Identification of soybean in Argentina using Sentinel-2 composite images

Linsheng Huang[1], Yue Chen[1,2], Yuhao Pan[3], Zihang Lou[4], Shijun Zheng[5], Xiaoyang Zhang[6], Le Yu[5,7,8], Shengwei Liu[2,9], Dailiang Peng[2*]

(1. *National Engineering Research Center for Agro-Ecological Big Data Analysis & Application, Anhui University, Hefei 230601, China*;
2. *Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China*;
3. *Research Area of Ecology and Biodiversity, School of Biological Sciences, The University of Hong Kong, Pokfulam, Hong Kong, China*;
4. *College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China*;
5. *Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modeling, Institute for Global Change Studies, Tsinghua University, Beijing 100084, China*;
6. *Geospatial Sciences Center of Excellence, Department of Geography & Geospatial Sciences, South Dakota State University, Brookings, SD 57007, USA*;
7. *Ministry of Education Ecological Field Station for East Asian Migratory Birds, Beijing 100084, China*;
8. *Tsinghua University (Department of Earth System Science)- Xi'an Institute of Surveying and Mapping Joint Research Center for Next-Generation Smart Mapping, Beijing 100084, China*;
9. *School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, Henan, China*)

Abstract: Soybean is one of the most important oil crops, and Argentina is the third-largest soybean producer in the world, accounting for 17% of the global soybean yield. Timely and accurate information on soybean spatial distribution is critical for ensuring global food security. Sentinel-2 multispectral data and machine learning classification models are used to investigate the potential of soybean identification in the early stage of the growing season in Argentina, with the help of Google Earth Engine (GEE). The earliest time window and optimal feature set for soybean identification are explored. Results are as follows: 1) the random forest (RF) classification model demonstrated the highest level of classification accuracy compared to the backpropagation neural network (BPNN), support vector machine (SVM), and naive Bayes (NB) models; 2) Soybean can be accurately identified as early as the end of February (filling stage), which is approximately one month before harvest; 3) The optimal feature-subset can reduce the amount of input data by 80% while maintaining high classification accuracy. The overall accuracy (OA) of the RF classification model is 85.87%, and the relative error between the estimated soybean planting area and the agricultural statistics is 3.45%. This study provided a high-precision method for early-season identification of soybeans over large scales. The results can provide a data support for early futures trading and agricultural insurance, as well as a reference for policy-making to ensure global soybean food security.

Keywords: soybean, machine learning, time window, feature selection, Sentinel-2, Google Earth Engine
DOI: 10.25165/j.ijabe.20241705.7634

Citation: Huang L S, Chen Y, Pan Y H, Lou Z H, Zheng S J, Zhang X Y, et al. Identification of soybean in Argentina using Sentinel-2 composite images. Int J Agric & Biol Eng, 2024; 17(5): 266–274.

1 Introduction

Soybean is an important oil crop, which can provide high-quality protein. As a nitrogen-fixing plant, soybeans can reduce the use of fertilizers in fields[1]. In the next 20 years, soybean demand will increase to 1.5 times the current global yield[2]. Since the 21st century, Argentina has rapidly expanded its soybean planting area and has become the third largest soybean producer in the world, accounting for 17% of the global soybean yield[3]. Some scholars predict that by 2030, Argentina will surpass the United States and Brazil to become the largest soybean producer[4]. Argentina is also the third largest soybean exporter in the world and is the main supplier of soybean products including protein derivatives, soybean meal, soybean oil, and biodiesel in the European market[5]. Accurately obtaining soybean spatial distribution is the basis of yield estimation, water resource management, and disaster assessment[6]. Meanwhile, obtaining soybean planting area information prior to harvest can not only provide data support for early futures trading and agricultural insurance but also global food security and agricultural policy.

Satellite remote sensing technology can provide dynamic observations with wider coverage and higher efficiency than traditional statistical reports or field surveys. At present, the two major strategies for soybean identification based on remote sensing are: 1) selecting a single-date image within the key phenological period of soybean[7-9], and 2) using time-series images during one or

Received date: 2022-04-25 **Accepted date:** 2023-03-24
Biographies: Linsheng Huang, PhD, Professor, research interest: remote sensing information processing, Email: linsheng0808@163.com; **Yue Chen**, Master, research interest: crop classification and growth monitoring, Email: 18297935328@163.com; **Yuhao Pan**, PhD candidate, research interest: vegetation remote sensing, Email: panyuhao20@mails.ucas.ac.cn; **Zihang Lou**, PhD candidate, research interest: agricultural remote sensing, Email: louzihang21@mails.ucas.ac.cn; **Shijun Zheng**, PhD candidate, research interest: machine learning and deep learning, Email: zhengshijun19@mails.ucas.ac.cn; **Xiaoyang Zhang**, PhD, Professor, research interest: land cover and land use change, Email: xiaoyang.zhang@sdstate.edu; **Le Yu**, PhD, Associate Professor, research interest: cropland mapping and land cover, Email: leyu@tsinghua.edu.cn; **Shengwei Liu**, Master, research interest: crop classification and growth monitoring, Email: 212004010029@home.hpu.edu.cn.

***Corresponding author: Dailiang Peng**, PhD, Professor, research interest: vegetation remote sensing. Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China. Tel: 86-10-82178166, Email: pengdl@aircas.ac.cn.

multiple growing seasons of soybean[10-14]. The first strategy is time-efficient and user-friendly. However, in many places, the clouds may prevent or delay the acquisition of images, leading to low interannual transferability of the single-date image method[15]. In addition, the classification results of the single-data image are inaccurate for areas with complex planting structures and high spectral similarities[16]. The second strategy mainly uses time-series images to capture the dynamic information of soybeans at different growth stages[6,16,17]. This approach has been widely studied in the large-scale identification of soybeans. In the early 21st century, much of the research relied on low spatial resolution data, such as MODIS and AVHRR[18-22]. Picoli et al.[23] used a support vector machine model, combined with MODIS time series to map nine land cover categories in the state of Mato Grosso, Brazil. However, the accuracy of soybean planting area extraction using medium to low spatial resolution time series data was hampered by the problem of mixed pixels due to the relatively low spatial resolution. In recent years, the launch of high-resolution and high-frequency satellites has improved soybean identification technology[24,25]. Paludo et al.[26] used Landsat-8, Sentinel-2, and SRTM data, combined with a simple non-iterative clustering segmentation method and a naive Bayes classifier to identify soybean and maize planting areas in the state of Paraná, Brazil. However, most studies are based on remote sensing images of the entire growth cycle of soybean, and the spatial distribution information of soybean can only be obtained after soybean harvest or even several months after harvest[27-33]. For example, the cropland data layer (CDL) of the United States Department of Agriculture (USDA) is not released until the beginning of the next calendar year. To some extent, this approach restricts production management and decision-making of soybeans in early or in-season of land lease, water and fertilizer management, crop insurance, harvest, and transportation coordination. Therefore, studying the precise identification of soybeans in the early growing season over large scales is urgent.

