Published June 6, 2019

A Spatialization Method for Grain Yield Statistical Data: A Study on Winter Wheat of Shandong Province, China

Guofeng Xiao, Xiufang Zhu,* Chenyao Hou, Ying Liu, and Kun Xu

ABSTRACT

Grain yield data based on administrative divisions (counties, cities, etc.) for statistics lack spatial information, which can be effectively solved by grain yield spatialization. This paper proposes a spatialization method for grain yield based on the Moderate Resolution Imaging Spectroradiometer (MODIS) Normalized Difference Vegetation Index (NDVI) time series data. The method was tested by taking winter wheat (*Triticum aestivum* L.) in Shandong Province in China as an example. First, the classification and regression tree (CART) algorithm was trained to extract the winter wheat planting pixels in 2016. The average NDVIs of the different growing stages (returning green, jointing, heading, and milk ripening) were calculated from the MODIS NDVI time series data. The relationship between winter wheat yield and NDVI variables (including single-phase NDVI and the average NDVI of different growing stages) was analyzed by univariate and multiple linear regressions. The NDVI variable with the highest correlation to winter wheat yield and the minimum root mean square error of the fitting equation were chosen as input to build the spatialization model. The results show that the classification accuracy of winter wheat estimated with the confusion matrix was 82.51% and that the average precision of planting acreage compared with county-level statistical data was 87.64%. The average relative error of yield spatialization at the county level was 22.71%. The method developed in this paper is easy to operate and popularize, and it can provide a technical reference for producing high-resolution crop yield distribution maps of long time series through spatialization.

Core Ideas

- · Dividing the study area into subregions and classifying them improved winter wheat classification accuracy.
- · Single-phase Normalized Difference Vegetation Indices acquired on 6 March, 23 April, 25 May, and 29 June were the best variables for building the wheat yield spatialization model.
- · The extraction accuracy of winter wheat area greatly impacted the spatialization of yield.
- · The proposed model can provide a technical reference for producing high-resolution crop yield distribution maps.

Published in Agron. J. 111:1892–1903 (2019)
doi:10.2134/agronj2018.09.0555

Agriculture is fundamental for human society. Agricultural production statistics are of paramount importance for societal, economic, agricultural, and policy concerns (Carletto et al., 2015). Therefore, governments are committed to collecting data on agricultural production to understand agricultural development and assist in formulating agricultural development policies. However, agricultural production statistics are reported on a geopolitical basis, such as by country, province, or city and are generally available in tabular form, which cannot provide fine-scale distribution information within geopolitical units (You and Wood, 2006). Fine-scale distribution information on agricultural production is valuable for a wide range of applications, such as field management, crop yield gaps, and agricultural insurance (Lobell, 2013). The spatialization of tabular crop production statistics from geopolitical units to subregions or even individual pixels is an effective method for producing fine-scale distribution information on crop production within geopolitical units (You and Wood, 2006).

The concept of 'spatialization' was proposed in the early 1990s (Tobler et al., 1995, 1997). It uses certain methods or parameters to build a model, then the model is applied to describe the distribution of data over a certain time and space scale. Over the past few decades, the spatialization of social and economic statistics has become a hot topic in many disciplines. The methods used for spatializing statistical data can be generally divided into four categories: spatial interpolation models (Huffman et al., 2007; Vicente-Serrano et al., 2003; Stahl et al., 2006), spatial allocation models based on land use and land cover data (Matlock et al., 1996; Yang et al., 2009), multisource data fusion models (Sutton et al., 2001; Wu et al., 2006; Azar et al., 2013), and remote sensing inversion models (Elvidge et al., 1997; Ghosh et al., 2010; Li et al., 2013).

Currently, the spatialization of statistical data mainly focuses on population data (Liu et al., 2008; Fisher and Langford, 1995; Mennis, 2003) and gross domestic product data (Yue et al.,

G. Xiao, X. Zhu, State Key Lab. of Earth Surface Processes and Resource Ecology, Beijing Normal Univ., Beijing 100875, China; X. Zhu, C. Hou, Key Lab. of Environmental Change and Natural Disasters of the Ministry of Education, Faculty of Geographical Science, Beijing Normal Univ., Beijing 100875, China; Y. Liu, K. Xu, Beijing Engineering Research Center for Global Land Remote Sensing Products, Institute of Remote Sensing Science and Engineering, Faculty of Geographical Science, Beijing Normal Univ., Beijing 100875, China. Received 3 Sept. 2018. Accepted 22 Feb. 2019. *Corresponding author (zhuxiufang@bnu.edu.cn).

Abbreviations: CART, classification and regression tree; GF, Gaofen remote sensing satellite; MODIS, Moderate Resolution Imaging Spectroradiometer; NDVI, Normalized Difference Vegetation Index; RMSE, root mean square error; SR, subregion.

