as a single threshold. According to Table 2 that, the overall classification accuracy of the maximum likelihood classification method, support vector machine method, and random forest method are 80.58%, 87.95%, and 95.96% respectively, while their Kappa coefficients are 0.61, 0.76, and 0.86 respectively. It shows that the classification accuracy of the random forest method is higher than the other two methods. This means that the random forest method is much more capable in crop classification and identification than traditional supervised classification methods.

TABLE 2. Confusion matrix of three classification methods based on original image.

Crop	Method	Wheat	Rice	Peas	Other crops	Woodland	Waters	Naked land	Water meadow	Ice	Total	Mapping accuracy（%）
Wheat	ML	200	0	0	0	0	0	0	0	0	200	99.50
	SVM	3338	0	0	0	0	0	0	0	0	3338	70.56
	RFC	4719	9	1	8	18	0	0	4	0	4759	99.75
Rice	ML	0	610	0	3	3	0	0	81	0	697	77.12
	SVM	0	94	0	0	0	0	0	0	0	94	11.88
	RFC	0	703	0	0	0	0	0	32	0	735	88.87
Peas	ML	0	6	100	82	1	0	0	5	0	194	96.15
	SVM	0	0	0	0	0	0	0	0	0	0	0
	RFC	0	0	63	0	0	0	0	0	0	63	60.58
Other crops	ML	1	87	4	318	0	0	0	16	0	426	76.81
	SVM	0	0	0	0	0	0	0	0	0	0	0
	RFC	0	0	0	142	0	0	0	0	0	142	34.30
Woodland	ML	0	0	0	2	2150	0	845	4	0	3001	89.14
	SVM	22	4	0	0	1482	0	11	1	0	1520	61.4
	RFC	2	1	0	0	1844	0	7	3	0	1857	76.45
Waters	ML	0	0	0	0	0	1069	75	0	0	1144	98.25
	SVM	0	1	0	0	0	816	0	0	1	818	75
	RFC	0	0	0	0	0	322	0	0	0	322	29.60
Naked land	ML	0	0	0	0	237	19	33036	16	0	33308	97.29
	SVM	0	0	0	0	192	187	33108	10	0	33497	97.50
	RFC	5	68	40	260	543	766	33950	20	1	35653	99.98
Water meadow	ML	0	88	0	9	21	0	1	612	0	743	83.65
	SVM	0	0	0	0	0	0	0	0	0	0	0
	RFC	5	10	0	4	7	0	0	687	0	713	92.09
Ice	ML	0	0	0	0	0	0	0	0	683	683	100
	SVM	0	0	0	0	0	0	0	0	676	676	98.98
	RFC	0	0	0	0	0	0	0	0	682	682	99.85
Total	ML	201	791	104	414	1212	1088	33957	746	683	40396	
Users accuracy (%)	ML	100	87.52	51.55	74.65	71.64	93.44	99.18	83.98	100		
	SVM	100	100	0.00	0.00	97.50	99.76	98.84	0.00	100		
	RFC	99.16	95.65	100	100	99.30	100	95.22	96.35	100		
Overall accuracy (%)	ML	80.58					Kappa coefficient	ML	0.6173			
	SVM	87.95						SVM	0.7677			
	RFC	95.96						RFC	0.8653			

MLC: maximum likelihood classification; SVM: support vector machine classification; RFC: random forest classification.



a: maximum likelihood classification result; b: support vector machine classification result; c:random forest classification result.

FIGURE 5. Classification results by three methods based on original image.