Several recent signs of progress make it possible for us to identify soybeans within the growing season over large scales. Sentinel-2 series of satellites with 5 d revisit cycles and 10 or 20 m spatial resolution can provide abundant temporal and spectral features[34,35], which are widely used in soybean classification[36-38]. In addition, Google Earth Engine (GEE), as a cloud data platform, has powerful computing capabilities, which can not only conveniently call, analyze, and process various satellite images and geospatial datasets, but also provide various classification algorithm interfaces[39,40]. Abundant satellite data and powerful computing platforms provide strong support for large-scale soybean identification.

However, challenges will inevitably emerge when identifying soybeans in the early stage of the growing season in Argentina. 1) The uncertainty of effective observation frequency in time and space can affect the classification results due to the different satellite orbits and location of cloud contamination[36]; 2) Fewer satellite observations can be utilized in the early season soybean identification than in the post-season soybean identification; 3) Numerous features are used to identify soybeans, such as vegetation index and texture[7,36,41]. Nevertheless, these high-dimensional input features increase the complexity of classification[30]. This study attempted to solve these challenges by using the following methods: 1) To obtain a homogeneous time series, the equal temporal interval composite method was used to build regular interval time-series images on a large scale[37,42,43]; 2) To fully dig out the spectral information in the early stage of the growing season, the incremental time window method was used to study the relationship between identification accuracy and seasonal variation[44] to explore the earliest time window of soybean identification; 3) To evaluate the potential of using only Sentinel-2 multispectral bands for early-season identification of soybean in Argentina, while avoiding the Hughes effect, the importance of each Sentinel-2 multispectral band for soybean identification in early-season was evaluated. Then, feature selection was used to decrease redundant information[36,45] to find the optimal feature subset for early-season identification of soybeans.

The main purpose of this study is to accurately identify soybeans in the early stage of the growing season in Argentina. The GEE platform was used to construct composite images of the Sentinel-2 multispectral bands in the soybean growing season 2019/2020 in Argentina's main agricultural areas. Subsequently, the machine learning classification models were used to explore the earliest time window for soybean identification, and the optimal feature subset was selected by assessing the importance of all spectral and temporal features during the time window. Finally, the spatial distribution of soybeans in Argentina was mapped.

2 Materials and methods

2.1 Study area

Argentina is located in the southeast of South America, facing the Atlantic Ocean in the east, Antarctica across the sea in the south, Chile in the west, and on the border with Bolivia and Paraguay in the north. With an area of 2.78 million km^2, it is the second-largest country in Latin America and eighth in the world. As shown in Figure 1a, the study area covers the main agricultural areas of Argentina according to the Buenos Aires Grain Exchange (2019) zonation[46], containing the main agricultural areas of 15 provinces. The geographical range is 22°0′S-41°21′S，56°42W-67°23′W, with a total area of approximately 1.3 million km^2, mainly located in temperate and subtropical climates with fertile soil and abundant rain. It is suitable for agricultural development. In addition to soybean, the main crops include maize and wheat. Soybean is mainly sown at the beginning of November and harvested at the beginning of April of the following year. The soybean planting area of the study area accounts for 99.8% of that in the whole extent of Argentina[47].

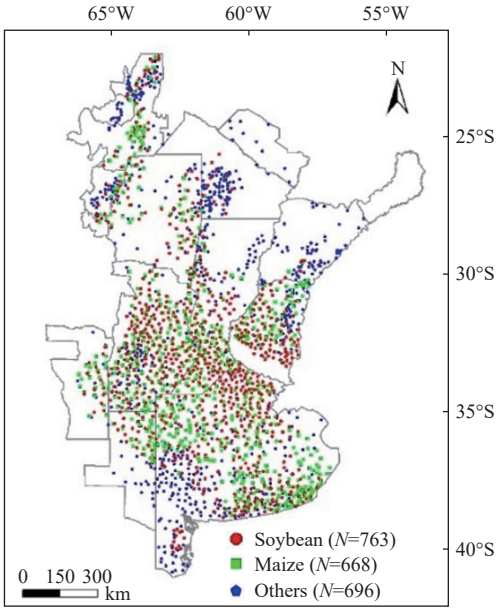


Figure 1 Distribution and number of samples in Argentina's main agricultural areas

2.2 Data

Sentinel-2 satellite series are composed of two identical satellites, A and B, with a temporal resolution of 5 d and 13 multispectral bands with a spatial resolution of 10, 20, and 60 m. This research selected November 1, 2019, to April 30, 2020, as the study period, considering the growth period of soybeans. Sentinel-2 top-of-atmosphere (TOA) reflectance data were used instead of surface reflectance (SR) due to the limited temporal availability of SR imagery in the GEE platform for the study area. The reliability of TOA reflectance data for image classification has been proven in previous studies[48,49], and lots of recent efforts have used S2 TOA reflectance data to identify crops[36,40,42,45,50-52]. The quality assessment (QA60) band was used to limit the cloud cover percentage to less than 10%[40,42]. All the band values were divided by 10 000 to obtain the band reflectance value. All 10 and 20 m spatial resolution bands of Sentinel-2 were selected, including blue, green, red, red-edge 1 (RE1), red-edge 2 (RE2), red-edge 3 (RE3), near-infrared (NIR), narrow infrared (Narrow NIR), shortwave infrared 1 (SWIR 1), and shortwave infrared 2 (SWIR2), a total of 10 bands were obtained. The cropland data were from GlobeLand30 products[53], which are used to mask non-farmland areas.

Ground reference data were obtained from the national crop map published by the National Institute of Agricultural Technology of Argentina[54]. Three categories of sample types were selected including soybean, maize, and others, considering that the phenological periods of maize and soybean are close. The other types mainly consist of sunflower, peanut, cotton, sugarcane, and sorghum. The following steps are performed to determine the sample data. First, larger sample plots (greater than 25 hm^2) which are the same crops in growing seasons 2018/2019 and 2019/2020 were selected as the sampling area for each sample type. Second, evident error points were visually removed from the high-resolution Google Earth image. The distribution and number of samples are shown in Figure 1. A total of 70% of the samples are randomly selected as training samples and the remaining 30% as validation samples.

The agricultural statistical data include the national and provincial soybean planting area data of Argentina, which are from the Ministry of Agriculture, Livestock and Fisheries[47].

2.3 Classifier setting

To evaluate different machine learning classification models for soybean identification in Argentina, four widely used models in crop remote sensing identification were selected in this study. The performance of each model was compared by calculating the overall accuracy (OA) and Kappa coefficient. Due to the large scale of the study area (approximately 1.3 million km^2), some classifiers require downloading images for local processing, which requires significant computational time and storage space. Consequently, the comparison experiment of classifiers was conducted in the top three soybean-producing provinces (Buenos Aires, Cordoba, and Santa Fe), which account for 75% of the total soybean planting area. Additionally, to ensure accurate identification, this study selected the median composite image[37,45] of Sentinel-2 in February 2019[7,13,15], during the filling stage of soybean, as the input image. The classification models and parameter settings are described as follows:

Back-propagation Neural Network (BPNN) is a multi-layer feedforward neural network training based on the error backpropagation algorithm[55]. Support Vector Machine (SVM) is a binary classification algorithm that aims to find the optimal separating hyperplane in a high-dimensional space, maximizing the distance between samples on both sides of the hyperplane to divide different classes[56]. Naive Bayes (NB) classification algorithm is based on the Bayes theorem and the assumption of feature independence[26]. Random Forest (RF) is a classification algorithm based on multiple decision trees, using a random and put-back sampling method when selecting a subset of training samples. Each tree is constructed from approximately two-thirds of the training samples, with the remaining third used for test classification, which are known as out-of-bag (OOB) samples. Through the OOB samples, the classification performance of each tree can be evaluated to determine the final performance of the classifier[57]. RF can quantify the importance of variables, which makes it useful for feature ranking or selection[1,50]. The importance of features was evaluated based on the mean decrease Gini (MDG)[58,59].