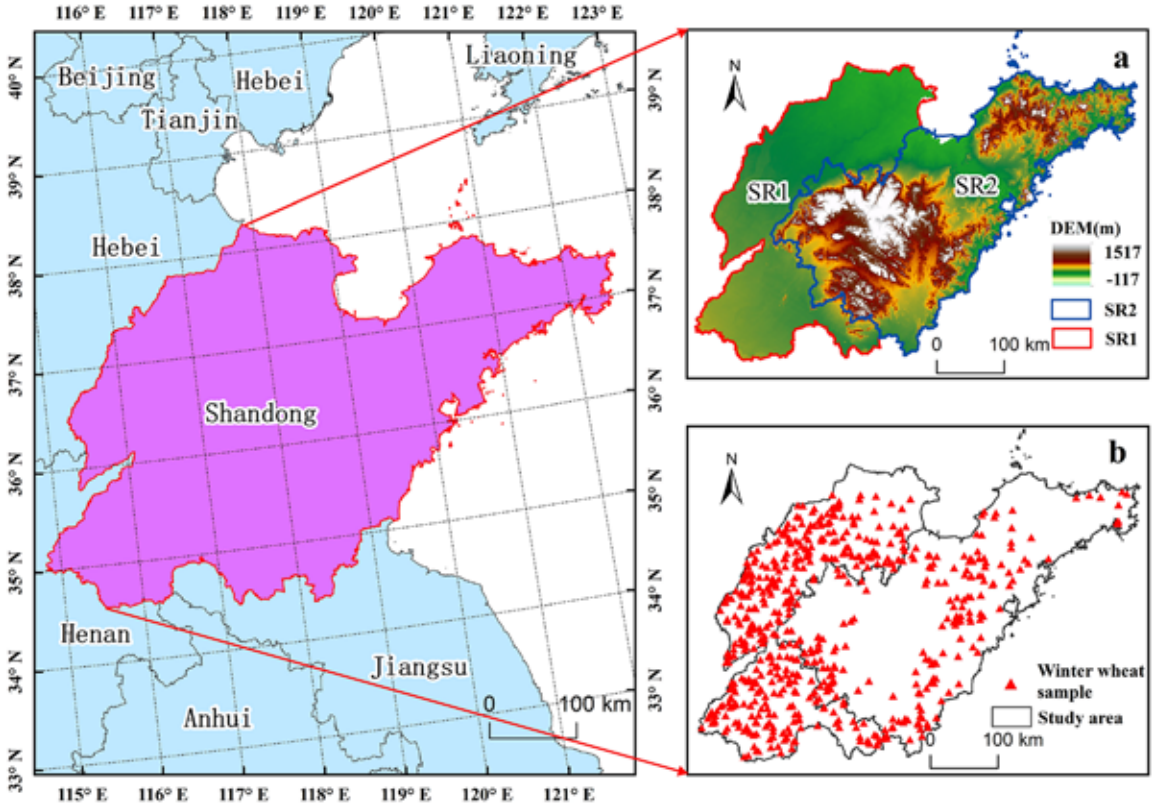

Fig. 1. Study area location. (a) Distribution of Subregion 1 (SR1, plains) and Subregion 2 (SR2, hilly region). (b) Distribution map of the winter wheat sample.

2014; Elvidge et al., 2001; Henderson et al., 2003). There is little research on the spatialization of agricultural economic statistics, because agricultural production activities are affected by many factors, including geographical location, climatic characteristics, and soil hydrology; therefore, it is difficult to spatialize. Of those that have considered the spatialization of agricultural economic statistics, researchers have paid more attention to the spatialization of crop planting areas (You and Wood, 2005; You et al., 2009; Khan et al., 2010; Monfreda et al., 2008) and the spatialization of agricultural production inputs (Potter et al., 2010; Zhu et al., 2012); the spatialization of grain yield statistical data is rare. On the basis of a 1-km resolution population density map, Liu and Li (2012) built a regression model using the population density as the dependent variable and the grain yield per unit of arable land as the independent variable. The model was used to produce a grain output map for China, with a 1-km spatial resolution, in 2000. On the basis of this 1-km resolution land cover map, Ji et al. (2015) developed the relationship between the acreages of different farmland types and grain yield to spatialize the grain yield of China in 2005. Their study focused on the total grain yield rather than the yield of a given crop type. However, a spatial distribution map of the yield of different crop types is more useful than a spatial distribution map of the total grain yield for research on agriculture activities.

Remote sensing technology has become one of the methods used for studying regional spatial patterns of crops and their dynamic changes due to its large-scale, high efficiency, and rapidity (Lei et al., 2012). Moderate Resolution Imaging Spectroradiometer (MODIS) NDVI is a very popular data source for crop distribution maps and crop yield estimations for large-scale studies (Fritz et al., 2008; Potgieter et al., 2013). Lots of studies have explored the empirical relationship between NDVI and crop yield (Mkhabela et al., 2005; Huang et al., 2014; Mashaba et al., 2017). However, different studies have used different NDVI variables, such as original NDVI, average or accumulated NDVI over the growth period, average or accumulated NDVI over key growing stages, etc.

This paper takes winter wheat in Shandong Province as an example, analyzes the relationship between NDVI variables with winter wheat yield, and chooses the best NDVI variables to build a spatialization model of winter wheat yield and make a yield distribution map of winter wheat with 250-m resolution. The remote sensing data used in this paper have the advantages of global coverage, free access, high spatial resolution (250 m), and a long coverage time (2000–2019). The method developed in this paper is easy to operate and popularize, and it can provide a technical reference for producing high-resolution (up to the 250-m pixel level) crop yield distribution maps of long time series through spatialization. The resulting high-resolution crop yield distribution map of long time series can help us to analyze the temporal and spatial changes of grain yield, and provide basic data for the agricultural insurance industry, agricultural planting systems, and agricultural disaster assessment.

STUDY AREA AND DATA

Study Area

Shandong Province is located on the east coast of China and the lower reaches of the Yellow River, ranging from 34°22.9′ to 38°24.01′N and 114°47.5′ to 122°42.3′E (Fig. 1). It has a land area of 155,800 km^2, 17 cities, and 137 counties. Shandong Province belongs to the warm temperate monsoon climate type. Its rainfall is concentrated within a short period; rain and heat occur within

Table 1. Data details and sources.

Data name	Year	Spatial resolution	Time resolution	Source	Application
Moderate Resolution Imaging Spectroradiometer normalized difference vegetation index	2016	250 m	16 d	USGS†	Extracting the planting area and building the yield spatialization model for winter wheat
Digital Elevation Model	–	90 m	–	USGS	Subdividing the study area
Winter wheat planting area statistics in Shandong Province	2016	–	–	Shandong Provincial Bureau of Statistics‡	Verifying the accuracy of extracted winter wheat area
Winter wheat yield statistics in Shandong Province	2016	–	–	Shandong Provincial Bureau of Statistics	Targeted data for spatialization
Landsat-8 Operational Land Imager	2016	30 m	–	USGS	Selection of winter wheat samples
Gaofen-1	2016	16 m	–	Geospatial data cloud§	Selection of winter wheat samples
Gaofen-2	2016	4 m	–	Geospatial data cloud	Selection of winter wheat samples
Google Earth imagery	2016	–	–	Google Earth software, Google, Santa Clara, CA,	Selection of winter wheat samples

† US Geological Survey (USGS) (http://glovis.usgs.gov/, accessed 10 May 2019).
‡ http://xxgk.stats-sd.gov.cn/ (accessed 10 May 2019)
§ http://www.gscloud.cn/ (accessed 10 May 2019).

the same period. The frost-free period is increasing from the northeast coast to the southwest. Its light resources are abundant and the heat conditions can meet the needs of two crops per year. Shandong is one of the major food production provinces in China. The grain crops are divided into summer grain and autumn grain. The summer grain is mainly winter wheat; the autumn grain is mainly corn (*Zea mays* L.), potato (*Solanum tuberosum* L.), soybean [*Glycine max* (L.) Merr.], rice (*Oryza sativa* L.), millet [*Setaria italica* (L.) P. Beauv.], sorghum [*Sorghum bicolor* (L.) Moench.], and small grains. Among these, wheat, corn, and potato are the three major food crops in Shandong.

Datasets

The data used in this paper include: (i) the MODIS-NDVI (MOD13Q1) time-series data, (ii) the Digital Elevation Model, (iii) winter wheat planting area statistics for Shandong Province, (iv) winter wheat yield statistics for Shandong Province, and (v) Landsat-8 Operational Land Imager, Gaofen (GF) 1, GF-2, and Google Earth imagery. The details of the data, including the sources, are shown in Table 1.