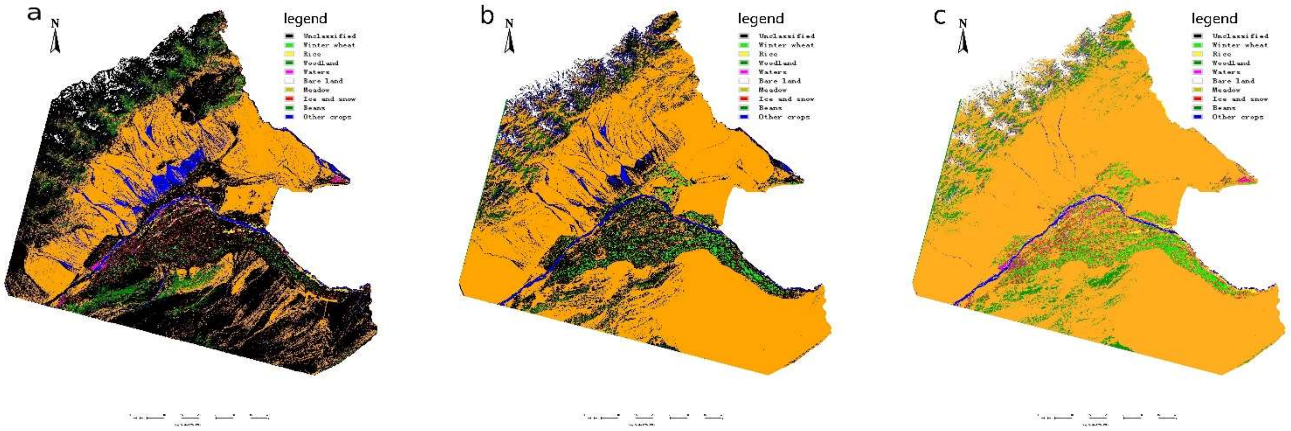
Comparison of Crop Classification Results Obtained by 3 Classification Methods after Adding with Auxiliary Classification Information

We calculated the NDVI, NDWI, EVI, and WDRVI of the original image separately, and conducted principal component analysis to extract the first 4 bands from the principal component brands. Then, we obtained 6 auxiliary classification bands, which were superimposed with the 7 bands of the original OLI image to form a 13-band image that was about to be classified. The same as the original image classification, 3 types of classification methods were applied respectively to classify crops of the same ground sample data and compare to the true-value image. As for the classification accuracy, please refer to Table 3. After adding the auxiliary information, the classification accuracy of the maximum likelihood classification and the support vector machine have been significantly improved. The overall accuracy of the MLC method increased by 3 percent while its Kappa coefficient increased by 0.06; the overall accuracy of the SVM method increased by 3.02 percent while its Kappa coefficient increased by 0.13; and the overall classification accuracy of the random forest method increased from 95.96% to 96.81%, indicating an increase of 0.85 percent, while its Kappa coefficient increased from 0.86 to 0.88. This shows that the adding of auxiliary information can improve the capability and accuracy of crops classification and identification to a certain extent.

TABLE 3. Confusion matrix of three methods based on the stacked image.

Crop	Method	Wheat	Rice	Peas	Other crops	Woodland	Waters	Naked land	Water meadow	Ice	Total	Mapping accuracy（%）
Wheat	ML	4225	1	0	2	0	0	0	8	0	4236	89.30
	SVM	3494	0	0	0	0	0	0	0	0	3494	73.85
	RFC	197	1	0	0	0	0	0	0	0	198	98.01
Rice	ML	1	548	0	1	0	0	0	78	0	628	69.28
	SVM	0	239	0	0	0	0	0	20	0	259	30.21
	RFC	0	759	0	0	0	0	0	21	0	780	95.95
Peas	ML	0	2	94	67	0	0	0	1	0	164	90.38
	SVM	0	0	0	0	0	0	0	0	0	0	0
	RFC	0	0	95	0	0	0	0	0	0	95	91.35
Other crops	ML	0	57	4	293	0	0	10	10	0	364	70.77
	SVM	0	0	0	2	0	0	0	0	0	2	0
	RFC	0	2	1	370	0	0	0	0	0	373	89.37
Woodland	ML	1	0	0	2	1911	0	728	3	0	2645	79.23
	SVM	30	6	0	0	1570	0	19	2	0	1627	65.09
	RFC	4	6	0	1	2028	0	8	7	0	2054	84.08
Waters	ML	0	0	0	0	0	936	46	0	0	982	86.03
	SVM	0	1	0	0	0	823	2	0	1	827	76
	RFC	0	1	0	0	0	323	0	0	0	312	29.69
Naked land	ML	0	0	0	0	203	12	28343	16	0	28574	83.47
	SVM	0	0	0	0	200	192	33121	10	0	33643	97.89
	RFC	0	14	7	42	383	765	33949	14	1	35175	99.98
Water meadow	ML	136	85	0	0	14	0	1	588	0	833	78.82
	SVM	0	0	0	0	0	0	0	0	0	0	0
	RFC	0	8	1	1	1	0	0	704	0	715	94.37
Ice	ML	0	0	0	0	0	0	0	0	611	611	89.46
	SVM	0	0	0	0	0	0	0	0	676	676	98.98
	RFC	0	0	0	0	0	0	0	0	682	682	99.85
Total	ML	4363	693	98	365	2128	948	29128	1408	613	39037	
Users' accuracy (%)	ML	100	87.26	57.32	80.49	72.25	95.32	99.19	70.59	100		
	SVM	100	92.28	0.00	100.00	96.50	99.52	98.81	0.00	100		
	RFC	99.49	97.31	100	99.20	98.73	100	99.98	94.37	99.85		
Overall accuracy (%)	ML	83.58					Kappa coefficient	ML	0.6747			
	SVM	90.97						SVM	0.8954			
	RFC	96.81						RFC	0.8791			