In this study, the BPNN algorithm (Neural Net Classification) in ENVI 5.3 software was used, with the number of iterations increasing from 500 to 2000 in increments of 500 to balance computational complexity and classification accuracy. The final number of iterations was set to 1000, while other parameters remained at their default values. The SVM algorithm (ee.Classifier.libsvm) in GEE was used, with the radial basis function (RBF) selected, and the cost parameter increased from 50 to 300 in increments of 50 to balance computational time and classification accuracy. The final cost parameter was set to 100, while other parameters were kept at their default values. The NB algorithm (ee.Classifier.smileNaiveBayes) was called in GEE, with only one parameter requiring adjustment, which was kept at its default setting. The RF algorithm (ee.Classifier.smileRandomForest) was called in GEE, with the number of trees increased from 100 to 500 in increments of 100 to balance computational complexity and classification accuracy. The final number of trees was set to 200, while other parameters were kept at their default values[45,60].

2.4 Time interval of composite images setting

The equal temporal interval median composite method was used to build regular interval time-series images in the study area[37,42,52]. First, during the study period, the median value of observations in the temporal intervals of 5, 15, and 30 d was obtained to build composite images in each temporal interval which were called single-period images (for example, November 1, 2019, to November 5, 2019, is the first single-period image of a 5-day temporal interval). Then, the percentage of pixels that have at least one effective observation within each temporal interval was counted to illustrate the coverage of Sentinel-2 images in the study area. To capture more detailed spectral information in the early stage of the growing season, the temporal interval of single-period images should be sufficiently narrow. However, an extremely narrow temporal interval may not be sufficient to fill in the gaps. Although the method of data interpolation can be used to fill in the gap, it is generally time-consuming and the results have some uncertainty for the data missing over a long time and a large range[51,52,61]. Therefore, the narrower interval and the smaller degree of data missing were used to determine the time interval of composite images.

2.5 Time window of soybean identification selection

The incremental time window method was used to explore the earliest identifiable time of soybean[50,62]. Starting from November 1, 2019, the time interval determined in Section 2.4 is gradually extended to 28 April 2020 to form equal temporal interval time-series images with different lengths. The influence of these time-series images on classification accuracies was assessed to determine the earliest time window of soybean identification. The main steps are the follows: 1) Gap filling is performed on the missing part of

each single-period image using the average value of the two images before and after[40]; 2) The overall accuracy (OA) and kappa coefficient (KC) of RF classification of each single-period image are calculated; 3) Starting from the first single-period image and sequentially adding the subsequent single-period image, the different lengths of equal temporal interval time-series images are formed; 4) The OA and KC of equal temporal interval time-series images with different lengths were calculated, and the soybean planting area was estimated at the same time; 5) The classification accuracy reaches the stable level or no longer significantly increases to determine the time window. The earlier time window is more valuable to the decision-making activities.

2.6 Optimal feature-subset construction

Feature selection was used to decrease the number of features and calculation runtime[63]. Sequential forward selection (SFS) is used to determine the dimension of the optimal feature subset for soybean identification[7]. SFS is a greedy algorithm for finding the optimal subset of features. The feature subset starts from an empty set, and then a new feature is added each time. The final feature subset is determined according to the optimal feature function. The main steps are as follows: 1) All the features in the time window determined in Section 2.5 were input into the RF classifier to obtain the importance score of each feature, then, the Min-Max Scaling was used to map the score to the range of[36]; 2) The importance of all features was arranged in descending order; 3) The feature subset starts from the empty set. According to the feature sequence in step 2, one new feature was added to the feature subset each time, then the OA of the current feature subset was calculated, and the soybean planting area was estimated at the same time; 4) The OA trend after each new feature was added was observed until the OA reached a stable level, or it no longer increased significantly. This condition indicated that relatively high accuracy can be obtained with fewer features and that the amount of input data and the calculation cost are significantly reduced.

In this study, two methods were mainly used to evaluate the accuracy of soybean identification, 1) by validation samples, a confusion matrix was generated to evaluate the accuracy of classification, and 2) by using a soybean distribution map, soybean planting area at national and provincial level is estimated, and then the accuracy of area extraction is evaluated by comparing with agricultural statistical data. The accuracy indexes used were the OA, KC, relative error, and coefficient of determination (R^2).

3 Result analysis and discussion

3.1 Classifier selection

Figure 2 shows the classification accuracy of different models, and it is evident that the RF classifier outperformed the other three models, achieving the highest OA and Kappa of 80.20% and 0.70, respectively. The BPNN classifier had a relatively high OA of 79.39%, while its Kappa was close to SVM. On the other hand, the NB classifier exhibited the lowest OA and Kappa among all classifiers, indicating its poor performance in classification. Moreover, the RF classifier is convenient to operate in the GEE platform, and the importance of each feature can be calculated using the explain function, which facilitates feature selection. Previous studies have extensively employed the RF classifier for soybean remote sensing identification[1,6,30,31,36,45,64]. Based on these findings, the RF classifier was selected for further research in this study.

3.2 Determination of time interval of composite images

Figure 3 shows that during the study period, the percentage of pixels which has at least one effective observation within each temporal interval in the study area, including 5, 15, and 30 d. A large degree of missing data was observed in multiple consecutive periods in the 5 d temporal interval, especially in the early stage of the growing season. For example, in the period within 20-24 d of the year (DOY) in 2020, only 43.45% of the pixels obtained at least one effective observation. This indicates that 56.55% of the pixels have encountered missing data. Thus, the 5 d temporal interval is ruled out. With increasing composite temporal interval, more pixels can satisfy the requirement for at least one effective observation in every period. In addition, during the 15 d temporal interval, slight data loss is observed in several periods (the maximum percentage of missing pixel data is 4.86%). In the 30 d temporal interval, every period at approximately 100% of the pixels obtain at least one effective observation, that is, almost no data is missing. Compared with the 30 d temporal interval, the 15 d temporal interval can capture more detailed image information. Therefore, to explore a shorter time window for soybean identification, the 15 d temporal interval was finally selected as the time interval for composite images.