METHODS

The spatialization process of the winter wheat yield, based on MODIS-NDVI time series data, is shown in Fig. 2. It mainly includes: (i) preprocessing of MODIS-NDVI time series data, the division of the research area and examination of winter wheat statistical data on planting area and yield; (ii) selection of winter wheat samples, based on high-resolution remote sensing data, and extracting winter wheat planting area, based on the classification and regression tree (CART) algorithm; (iii) calculation of the average NDVI during different winter wheat growing stages and analysis of the relationship between the NDVI variables and winter wheat yield; (iv) building the spatial model for winter wheat yield spatialization; and (v) verifying the accuracy of winter wheat yield spatialization.

Data Preprocessing

First, MRT software (https://modis.gsfc.nasa.gov/tools/, accessed 14 May 2019) was used to convert the MODIS-NDVI time-series data in 2016 to Albers equal-area projections, with a resampling resolution of 250 m. The Savitzky-Golay filter method developed by Chen et al. (2004) was used to eliminate noise and smooth the MODIS-NDVI time series data. The MODIS-NDVI data during the winter wheat growing season in 2016 in Shandong Province were extracted for classification. Based on the phenological period information for winter wheat in Shandong Province, the average NDVI of the different growing stages (returning green, jointing, heading, and milk ripening stages) were calculated.

Considering the topographic characteristics of Shandong Province and the complexity of winter wheat planting plots, the study area was divided into a plain–simple landform region [referred to as Subregion (SR) 1] and a plain–complex hilly landform region (SR2) based on the Digital Elevation Model and administrative boundary data(Fig. 1a). The training and validation samples of winter wheat were selected by referring to high-resolution remote sensing data (Landsat-8 Operational Land Imager, GF-1, GF-2, etc.) (Fig. 1b). We examined the completeness and regularity of the data on winter wheat acreage and yield. Spatial linking of the statistical data and vector boundary data ensured there were statistical data on the winter wheat area and yield in each region.

Classification with the CART Algorithm

The CART algorithm is a decision tree construction algorithm first proposed by Breiman et al. (1984). It is based on two recursive division segmentation techniques that divide the sample set into two subsets, giving two branches to each nonleaf node of the decision tree. The decision tree generated by the CART algorithm is a simple binary tree and thus there can only be a 'yes' or 'no' answer to every step.

The CART algorithm uses the Gini coefficient (Gini index) in economics as the criterion for selecting the best test variables. The selection criterion is that each subnode must achieve the highest purity or, in other words, all the elements of the subnodes must belong to the same category. Assuming the dataset S, there are n class categories $\{C_1, C_2,..., Cn\}$ in total, and each class is a sample subset $C_i = S_i$ $(1 \le i \le n)$. $|S|$ is the number of samples in the sample set, $|Ci|$ is the number of subsets Ci in the sample set S, $\rho_i = |C_i|/|S|$ is the probability that the sample in the sample set belongs to class C_i, and the Gini coefficient of dataset S can be expressed as:

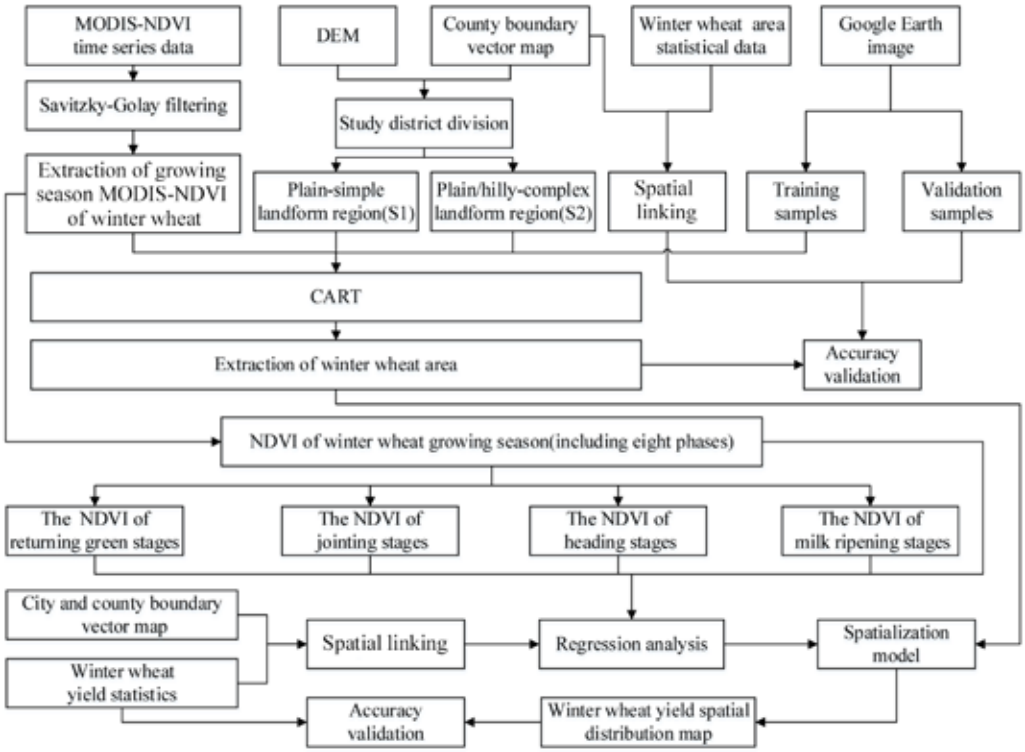

Fig. 2. Technical flowchart.

$$Gini(S) = 1 - \sum_{i=1}^{n} \rho_i^2 \,. \qquad [1]$$

Based on high-resolution remote sensing data and Google Earth imagery, training samples were selected for the two subregions. Based on the MODIS-NDVI data of the growing season of winter wheat, the CART classification method was used to establish the decision tree automatically for winter wheat planting area identification, and the winter wheat planting areas in SR1 and SR2 were extracted. The classification results of the two subregions were merged into the distribution map of winter wheat in Shandong Province.

The classified winter wheat planting area of each county was evaluated by using the county boundary map and was compared with the statistical data on the winter wheat planting area. The accuracy of the classified planting area (the absolute value of the difference between the statistical acreage and the classified acreage divided by the statistical acreage) in each county was calculated; counties with a classification accuracy less than 60% were selected. The decision tree was re-established for the images of the selected counties to improve the classification accuracy.