MLC: maximum likelihood classification; SVM: support vector machine classification; RFC: random forest classification.



a:maximum likelihood classification result; b:support vector machine classification result; c:random forest classification result.

FIGURE 6. Classification results of three methods based on the stacked image.

For the three methods, the red band and the 2 short-wave infrared bands in the original OLI data are more important to the classification, while the importance of coast, blue, green, and red bands are lower; After adding NDVI, EVI, WDRVI, NDWI and the first 4 bands of PCA, the red band and the 2 short-wave infrared bands are still more important to the classification that others. However, the band operation and principal component help generate new feature bands. Under the condition that the original information has been fully utilized, these changes could improve the valid information in data to certain extent. This shows that the adding of auxiliary information can enhance the capability and accuracy of crop identification and classification to some certain extent.

CONCLUSIONS

As a key algorithm in machine learning, the random forest method is featured in easy implementation, low computing cost, and less parameter adjustment, as well as rapid processing of massive high-dimensional data. Compared to maximum likelihood classification and support vector machine classification, the random forest method can achieve higher classification accuracy in an appropriate time. Meanwhile, by taking random sampling for both samples and feature attributes, this method achieves better generalization performance for the classification model acquired through final training, so that it can better process the data with noise (such as cloud and fog, etc.) and acquire comparatively high accuracy.

In this paper, all the maximum likelihood classification, support vector machine, and random forest classification used default parameters. As a matter of fact, the maximum likelihood classification and random forest classification could acquire comparatively high classification accuracy by general default parameters. And the parameter adjustment has little influence on the classification results. However, different kernel functions and corresponding parameter settings may have great influences on the results acquired by support vector machine classification, that is, higher classification accuracy may be achieved through parameter optimization. But the time spent on parameter optimization may be multiple times more than classification time, and the optimized parameters may not be the optimal among other image classifications. All these factors limit the application of support vector machine in practical business of agricultural condition remote sensing. On the other side, the random forest classification basically needs no parameter adjustment, but has even higher classification accuracy than the maximum likelihood classification method which also needs no parameter adjustment. Therefore, the random forest method can be an alternative of the maximum likelihood classification method to be widely used in the practical work of agricultural condition remote sensing.

Another advantage of the random forest method is that it can be used to sort the importance of image features. When the input image has many feature dimensions, the feature importance can be used for feature filtering, thus, to eliminate unrelated features, reduce computation, and improve identification accuracy. This is quite a significant advantage for application fields that have a great number of data features, such as multi-time-series image classification, and object-oriented classification, etc.

REFERENCES

Berhane TM, Lane CR, Wu Q, Autrey BC, Anenkhonov OA, Chepinoga VV, Liu H (2018) Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory. Remote sensing 10:580.

Cong M, Cui J, Peng X, Ji W (2018) Preliminary analytical method for unsupervised remote sensing image classification based on visual perception and a force field. Geocarto International 33:1350-1366.

Cong P, Chen K, Qu L, Han J (2019) Dynamic changes in the wetland landscape pattern of the Yellow River Delta from 1976 to 2016 based on satellite data. Chinese Geographical Science 29:372-381.

Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20:273-297. DOI: 10.1007/BF00994018.

Deschamps B, McNairn H, Shang J, Jiao X (2012) Towards operational radar-only crop type classification: comparison of a traditional decision tree with a random forest classifier. Canadian Journal of Remote Sensing 38:60-68.

Fu B, Wang Y, Campbell A, Li Y, Zhang B, Yin S, Xing Z, Jin X (2017) Comparison of object-based and pixel-based Random Forest algorithm for wetland vegetation mapping using high spatial resolution GF-1 and SAR data. Ecological indicators 73:105-117.

Gašparović M, Zrinjski M, Barković Đ, Radočaj D (2020) An automatic method for weed mapping in oat fields based on UAV imagery. Computers and Electronics in Agriculture 173:105385.