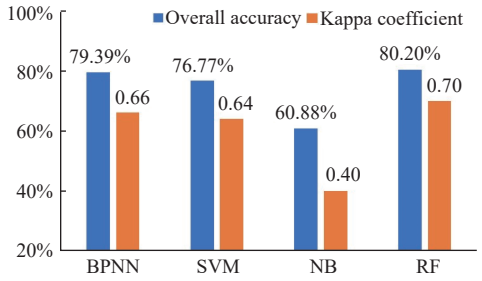
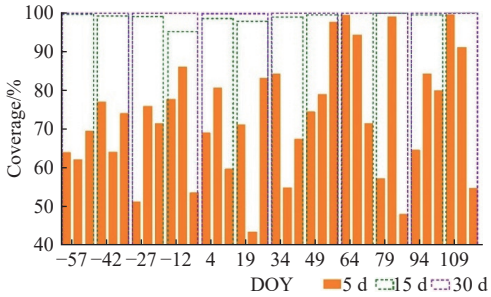


Figure 2 Overall accuracy and Kappa coefficient of classification results from different models



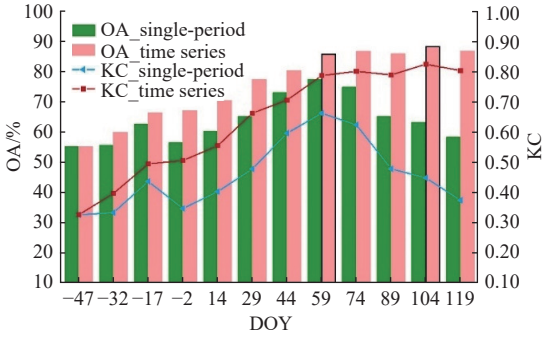
Note: DOY in this figure represents the deadline of each period in the 5 d temporal interval (DOY of −57 represents the period of −61 to −57 d of 2020), it is displayed once every 15 d, same below. The date on the horizontal axis takes January 1, 2020 as 1, forward 2020 as positive and backward to 2019 as negative.

Figure 3 Percentage of pixels with effective observations in the study area (5, 15, 30 d temporal intervals)

3.3 Determination of time window of soybean identification

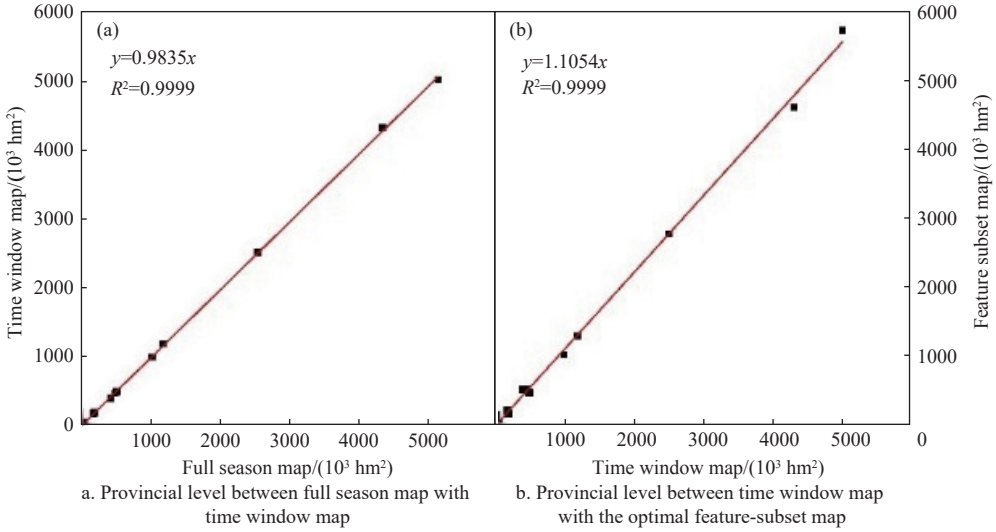
The OA and KC of RF classification of each 15 d single-period image and 15 d time-series images of different lengths are shown in Figure 4. From the single-period images, the OA and KC initially increase and then decrease. Among them, the classification accuracy of single-period images with DOY of 45 to 59 is the highest (OA: 77.61%, KC: 0.66). Moreover, the OA and KC of different lengths of time-series images indicate that with the addition of more images, the classification accuracy continuously improves and finally tends to be stable. Subsequently, compared with single-period images (the maximum OA is 77.61%), the classification accuracy of time-series images is higher (the maximum OA is 88.43%), consistent with the previous research

results[37,65]. This means that time-series images can provide richer dynamic information, which enables the classifier to easily capture the changes between different categories, thus improving the classification accuracy. However, it eventually tends to be saturated, with the increase in data. Specifically, when the time-series images with DOY of −61 to 59 are used for classification (the first black frame from the left in Figure 4), acceptable classification accuracy is obtained (OA: 85.90%, KC: 0.79). After adding more images, the classification accuracy fluctuates in a small range. Among them, the classification accuracy of the time-series images with DOY of −61 to 104 is the highest (the second black frame from the left in Figure 4), but the improvement is limited compared with the classification accuracy with DOY of −61 to 59 (the OA and KC are increased by 2.54% and 0.04, respectively), and the time window is extended by 45 d. The estimated soybean planting area is calculated from the two time-series images. According to Figure 5a, the R^2 of the two is 0.99. Thus, the time-series images with DOY of −61 to 59 can achieve the approximate effect of using the maximum OA of the full season. Therefore, to identify the soybean as early as possible, the DOY of −61 to 59 (November 1, 2019, to February 28, 2020) is selected as the time window, during which soybean is from the sowing stage to the filling stages, approximately one month earlier than the soybean harvest, which is close to the previous research results[15,22].



Note: DOY indicates the deadline for each single-period image and from November 1, 2019 to the current deadline of the time-series images (DOY −47 represents the period −61 to −47 d of 2020). KC: Kappa coefficient.

Figure 4 OA and KC of RF classification of 15 d single-period images and 15 d time-series images with different lengths



a. Provincial level between full season map with time window map

b. Provincial level between time window map with the optimal feature-subset map

Note: The time window and full season refer to the period in the DOY of −61 to 59 and DOY of −61 to 104, respectively.

Figure 5 Regression of estimated soybean area at provincial level between full season map with time window map and between time window map with the optimal feature-subset map

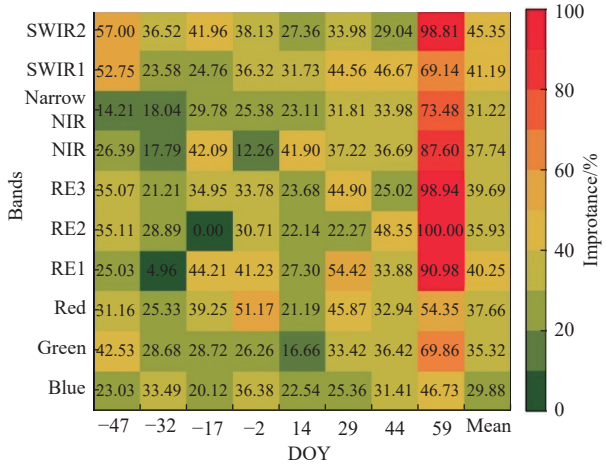
3.4 Determination of optimal feature subset of soybean identification

All the features within the time window are input into the RF classifier, and the importance of each feature was obtained, as shown in Figure 6, where the rightmost column (mean) refers to the mean of each spectral band in the horizontal axis direction. This column indicates that the shortwave infrared and red-edge bands are more important to the classification than the near-infrared and visible bands, consistent with the previous research conclusions[6,37,65]. Among them, the important mean values are shortwave infrared 2 (45.35%)>shortwave infrared 1 (41.19%)>red-edge 1 (40.25%)>red-edge 3 (39.69%)>red-edge 2 (35.93%). Compared with individual features, the spectral bands with DOY of 45 to 59 (in late February) are highly important, during the period soybean is in filling stages. Therefore, the images at the peak of the crop growth period are crucial in crop identification[6].