Selection of Regression Factors

Existing studies have shown that the winter wheat yield is closely related to the NDVI during the growing season (Mkhabela et al., 2011; Hansen and Schjoerring, 2003). Therefore, we extracted eight 16-d MODIS-NDVI data periods during the winter wheat growing season in Shandong Province, the starting dates of which were 6 March 6, 22 March, 7 April, 23 April, 9 May, 25 May, 10 June, and 26 June. The abbreviation $NDVI_{9May}$ represents the 16-d MODIS-NDVI during 9 to 24 May. By combining the winter wheat phenology information from Shandong Province, we calculated the average NDVI of the returning green stage (the mean of $NDVI_{6Mar}$ and $NDVI_{22Mar}$), jointing stage (the mean of $NDVI_{7Apr}$ and $NDVI_{23Apr}$), heading stage (the mean of $NDVI_{9May}$ and $NDVI_{25May}$), and the milk ripening stage (the mean of $NDVI_{10June}$ and $NDVI_{26June}$). The eight single-phase NDVIs and the NDVIs of the four phenological stages in the winter wheat growing season were aggregated at county level as input variables.

In SPSS software (version 25.0, IBM Corp., Armonk, NY), univariate linear regression was established between each NDVI variable and the winter wheat yield and the correlation coefficient was determined and tested for significance. Multivariate linear regressions between NDVI variables and winter wheat yield were established from the independent variable to determine the number of input variables and correlation coefficients. The equations for the univariate linear regression and multiple regression are as follows:

$$Y = aX_i + b \,, \qquad [2]$$

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \,, \qquad [3]$$

where Y represents the winter wheat yield, X_i represents the NDVI of the different phases, and a, b, and $\beta_0 \ldots \beta_n$ are constants. Finally, the correlation coefficient between each input variable and the winter wheat yield, the regression's standardized residual histogram, and the minimum root mean square error (RMSE) of the fitting equation were analyzed. The NDVI variable with the highest correlation coefficient for winter wheat yield and the minimum RMSE of the fitting equation were selected as the spatialization factors of winter wheat yield.

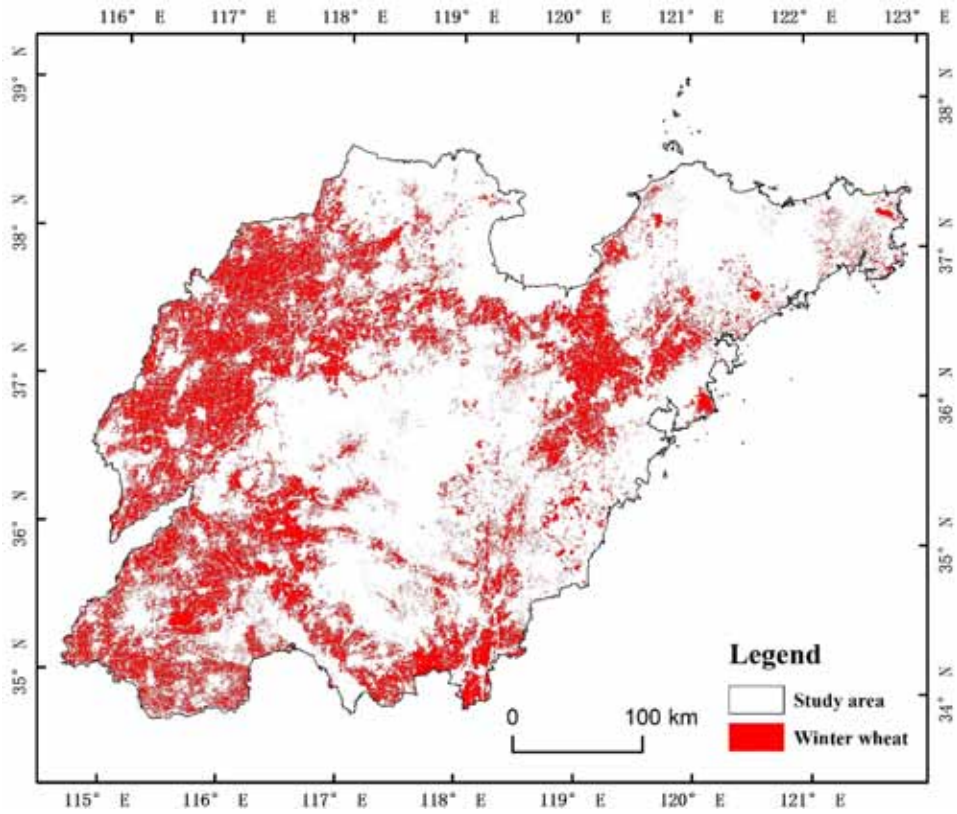

Fig. 3. Distribution map of winter wheat planting area.

Building the Spatial Model

According to the analysis results in the Selection of Regression Factors section above, the NDVI of the four single-phase stages were chosen as the optimal input variables for building the spatialization model of winter wheat yield (details on the analyses are given in the Spatialization Model Factors section). For a given county, the spatial model is as follows:

$$Y_j' = Y \times \sum_{i=1}^{4} \left(\frac{NDVI_{i,j}}{\sum_{j=1}^{n} NDVI_{i,j}} \times \alpha_i \right); \quad [4]$$

$$\alpha_i = p_i \Big/ \sum_{i=1}^{4} p_i, \quad [5]$$

where Y_j' represents the spatialized yield of winter wheat pixel j ($j = 1,...n$), n is the number of pixels classified as winter wheat in the given county, Y represents the statistical winter wheat yield of the given county, $NDVI_{i,j}$ represents the NDVI of the i^{th} stage, i (= 1, 2, 3, 4) represents the four single-phase stages, $\sum_{j=1}^{n} NDVI_{i,j}$ represents the sum of the NDVIs of all winter wheat pixels in a given county in the i^{th} stage, p_i represents the correlation coefficient between the NDVI and the statistical yield in the i^{th} stage, and α_i represents the correlation coefficient normalization result for the i^{th} stage.

Validation of the Accuracy

In this paper, validation of the accuracy included two steps: verifying the planting area classification and verifying the yield spatialization results. The verification of the planting area classification was performed by the standard confusion matrix method, in which 595 validation samples of winter wheat were randomly selected in Shandong Province (Fig. 1) for calculating the confusion matrix. In total, 403 samples were located in SR1 and 192 samples are in SR2. The verification method for the spatialization results usually made a comparison between the fine-scale data (such as the township level) and the spatialized results from the data at a coarser scale (county level). For this study, there were no official statistical winter wheat yield data at township level in Shandong Province; therefore, we used indirect methods to verify the accuracy of our spatial model. The indirect validation method involved making 250-m resolution spatial distribution map of the winter wheat yield based on our spatialization model with winter wheat yield at the municipal level. The total winter wheat yield at county level was calculated from the distribution map and was compared with statistical data on the total yield of winter wheat at the county level.

RESULTS AND DISCUSSION

Winter Wheat Distribution

The winter wheat distribution map of Shandong Province is shown in Fig. 3. Winter wheat in Shandong Province is mainly distributed in the western plain area, the southwest plain area, and the flat terrain in the middle. The plains area, in which the main industry is agriculture and other primary industries, is main area for winter wheat planting. The terrain in the central hilly region is complex, the land is fragmented, and the natural conditions are poor, all of which are not conducive to the cultivation of winter wheat. The coastal areas are economically developed regions, where the main industries are in the secondary and tertiary sectors and the winter wheat acreage is small.