Gu X, Gao X, Ma H, Shi F, Liu X, Cao X (2019) Comparison of machine learning methods for land use/land cover classification in the complicated terrain regions. Remote Sensing Technology and Application.

Guermazi E, Bouaziz M, Zairi M (2016) Water irrigation management using remote sensing techniques: a case study in Central Tunisia. Environmental Earth Sciences 75:202.

Jamali A (2019) Evaluation and comparison of eight machine learning models in land use/land cover mapping using Landsat 8 OLI: a case study of the northern region of Iran. SN Applied Sciences 1:1-11.

Kandrika S, Roy PS (2008) Land use land cover classification of Orissa using multi-temporal IRS-P6 awifs data: A decision tree approach. International Journal of Applied Earth Observation and Geoinformation 10:186-193.

Khamparia A, Singh A, Luhach AK, Pandey B, Pandey DK (2020) Classification and identification of primitive Kharif crops using supervised deep convolutional networks. Sustainable Computing: Informatics and Systems 28:100340.

Khorram S, Nelson SA, van der Wiele CF, Cakir H (2017) Processing and applications of remotely sensed data. Handbook of Satellite Applications:1017.

Li-ping C, Yu-jun S (2018) Comparison of object-oriented remote sensing image classification based on different decision trees in forest area. Yingyong Shengtai Xuebao 29(12):3995-4003.

Muhammad S, Zhan Y, Wang L, Hao P, Niu Z (2016) Major crops classification using time series MODIS EVI with adjacent years of ground reference data in the US state of Kansas. Optik 127:1071-1077.

Murmu S, Biswas S (2015) Application of fuzzy logic and neural network in crop classification: a review. Aquatic Procedia 4:1203-1210.

Ojaghi S, Ahmadi FF, Ebadi H (2016) A new method for semi-automatic classification of remotely sensed images developed based on the cognitive approaches for producing spatial data required in geomatics applications. Arabian Journal of Geosciences 9:1-12.

Parida BR, Ranjan AK (2019) Wheat acreage mapping and yield prediction using Landsat-8 OLI satellite data: a case study in Sahibganj Province, Jharkhand (India). Remote Sensing in Earth Systems Sciences 2:96-107.

Peña JM, Gutiérrez PA, Hervás-Martínez C, Six J, Plant RE, López-Granados F (2014) Object-based image classification of summer crops with machine learning methods. Remote sensing 6:5019-5041.

Piramanayagam S, Saber E, Schwartzkopf W, Koehler FW (2018) Supervised classification of multisensor remotely sensed images using a deep learning framework. Remote sensing 10:1429.

Sanhouse-García AJ, Rangel-Peraza JG, Bustos-Terrones Y, García-Ferrer A, Mesas-Carrascosa FJ (2016) Land use mapping from CBERS-2 images with open source tools by applying different classification algorithms. Physics and Chemistry of the Earth, Parts A/B/C 91:27-37.

Singla SK, Garg RD, Dubey OP (2019) Streamlining multitemporal vegetation indices for dependable crop growth monitoring in Himalayan foothill region. Sādhanā 44:139.

Testa S, Soudani K, Boschetti L, Mondino EB (2018) MODIS-derived EVI, NDVI and WDRVI time series to estimate phenological metrics in French deciduous forests. International Journal of Applied Earth Observation and Geoinformation 64:132-144.

Wang H, Wang T, Zhou Y, Zhou L, Li H (2019) Information classification algorithm based on decision tree optimization. Cluster Computing 22:7559-7568.

Wu Q, Zhong R, Zhao W, Song K, Du L (2019) Land-cover classification using GF-2 images and airborne lidar data based on Random Forest. International Journal of Remote Sensing 40:2410-2426.

Xu J, Zhao H, Yin P, Jia D, Li G (2018) Remote sensing classification method of vegetation dynamics based on time series Landsat image: a case of opencast mining area in China. EURASIP Journal on Image and Video Processing 2018:1-10.

Zhan Y, Muhammad S, Hao P, Niu Z (2018) The effect of EVI time series density on crop classification accuracy. Optik 157:1065-1072.

Zhou Z, Huang J, Wang J, Zhang K, Kuang Z, Zhong S, Song X (2015) Object-oriented classification of sugarcane using time-series middle-resolution remote sensing data based on AdaBoost. PloS one 10:e0142069.