According to the SFS, the features in descending order of importance are added to the RF classifier in turn, then the relationship between the number of classification features and the OA is observed, as shown in Figure 7. First, the red-edge 2 with DOY of 45 to 59 is added, the OA is 48.01%. Subsequently, the red-edge 3 with the DOY of 45 to 59 is added, the OA reaches 63.96%. When the ninth feature (shortwave infrared 2 with DOY of −61 to −47) is added, the OA reaches a local maximum of 83.87%, then it continues to fluctuate in a small range. When the 16th feature (shortwave infrared 1 with DOY of 30 to 44) is added, the OA reaches 85.87%, which is close to the OA of classification using 80 features. When the 36th feature (red-edge 3 with DOY of −31 to −17) is added, the local maximum OA reaches 87.48%, which is slightly higher than the accuracy when 80 features are input, thereby verifying the existence of the Hughes effect, that is, using all data into the classifier not only wastes calculation time, but also reduces classification accuracy[6,45,66].

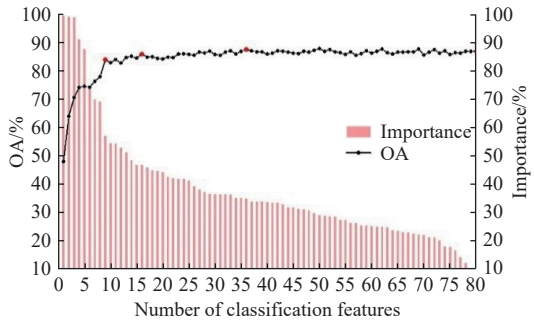
The OA is relatively high (83.87%, 85.87%, and 87.48%) when the first nine, 16, and 36 features are added. The relative errors between the estimated soybean planting area and agricultural

statistical data are 9.03%, 3.45%, and −6.87%. Therefore, after weighing the amount of input data and the classification effect, the feature subset consisting of 16 features is finally selected. The soybean planting area estimated by the optimal feature subset is compared with that estimated by all features in the time window. According to Figure 5b, the R^2 is 0.99. This shows that the optimal feature subset can represent all the features in the time window while reducing the amount of calculated data by 80% (the number of features decreased from 80 to 16). The detailed data are compared, as shown in Table 1. The table lists that when the number of input features is large, similar results can be obtained with fewer inputs through feature selection while reducing computational complexity and time cost. In addition, the optimal feature subset mainly includes the following 16 features: shortwave infrared 1 and shortwave infrared 2 with DOY of −61 to −47, red with DOY of −16 to −2, red-edge 1 with DOY of 15 to 29, shortwave infrared 1 and red-edge 2 with DOY of 30 to 44, and all 10 and 20 m resolution bands of Sentinel-2 with DOY of 45 to 59. Among them, DOY of 45 to 59 (in late February), accounting for 10 features, fully illustrates the importance of bands in late February for classification. At the same time, shortwave infrared 1 appears three times, and shortwave infrared 2, red-edge 1, and red-edge 2 appear two times, verifying that shortwave infrared and red-edge bands contribute significantly to soybean identification.



Note: The rightmost column (mean) refers to the mean of the horizontal axis direction of each spectral band. The color change from green to red shows the increasing trend of variable importance. The horizontal axis date refers to the deadline for each 15 d temporal interval and the vertical axis refers to the band names.

Figure 6 Importance of all features in the time window



Note: The histogram shows the importance of the newly added feature.

Figure 7 Relationship between the number of participated features in the RF classifier and the OA of classification

Table 1 OA and KC of classification and the relative error of estimated soybean planting area for different images

Images	Full season	Time window	Nine features	Optimal feature-subset	36 features
Number of features	110	80	9	16	36
OA/%	88.43	85.90	83.87	85.87	87.48
Kappa coefficient	0.83	0.79	0.76	0.79	0.81
Relative error/%	−5.14	−6.77	9.03	3.45	−6.87

Note: full season refers to the period from 1 November 2019 to 13 April 2020 (DOY of −61 to 104).

3.5 Soybean identification and accuracy evaluation

The spatial distribution map of soybeans in the study area in the growing season 2019/2020 is drawn using the optimal feature subset determined in Section 3.4, as shown in Figure 8. The OA is 85.87% and KC is 0.79. The estimated soybean planting area is 17.49 million hm^2, and the relative error is 3.45% compared with agricultural statistics. It was clear that the total soybean planting area is almost equal to the statistics. In order to further verify the mapping quality, the soybean planting area was calculated at the provincial level and compared with the statistics. The result is shown in Figure 9, the slope of linear fit is 1.04 and the R^2 is 0.99. Obviously, the remarkable correlation between the two was confirmed by the slope values and R^2, which are relatively close to 1. What's more, comparing soybean mapping with Sentinel 2 composite images also showed high consistency (Figure 10). Therefore, this method can extract soybeans effectively, rapidly, and accurately.

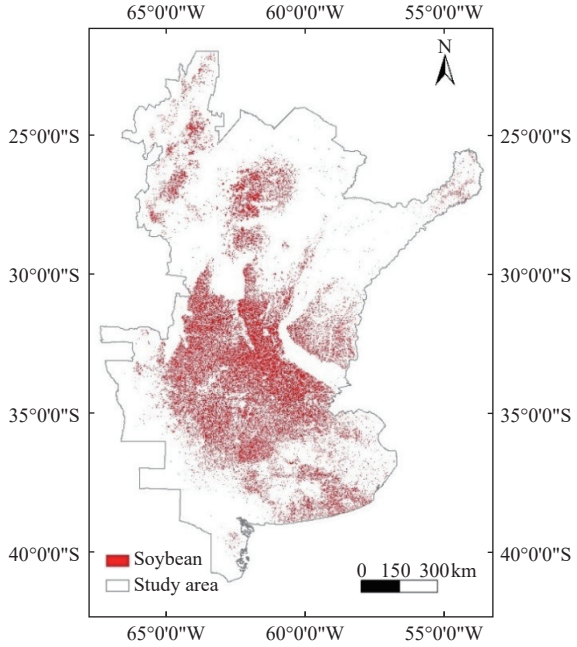


Figure 8 Spatial distribution map of soybeans in the study area in the growing season of 2019/2020

The soybean planting area in Argentina in the growing season 2018/2019 and 2020/2021 are estimated using the same policy to verify the robustness of this method. The results show that the estimated soybean planting area in the growing season 2018/2019 and 2020/2021 are 15.38 and 16.93 million hm^2, respectively, with a relative error of −9.59% and 1.70% compared with the agricultural statistical data. According to Figure 9, the R^2 of the provincial estimated soybean area and agricultural statistics area are 0.98 and 0.99, respectively. Therefore, this method has good robustness.