Table 2. Verification of the accuracy of winter wheat planting area.

Region	Winter wheat samples	Confusion matrix			Statistical data
		Producer accuracy	User accuracy	Overall accuracy	Average accuracy
	n	%			
Subregion 1†	403	88.63	77.84	86.03	92.88
Subregion 2	192	74.69	82.24	78.34	81.28
Subregion 1 + Subregion 2	595	83.26	78.91	82.51	87.64

† Subregion 1, plains region; Subregion 2, hilly region.

Two methods were used to verify the classification accuracy of the winter wheat planting area. One was the confusion matrix; the other was a comparison with statistical data (Table 2). It can be seen from Table 1 that the overall accuracy of classification in SR1 was higher than that in SR2. The omission error in SR1 is higher than that in SR2 and the commission error in SR2 is higher than that in SR1. The overall accuracy of identifying the winter wheat planting area in Shandong Province was 82.51%, the producer accuracy was 83.26%, and the user accuracy was 78.91%. In addition, in a comparison with the actual winter wheat planting area (the statistics), we found that the average accuracy of recognizing winter wheat planting area in SR1, SR2, and the whole province was 92.88, 81.28, and 87.64%, respectively.

Spatialization Model Factors

The correlation between different NDVI variables and winter wheat yield was obtained by univariate linear regression and multiple linear regression analysis (Table 3). From Table 3, we can see that in the univariate linear regression analysis, when the independent variable was a single-phase NDVI, the correlation between the $NDVI_{7Apr}$ and winter wheat yield was the highest ($R^2 = 0.901$, $P \le 0.05$). When the independent variable was the phenophase NDVI, the NDVI of the jointing stage had the highest correlation with winter wheat yield ($R^2 = 0.900$, $P \le 0.05$).

In the multiple linear regression analysis, when the input of the independent variables was the single-phase NDVI variables, the variables for the linear regression were selected as: $NDVI_{6Mar}$, $NDVI_{23Apr}$, $NDVI_{25May}$, and $NDVI_{26June}$. When the input of the independent variables was the phenophase NDVI variables, the variables of the linear regression were: returning green stage NDVI, jointing stage NDVI, heading stage NDVI, and milk ripening stage NDVI; the NDVIs of all four stages were used to build the linear regression model. When the input variables included both single-phase and phenophase NDVI variables, the variables in the linear regression were selected as: $NDVI_{6Mar}$, $NDVI_{23Apr}$, $NDVI_{25May}$, and milk ripening stage NDVI. All three multiple linear regression models passed the significance test ($R^2 = 0.903$, $P \le 0.05$).

To further select the variables involved in the spatialization of winter wheat yield, we analyzed the residuals of the univariate linear regression and the multiple linear regression; the regressions' standardized residual histogram is shown in Fig. 4, alongside the RMSE of the fitting equation. In the standardization residual histogram of the regression, the normal curve is a criterion for judging whether the standardized residual histogram conforms to a normal distribution. From Fig. 4, we can see that in the linear regression, the standardized residual histograms of both linear regression equations are in accordance with normal distribution. The SD of the standardized residual error histogram of the multiple linear regression was smaller and more consistent with the normal distribution, indicating that the multiple linear regression model is superior to the univariate linear regression model. In a comparison of the equation fitting effects of three multiple linear regressions, the RMSE of the fitted equation was calculated separately. When the independent variable was a single-phase NDVI variable, the RMSE was the smallest. The single-phase NDVI variables were: $NDVI_{6Mar}$, $NDVI_{23Apr}$, $NDVI_{25May}$, and $NDVI_{26June}$. Therefore, we

Table 3. Regression model correlation.

Methods		Independent†	R^2	Adjusted R^2
Univariate linear regression	Single-phase	$NDVI_{6Mar}$	0.864**	0.863
		$NDVI_{022Mar}$	0.894**	0.894
		$NDVI_{7Apr}$	0.901**	0.900
		$NDVI_{23Apr}$	0.897**	0.897
		$NDVI_{9May}$	0.888**	0.887
		$NDVI_{25May}$	0.872**	0.871
		$NDVI_{10June}$	0.856**	0.855
		$NDVI_{26June}$	0.861**	0.860
	Phenological stages	$NDVI_{rgs}$	0.885**	0.884
		$NDVI_{js}$	0.900**	0.899
		$NDVI_{hs}$	0.883**	0.882
		$NDVI_{mrs}$	0.861**	0.860
Multiple linear regression	Single-phase	$NDVI_{6Mar}$, $NDVI_{23Apr}$, $NDVI_{25May}$, $NDVI_{26June}$	0.903**	0.900
	Phenological stages	$NDVI_{rgs}$, $NDVI_{js}$, $NDVI_{hs}$, $NDVI_{mrs}$	0.903**	0.900
	Single-phase and phenological stages	$NDVI_{6Mar}$, $NDVI_{23Apr}$, $NDVI_{25May}$, $NDVI_{mrs}$	0.903**	0.900

**, Significant at the 0.01 probability level

† $NDVI_{rgs}$, NDVI during the returning green stage; $NDVI_{js}$, NDVI during the jointing stage; $NDVI_{hs}$, NDVI during the heading stage; $NDVI_{mrs}$, NDVI during the milk ripening stage.

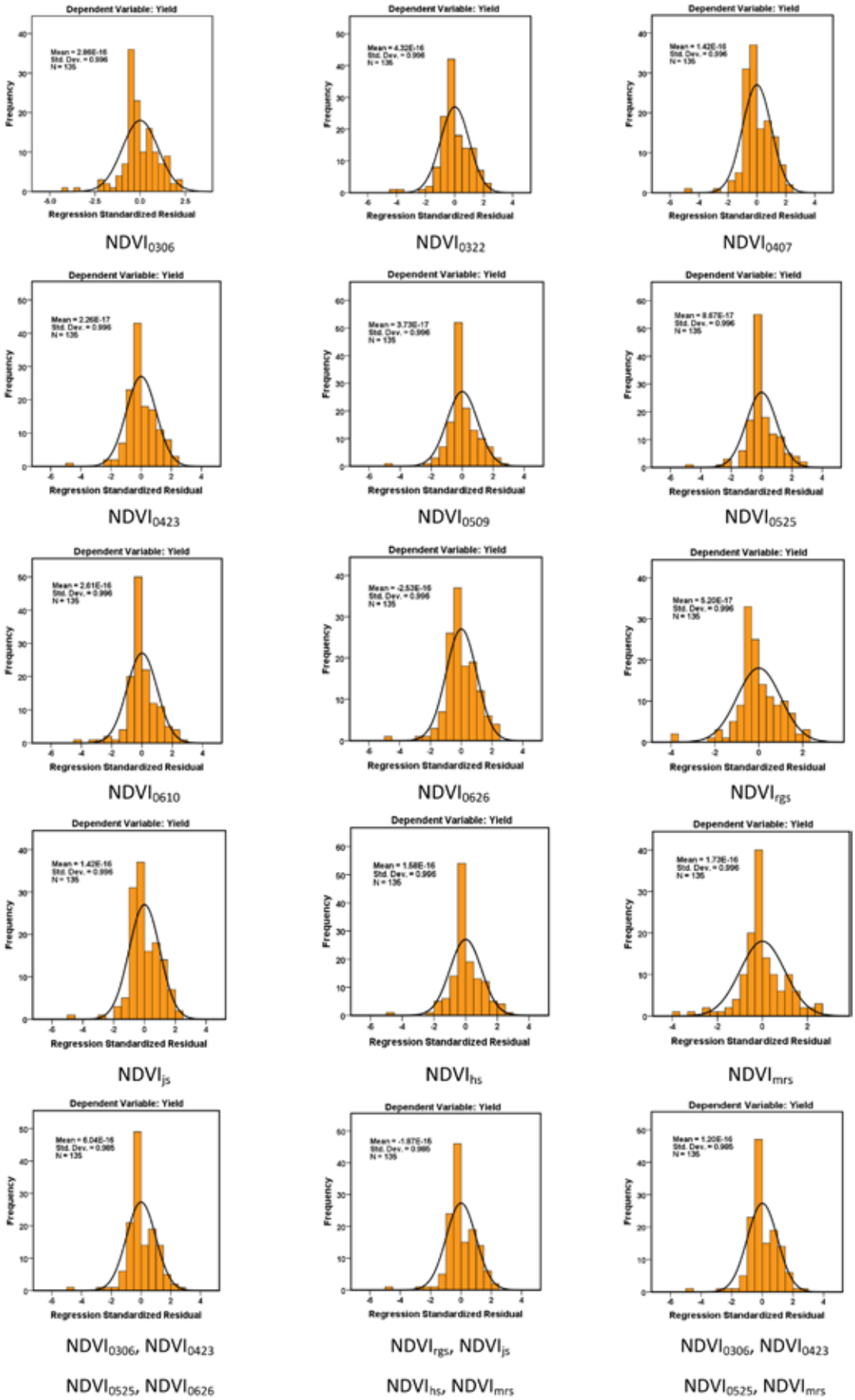

Fig. 4. Standardized residual histogram of the regressions.

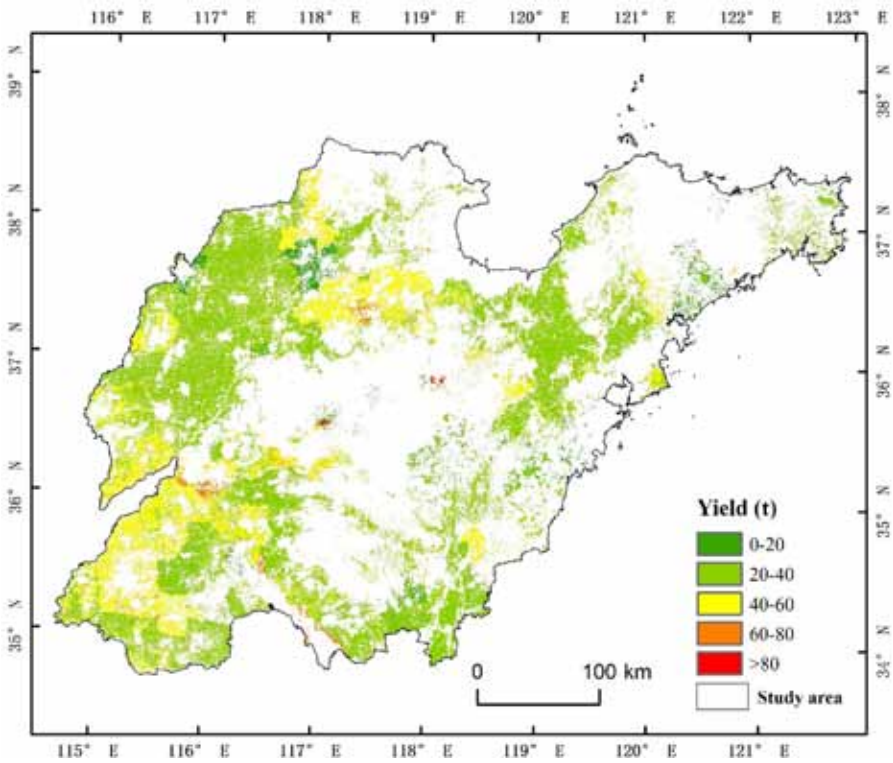

Fig. 5. Distribution map of winter wheat yield spatialized from county-level statistical data.

finally chose $NDVI_{6Mar}$, $NDVI_{23Apr}$, $NDVI_{25May}$, and $NDVI_{26June}$ to build the spatialization model.

Winter Wheat Yield Distribution

The 250-m resolution spatial distribution map of the winter wheat yield in Shandong Province is shown in Fig. 5. It can be seen from Fig. 5 that winter wheat yield was higher in the western and southwestern parts of Shandong Province and lower in the central hilly region and coastal fringe. The yield distribution generally shows a pattern of high in the west and low in the east.

Because of the lack of township-level winter wheat yield data, the model accuracy for the spatialization of winter wheat yield established at county level cannot be estimated. Thus we used indirect methods to verify the accuracy of the model. The spatialization model for winter wheat yield was based on municipal-level data and the spatial distribution map of winter wheat yield was obtained (Fig. 6). With the county-level winter wheat yield data as a reference, we calculated the relative error for each county. The average relative error was 22.71%. The scatter map of the spatialized yield and the statistical yield at county level is shown in Fig. 7. Most of the scatter points are close to the 1:1 line.

Uncertainty Analysis

Uncertainties in Extracting the Winter Wheat Planting Area

During the process of winter wheat planting area extraction, the remote sensing data pre-processing, the complexity of the types of land objects, the spatial resolution of the remote sensing images, and the classification method all influence the results of extracting the winter wheat planting area.

Data Preprocessing . The remote sensing data used in the experiment were MODIS product data (MOD13Q1), which comprise 16-d synthetic vegetation index products with a high time resolution. These can reflect the NDVI values of different phenological phases in the crop growing season and MODIS is one of the best data sources for extracting crop areas. In this experiment, the Savitzky–Golay filter was used to smooth the MODIS NDVI data. It improved the smoothness of the spectrum and reduced noise interference. However, the filtered results were influenced by the size of the filter window and the fitting polynomial order. Choosing an appropriate filter window size and the polynomial fitting order are important for reducing the noise of the winter wheat phenological curve and enhancing the fitting degree of the phenophase curve.