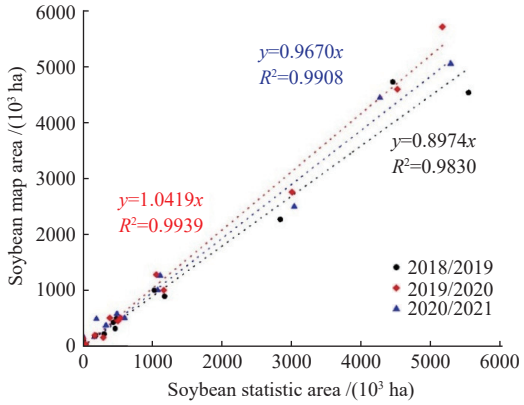


Figure 9 Comparison of the estimated soybean area from the distribution map with the agricultural statistics at a provincial level in the growing season of 2018/2019, 2019/2020, and 2020/2021

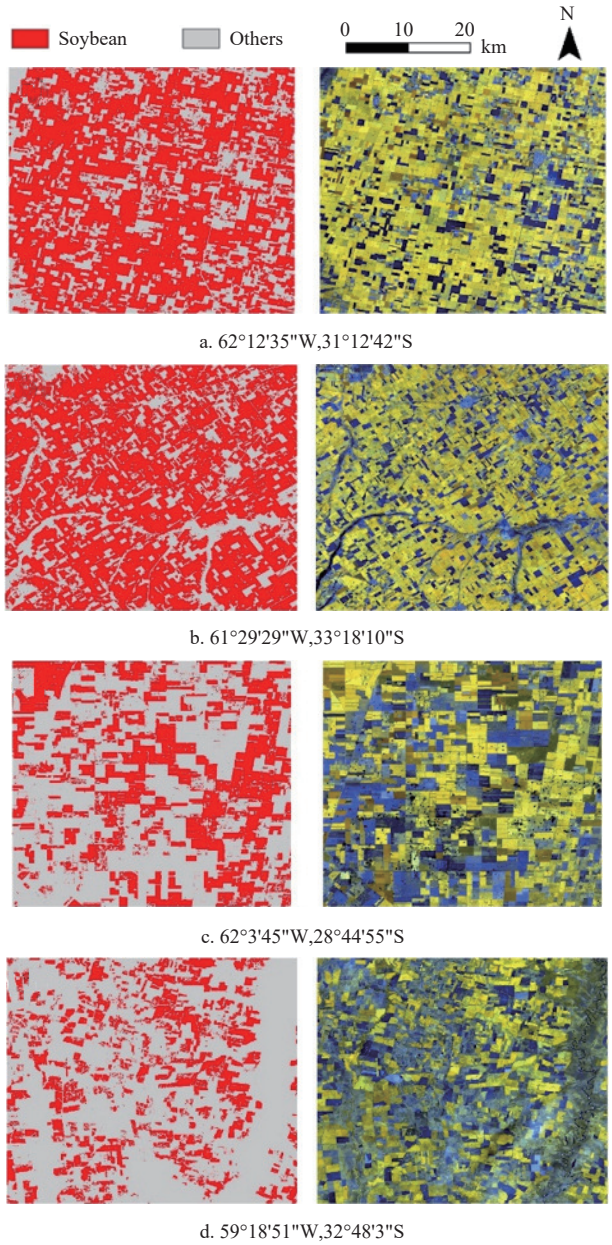


Figure 10 Comparisons between soybean mapping and Sentinel 2 imagery which is red-green-blue (R-G-B) composites of RE3, RE2, and RE1

Uncertainty and implication for future studies. First, due to the complex environmental and climatic conditions on the national scale, soybean sowing times and growth states can be varied (intraclass variability)[6] and lead to misclassification of classifiers. Second, the use of cropland data, Sentinel-2 TOA reflectance data, and image cloud removal processing methods may have a certain impact on soybean identification. In addition, the estimated soybean planting area has a certain error due to classification errors and the existence of mixed pixels[67]. In the future, when the ground survey work is relatively easy to carry out, the distribution map can be considered an auxiliary variable, and then the survey estimate can be used as a dependent variable for regression analysis to obtain more reliable soybean distribution and area information[10]. At the same time, the object-based approach can reduce the influence of pixel heterogeneity to a certain extent and improve the accuracy of soybean identification[41].

4 Conclusions

With the help of the Google Earth Engine (GEE) platform, Sentinel-2 multispectral data and machine learning classification model were used, the earliest time window and optimal feature-subset of soybean identification in Argentina were determined, and then the spatial distribution of soybean in the growing season 2019/2020 was mapped. The main conclusions of this study are as follows:

1) Among the four machine learning models, the random forest (RF) classifier demonstrated the highest level of classification accuracy, followed by the backpropagation neural network (BPNN) and support vector machine (SVM) classifiers, while the naive Bayes (NB) classifier exhibited poor performance;

2) Soybean can be accurately identified during the filling stage, which occurs approximately one month before harvest;

3) A vital temporal period for soybean identification in Argentina is in late February. Shortwave infrared and red-edge bands play a significant role in soybean identification. The input data can be reduced by 80% (from 80 to 16 features) by using the optimal feature subset after feature selection while maintaining high classification accuracy, the OA is 85.87% and the relative error of the estimated soybean planting area is 3.45%.

4) In addition, the soybean planting area in the growing season 2018/2019 and 2020/2021 are estimated using the same policy, these areas are in good agreement with the agricultural statistics (relative error<10%), thereby verifying the robustness of the method.

In general, this study found that Sentinel-2 composite images can be used to rapidly identify soybeans by using the optimal feature subset approximately 1 month prior to soybean harvest in Argentina. The research results provide a scientific reference for the production management and decision-making of large-scale crops in the early season or during the season. In the future, guided by these research conclusions, larger-scale crop identification and regional crop suitability evaluations can be carried out.

Acknowledgements

This work was funded by the Science and Disruptive Technology Program, AIRCAS(Grant No. 2024-AIRCAS-SDPT-15), and the National Natural Science Foundation of China (Grant No. 42471372).

[References]

[1] Li X Y, Yu L, Peng D L, Gong P. A large-scale, long time-series (1984–2020) of soybean mapping with phenological features: Heilongjiang Province as a test case. International Journal of Remote Sensing, 2021;

42(19): 7332–7356.
[2] King L A, Adusei B, Stehman S V, Potapov P V, Song X P, Krylov A, et al. A multi-resolution approach to national-scale cultivated area estimation of soybean. Remote Sensing of Environment, 2017; 195: 13–29.
[3] Food and Agriculture Organization of the United Nations (FAO). Available: http://www.fao.org/faostat/en/#data/QC. Accessed on [2024-09-04].
[4] Masuda T, Goldsmith P D. World soybean production: area harvested, yield, and long-term projections. International Food & Agribusiness Management Review, 2009; 12(4): 143–162.
[5] Hilbert J A, Luciana S, Manosalva J, Patricio G, Ariana C. Greenhouse gas emission from the cultivation of soybean in Argentina. 2021. Available: https://www.researchgate.net/publication/350767239_Greenhouse_gas_emission_from_the_cultivation_of_soybean_in_Argentina. Accessed on [2024-09-04].
[6] Zhang H Y, Kang J Z, Xu X, Zhang L P. Accessing the temporal and spectral features in crop type mapping using multi-temporal Sentinel-2 imagery: A case study of Yi'an County, Heilongjiang province, China. Computers and Electronics in Agriculture, 2020; 176: 105618.
[7] She B, Yang Y Y, Zhao Z G, Huang L S, Liang D, Zhang D Y. Identification and mapping of soybean and maize crops based on Sentinel-2 data. Int J Agric & Biol Eng, 2020; 13(6): 171–182.
[8] Wang L M, Liu J, Yang L B, Yang F G, Fu C H. Impact of short infrared wave band on identification accuracy of corn and soybean area. Transactions of the CSAE, 2016; 32(19): 169–178. (in Chinese)
[9] Wang L M, Liu J, Yang L B, Yang F G, Fu C H. Application of random forest method in maize-soybean accurate identification. Acta Agronomica Sinica, 2018; 44(4): 569–580. (in Chinese)
[10] Song X P, Potapov P V, Krylov A, King L A, Di Bella C M, Hudson A, et al. National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey. Remote Sensing of Environment, 2017; 190: 383–395.
[11] Bocco M, Ovando G, Sayago S, Willington E, Heredia S. Estimating soybean ground cover from satellite images using neural-networks models. International Journal of Remote Sensing, 2012; 33(6): 1717–1728.
[12] Willington E, Clemente J P, Bocco M. Determination of agricultural land use: incidence of atmospheric corrections and the implementation in multi-sensor and multi-temporal images. Revista de Teledetección, 2015; 44: 81–89.
[13] Hu Q, Ma Y X, Xu B D, Song Q, Tang H J, Wu W B. Estimating sub-pixel soybean fraction from time-series MODIS data using an optimized geographically weighted regression model. Remote Sensing, 2018; 10(4): 491.
[14] Wardlow B D, Egbert S L, Kastens J H. Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U. S. Central Great Plains. Remote Sensing of Environment, 2007; 108(3): 290–310.
[15] Zhong L H, Yu L, Li X C, Hu L N, Gong P. Rapid corn and soybean mapping in US Corn Belt and neighboring areas. Scientific Reports, 2016; 6: 36240.
[16] Hu Q, Wu W B, Song Q, Lu M, Chen D, Yu Q Y, et al. How do temporal and spectral features matter in crop classification in Heilongjiang Province, China? Journal of Integrative Agriculture, 2017; 16(2): 324–336.
[17] Ajadi O A, Barr J, Liang S Z, Ferreira R, Kumpatla S P, Patel R, et al. Large-scale crop type and crop area mapping across Brazil using synthetic aperture radar and optical imagery. International Journal of Applied Earth Observation and Geoinformation, 2021; 97: 102294.
[18] Arvor D, Jonathan M, Penello Meirelles M S, Dubreuil V, Durieux L. Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil. International Journal of Remote Sensing, 2011; 32: 7847–7871.
[19] Brown J C, Kastens J H, Coutinho A C, Victoria D D C, Bishop C R. Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data. Remote Sensing of Environment, 2013; 130: 39–50.
[20] Zhong L H, Hu L N, Yu L, Gong P, Biging G. Automated mapping of soybean and corn using phenology. ISPRS Journal of Photogrammetry and Remote Sensing, 2016; 119: 151–164.
[21] Kuchler P C, Bégué A, Simões M, Gaetano R, Arvor D, Ferraz R P D. Assessing the optimal preprocessing steps of MODIS time series to map cropping systems in Mato Grosso, Brazil. International Journal of Applied Earth Observation and Geoinformation, 2020; 92: 102150.
[22] Chen Y L, Lu D S, Moran E, Batistella M, Dutra L V, Sanches I D, da Silva R F B, Huang J F, et al. Mapping croplands, cropping patterns, and crop types using MODIS time series data. Internatinal Journal of Applied Earth Observation and Geoinformation, 2018; 69: 133–147.
[23] Picoli M C A, Camara G, Sanches I, Simões R, Carvalho A, Maciel A, et al. Big earth observation time series analysis for monitoring Brazilian agriculture. ISPRS Journal of Photogrammetry and Remote Sensing, 2018; 145(Part B): 328–339.
[24] Formaggio A R, Vieira M A, Rennó C D. Object based image analysis (OBIA) and data mining (DM) in Landsat time series for mapping soybean in intensive agricultural regions. In: 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich: IEEE, 2012; 2257-2260. doi: 10.1109/IGARSS.2012.6351047.
[25] da Silva Junior C A, Leonel-Junior A H S, Rossi F S, Filho W L F C, de Barros Santiago D, de Oliveira-Júnior J F, et al. Mapping soybean planting area in midwest Brazil with remotely sensed images and phenology-based algorithm using the Google Earth Engine platform. Computers and Electronics in Agriculture, 2020; 169: 105194.
[26] Paludo A, Becker W R, Richetti J, De Albuquerque Silva L C, Johann J A. Mapping summer soybean and corn with remote sensing on Google Earth Engine cloud computing in Parana state - Brazil. International Journal of Digital Earth, 2020; 13(12): 1624–1636.
[27] Boryan C G, Yang Z W. Implementation of a new automatic stratification method using geospatial cropland data layers in NASS area frame construction. In: 2014 IEEE Geoscience & Remote Sensing Symposium, Quebec City: IEEE, 2014; pp.2110–2113. doi: 10.1109/IGARSS.2014.6946882.
[28] Fisette T, Davidson A, Daneshfar B, Rollin P, Aly Z, Campbell L. Annual space-based crop inventory for Canada: 2009–2014. In: 2014 IEEE Geoscience & Remote Sensing Symposium, Quebec City: IEEE, 2014; 5095-5098. doi: 10.1109/IGARSS.2014.6947643.
[29] Defourny P, Bontemps S, Bellemans N, Cara C, Dedieu G, Guzzonato E, et al. Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. Remote Sensing of Environment, 2019; 221: 551–568.
[30] Liu X, X Yu L, Zhong L H, Hao P Y, Wu B, Wang H S, et al. Spatial-temporal patterns of features selected using Random Forests: a case study of corn and soybeans mapping in the US. International Journal of Remote Sensing, 2018; 40(1): 269–283.
[31] Wang S, Di Tommaso S, Deines J M, Lobell D B. Mapping twenty years of corn and soybean across the US Midwest using the Landsat archive. Scientific Data, 2020; 7(1): 307.
[32] De Abelleyra D, Veron S, Banchero S, Mosciaro M J, Propato T, Ferraina A, et al. First large extent and high resolution cropland and crop type map of Argentina. In: 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS), Santiago: IEEE, 2020; 392-396. doi: 10.1109/LAGIRS48042.2020.9165610.
[33] Junior C A D S, Lima M, Johann J A, Rossi F S, de Oliveira-Júnior J F, Junior A S H L, et al. SojaMaps: project of monitoring of soybean areas in Brazil using big data in cloud computing. In: XIX Simpósio Brasileiro de Sensoriamento Remoto, 2019.
[34] Immitzer M, Vuolo F, Atzberger C. First experience with Sentinel-2 data for crop and tree species classifications in Central Europe. Remote Sensing, 2016; 8(3): 166.
[35] Veloso A, Mermoz S, Bouvet A, Le Toan T, Planells M, Dejoux J F, et al. Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. Remote Sensing of Environment, 2017; 199: 415–426.
[36] You N S, Dong J W, Huang J X, Du G M, Zhang G L, He Y L, et al. The 10-m crop type maps in Northeast China during 2017-2019. Scientific Data, 2021; 8: 41.
[37] Luo C, Liu H J, Lu L P, Liu Z R, Kong F C, Zhang X L. Monthly composites from Sentinel-1 and Sentinel-2 images for regional major crop mapping with Google Earth Engine. Journal of Integrative Agriculture, 2021; 207(7): 1944–1957.
[38] Vuolo F, Neuwirth M, Immitzer M, Atzberger C, Ng W T. How much does multi-temporal Sentinel-2 data improve crop type classification? International Journal of Applied Earth Observation and Geoinformation, 2018; 72: 122–130.
[39] Luo C, Qi B S, Liu H J, Guo D, Lu L P, Fu Q, et al. Using time series Sentinel-1 images for object-oriented crop classification in Google Earth Engine. Remote Sensing, 2021; 13(4): 561.