Complexity of the Type of Land Objects. In this paper, the topography of the study area is mainly plains and hills. The plains are located in the western and southwestern parts of Shandong Province. This type of land is simple, the arable land is concentrated, and the accuracy of the extraction area of winter wheat is high. The hills in the central part of Shandong Province are undulating and the terrain is more complex, mostly in the mountains. Forestry is relatively well developed and the arable land is broken. The area of winter wheat is small and the classification accuracy of the winter wheat planting area is low. Through experiments, it has been found that the accuracy of winter wheat planting area in SR2 is lower than that in SR1. Through the introduction of an existing distribution map of farmland fields and the improvement of remote sensing image quality, the accuracy of extracted crop planting area in SR2 can be improved.

Spatial Resolution. The MODIS NDVI data used in this study had a spatial resolution of 250 m. The land objects in the hilly region are complex and the arable land is broken. The pixel mixing phenomenon is serious, aggravating classification errors. In future research, high spatial resolution remote sensing data (such as Landsat Thematic Mapper, GF-1, GF-2, etc.) can be

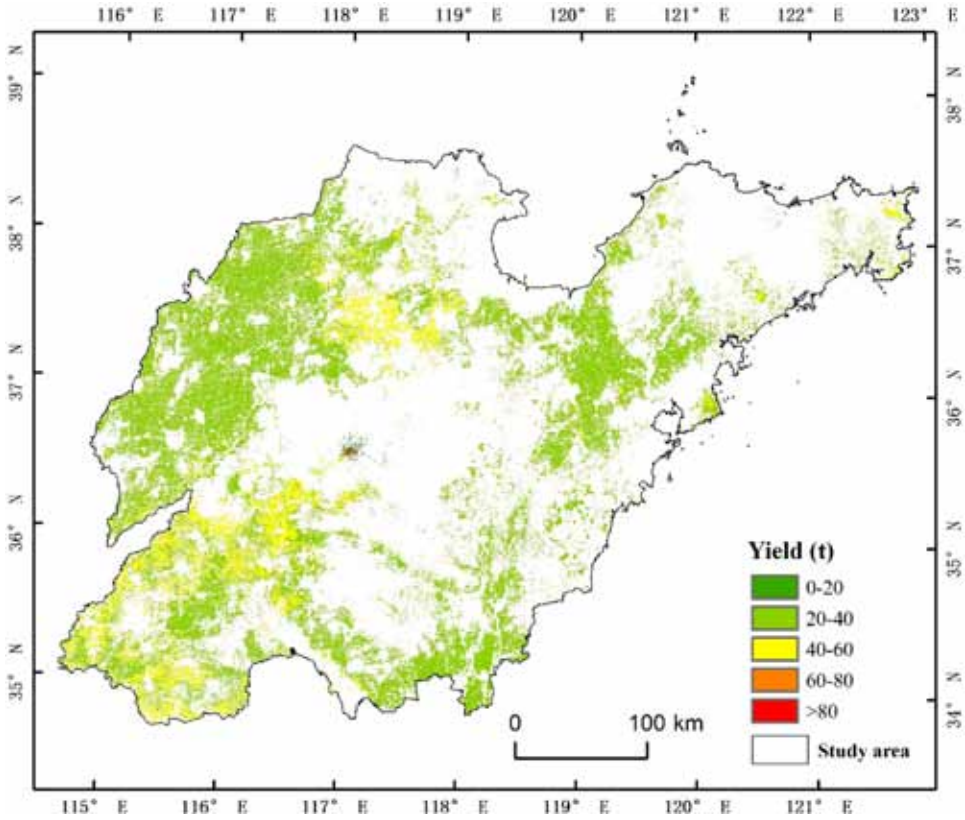

Fig. 6. Distribution map of winter wheat yield spatialized from municipal level statistical data.

introduced to reduce the problem of mixed pixels and improve the accuracy of extracting winter wheat planting area.

Method of Extracting Winter Wheat Planting Area. To improve the classification accuracy, researchers have proposed many different methods, such as an artificial intelligence neural networks, decision trees, support vector machines, and more. The decision tree classification method can make full use of the spectral features and other auxiliary information for the image and effectively solve the problem of different objects having the same spectrum and the same objects having a different spectrum. In the conventional decision tree method, a classification decision tree is built through empirical judgment, which is greatly influenced by the operator's experience and skills and is based on single-phase or multiphase remote sensing data, which only includes some information on the crop growing season. In this study, the MODIS NDVI data, covering the entire winter wheat growing season, were used in the classification to reduce the interference of other crop information. A large number of training samples were selected by high-resolution images and a decision tree was automatically established via the CART algorithm to avoid human interference. Besides this, we divided the study area into two subregions on the basis of factors, such as planting systems and the complexity of the land cover type, and established a decision tree respectively for each subregion to improve the classification accuracy.

Uncertainties about Winter Wheat Yield Spatialization

At present, the crop yield data were collected by administrative units without spatial location information. In this study, we built a spatialization model based on MODIS NDVI data and used it to spatialize the winter wheat statistical yield data to obtain a spatial distribution map of winter wheat yield. The spatialization model is influenced by various factors, as described below.

The Effect of Winter Wheat Area Extraction Accuracy on the Spatialization of the Yield. The classification of the winter wheat planting area is the basis for the spatialization of winter wheat yield. The accuracy of the winter wheat yield spatialization is impacted by the accuracy of planting area classification. The overall accuracy of the winter wheat planting area in SR2 is relatively low, mainly caused by omission errors. The classified planting area of winter wheat was lower than the officially reported statistical area, resulting in a higher average value for the yield in winter wheat pixels after spatialization. The user accuracy of the winter wheat planting area in SR1 is relatively low, showing more commission errors. In a comparison with the high-resolution images, it was found that the ridges between field plots were also classified as winter wheat planting areas. The classified winter wheat planting area was larger than the officially reported statistical area, resulting in a lower average yield in winter wheat pixels. In this study, we reclassified the counties where the winter wheat classification accuracy was lower than 60% to improve the overall accuracy. In future, more work should be done to improve the classification accuracy and ultimately increase the precision of the spatialization of winter wheat yield. Introducing a distribution map of farmland plot data and the spectral unmixing method might improve the classification of low-resolution images.

Variables Selected for the Spatialization of Winter Wheat Yield. The existing grain yield spatialization research (Liu and Li, 2012; Ji et al., 2015) built spatial models based on factors such as population density and acreage of different farmland types. In reality, the correlation between single crop yield, population density, and farmland area type is relatively low.