[40] Luo J S, Ma X W, Chu Q F, Xie M, Cao Y J. Characterizing the up-to-date land-use and land-cover change in Xiong'an New Area from 2017 to 2020 using the multi-temporal Sentinel-2 images on Google Earth Engine. ISPRS International Journal of Geo-Information, 2021; 10(7): 464.
[41] Song Q, Hu Q, Zhou Q B, Hovis C, Xiang M T, Tang H J, et al. In-season crop mapping with GF-1/WFV data by combining object-based image analysis and Random Forest. Remote Sensing, 2017; 9(11): 1184.
[42] Amani M, Kakooei M, Moghimi A, Ghorbanian A, Ranjgar B, Mahdavi S, et al. Application of Google Earth Engine cloud computing platform, Sentinel Imagery, and neural networks for crop mapping in Canada. Remote Sensing, 2020; 12(21): 3561.
[43] Kumari M, Pandey V, Choudhary K K, Murthy C S. Object-based machine learning approach for soybean mapping using temporal sentinel-1/sentinel-2 data. Geocarto International, 2021; 37(23): 6848–6866.
[44] Oldoni L V, Prudente V H R, Diniz J M F S, Wiederkehr N C, Sanches I D, Gama F F. Polarimetric Sar data from Sentinel-1a applied to early crop classification. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2020; XLIII-B3-2020: 1039–1046. doi: 10.5194/isprs-archives-XLIII-B3-2020-1039-2020.
[45] Yin L K, You N S, Zhang G L, Huang J X, Dong J W. Optimizing feature selection of individual crop types for improved crop mapping. Remote Sensing, 2020; 12(1): 162.
[46] Buenos Aires Grain Exchange. 2019. Available: http://www.bolsadecereales.com/ver-acerca-del-panorama-agricola-semanal-78#. Accessed on [2024-09-04].
[47] Ministry of Agriculture, Livestock and Fisheries | Presidency of the Nation, Argentina. Available: http://datosestimaciones.magyp.gob.ar/. Accessed on [2024-09-04].
[48] Emelyanova I, Barron O, Alaibakhsh M. A comparative evaluation of arid inflow dependent vegetation maps derived from LANDSAT top-of-atmosphere and surface reflectances. International Journal of Remote Sensing, 2018; 39(20): 6607–6630.
[49] Song C H, Woodcock C E, Seto K C, Lenney M P, Macomber S A. Classification and change detection using Landsat TM data: When and how to correct atmospheric effects. Remote Sensing of Environment, 2001; 75(2): 230–244.
[50] Chen Y S, Hou J L, Huang C L, Zhang Y, Li X H. Mapping maize area in heterogeneous agricultural landscape with multi-temporal Sentinel-1 and Sentinel-2 images based on Random Forest. Remote Sensing, 2021; 13(15): 2988.
[51] Hao P Y, Tang H J, Chen Z X, Liu Z J. Early-season crop mapping using improved artificial immune network (IAIN) and Sentinel data. PeerJ, 2018; 6: e5431.
[52] You N S, Dong J W. Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth Engine. ISPRS Journal of Photogrammetry and Remote Sensing, 2020; 161: 109–123.
[53] GlobeLand30: Global Geo-information Public Product, 2020. Available: https://www.webmap.cn/commres.do?method=globeIndex. Accessed on [2024-09-04].
[54] National Institute of Agricultural Technology of Argentina. Available: http://www.geointa.inta.gob.ar/. Accessed on [2022-04-23].
[55] Rumelhart D E, Hinton G E. Learning representations by back-propagating errors. Nature, 1986; 323: 533–536.
[56] Hermes L, Frieauff D, Puzicha J, Buhmann J M. Support vector machines for land usage classification in Landsat TM imagery. In: IEEE 1999 International Geoscience and Remote Sensing Symposium (IGARSS'99), 1999; 1: 348–350.
[57] Teluguntla P, Thenkabail P, Oliphant A, Xiong J, Gumma M K, Congalton R G, et al. A 30-m landsat-derived cropland extent product of Australia and China using Random Forest machine learning algorithm on Google Earth Engine cloud computing platform. ISPRS Journal of Photogrammetry and Remote Sensing, 2018; 144: 325–340.
[58] Tatsumi K, Yamashiki Y, Torres M A C, Taipe C L R. Crop classification of upland fields using Random forest of time-series Landsat 7 ETM+ data. Computers and Electronics in Agriculture, 2015; 115: 171–179.
[59] Rodriguez-Galiano V F, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez J P. An assessment of the effectiveness of a Random Forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 2012; 67: 93–104.
[60] Pelletier C, Valero S, Inglada J, Champion N, Dedieu G. Assessing the robustness of Random Forests to map land cover with high-resolution satellite image time series over large areas. Remote Sensing of Environment, 2016; 187: 156–168.
[61] Müller H, Rufin P, Griffiths P, Siqueira A J B, Hostert P. Mining dense Landsat time series for separating cropland and pasture in a heterogeneous Brazilian savanna landscape. Remote Sensing of Environment, 2015; 156: 490–499.
[62] Dong J, Fu Y Y, Wang J J, Tian H F, Fu S, Niu Z, et al. Early-season mapping of winter wheat in China based on Landsat and Sentinel images. Earth System Science Data, 2020; 12(4): 3081–3095.
[63] Fan D D, Li Q Z, Wang H Y, Zhang Y, Du X, Shen Y. Improvement in recognition accuracy of minority crops by resampling of imbalanced training datasets of remote sensing. National Remote Sensing Bulletin, 2019; 23(4): 730–742. (in Chinese)
[64] Zhong L H, Peng G, Biging G S. Efficient corn and soybean mapping with temporal extendability: A multi-year experiment using Landsat imagery. Remote Sensing of Environment, 2014; 140: 1–13.
[65] Cai Y P, Guan K Y, Peng J, Wang S W, Seifert C, Wardlow B, et al. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sensing of Environment, 2018; 210: 35–47.
[66] Löw F, Michel U, Dech S, Conrad C. Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using Support Vector Machines. ISPRS Journal of Photogrammetry and Remote Sensing, 2013; 85: 102–119.
[67] Husak G J, Marshall M T, Michaelsen J, Pedreros D, Funk C, Galu G. Crop area estimation using high and medium resolution satellite imagery in areas with complex topography. Journal of Geophysical Research, 2008; 113(D14): D14112.