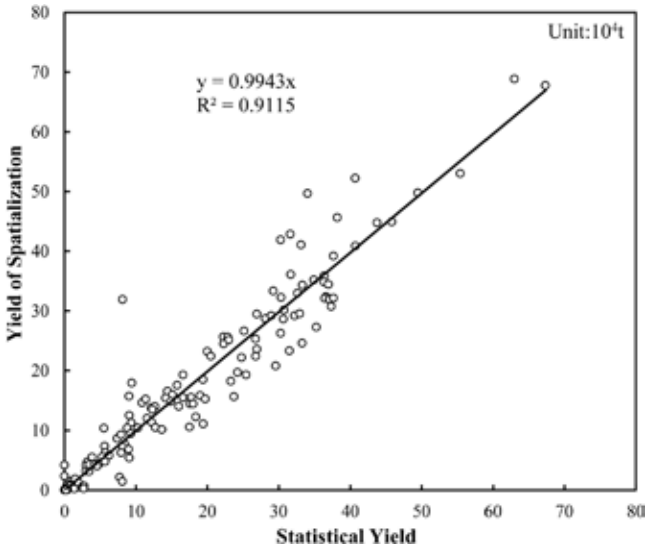

Fig. 7. Comparison of spatialized yield and statistical yield at county level.

Many published studies show that there is a significant correlation between winter wheat yield and the winter wheat growth season NDVI (Chen et al., 2004; Mkhabela et al., 2011). The equations regressing the NDVI variables of the winter wheat growing season and the yield of winter wheat were established. The correlation between the NDVI variables and winter wheat yield, the standardized residual histogram of the regression model, and the RMSE of the fitting equation were analyzed to determine the factors involved in the spatialized model of winter wheat yield in this paper. Currently, the factors involved in the spatial model mainly take into account the statistical relevance rather than the formative mechanism of crop yield. Gross primary productivity is highly correlated with biological productivity and commonly used in crop yield estimations (Lobell et al., 2003; Reeves et al., 2005). Gross primary productivity products are also available from MODIS. In future, we will try to build a crop yield spatialization model from the MODIS gross primary productivity products and compare it with the model based on MODIS NDVI data.

Limitations of the Spatialization Methods. In this paper, to ensure that the error of the spatialized model was distributed within the county administrative unit, our method used the weights of the variables in the county-level administrative unit to allocate the county-level yield. The spatialized results show a block phenomenon. However, this method can control the spatial model error within the minimum allocation unit and improve the accuracy of the spatial model.

Spatialization Differences in Statistical Data at Different Scales. In the spatialization model, the models constructed from data at different scales have several differences. The smaller the geopolitical unit of the statistical data inputted in the model, the higher the accuracy of the model and vice versa. Statistical data at the municipal level do not reflect the differences among counties within the municipality. Visually, the yield distribution map is smoother if it is based on the spatialization of municipal-level statistics than on county-level statistics (Fig. 5 and Fig. 6), which indicates that the local differences of the winter wheat yield spatialization results (Fig. 5) were larger for the county-level statistics than for the municipal-level statistics (Fig. 6).

CONCLUSIONS

Taking the winter wheat in Shandong Province as an example, this study proposed a method for spatialization based on NDVI. On the basis of the MODIS-NDVI time series data of the growing season of winter wheat, the planting area of winter wheat was extracted by the CART algorithm. The NDVI value of each phenological period was calculated by combining the phenological period information for winter wheat. We analyzed the relationship between winter wheat yield with NDVI variables (include single-phase NDVIs and the average NDVI of different growing stages) via univariate linear regression and multiple linear regression and chose the NDVI variables with the highest correlations to winter wheat yield and the minimum RMSE of the fitting equation as input variables to build the spatialization model and make a spatial distribution map of winter wheat yield. The main conclusions are as follows:

(i) On the basis of the MODIS NDVI data of the growing season of winter wheat, we extracted the winter wheat planting area via the CART classification algorithm. The spatial location accuracy estimated with the confusion matrix was 82.51% and the average precision of planting acreage compared with statistical data at the county level was 87.64%.

(ii) The identification accuracy of the winter wheat planting area in SR1 was higher than that in SR2. The commission error of the winter wheat planting area in SR2 was higher than that in SR1. The omission error of the winter wheat planting area in SR1 was higher than that in SR2.

(iii) After consideration of the correlation between winter wheat yield and NDVI variables, the standardized residual histogram of the regression model, the RMSE of the fitting equation, and the influencing factors of winter wheat yield, the final variables used to build the spatialization model were: $NDVI_{6Mar}$, $NDVI_{23Apr}$, $NDVI_{25May}$, and $NDVI_{26June}$.

(iv) The experiment used indirect methods for verifying the accuracy, the spatialization model of winter wheat yield was established from the municipal-level data and the spatial distribution map of winter wheat yield was obtained. We compared the spatialized yield with the officially reported yield at the county level and found the average relative error of spatialization to be 22.71%.

Spatialization of grain yield can provide spatial information for statistical data on grain yield and provide basic data for the development of agriculture. In future research, high spatial–temporal resolution data and accurate parcel data could be introduced to improve the accuracy of extracting grain-planting areas. Other variables (such as gross primary productivity) could be introduced to spatialize crop yields to improve the accuracy of grain yield spatialization.

AUTHOR CONTRIBUTIONS

GX and XZ conceived the study and designed the experiments; GX performed the experiments; KX conducted the data preprocessing; CH and YL analyzed the data; GX and XZ wrote the paper.

CONFLICT OF INTEREST DISCLOSURE

The authors declare that there is no conflict of interest.

ACKNOWLEDGMENTS

This work was supported by the National Key R &D Program of China (Grant No.2017YFC1502505), the National Natural Science Foundation for Distinguished Young Scholars of China (No. 41401479), and the Major Project of High-Resolution Earth Observation System. The authors would thank the United States Geological Survey for providing the MOD13Q1 data (http://glovis.usgs.gov/, accessed 10 May 2019) and Shandong Provincial Bureau of Statistics for providing the winter wheat planting area and winter wheat yield statistical data (http://xxgk.stats-sd.gov.cn/, accessed 10 May 2019).

REFERENCES

Azar, D., R. Engstrom, J. Graesser, and J. Comenetz. 2013. Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. Remote Sens. Environ. 130:219–232. doi:10.1016/j.rse.2012.11.022

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. Classification and regression trees. CRC Press, Boca Raton, FL.

Carletto, C., D. Jolliffe, and R. Banerjee. 2015. From tragedy to renaissance: Improving agricultural data for better policies. J. Dev. Stud. 51:133–148. doi:10.1080/00220388.2014.968140

Chen, J., P. Jönsson, M. Tamura, Z. Gu, B. Matsushita, and L. Eklundh. 2004. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. Remote Sens. Environ. 91:332–344. doi:10.1016/j.rse.2004.03.014

Elvidge, C.D., K.E. Baugh, E.A. Kihn, H.W. Kroehl, E.R. Davis, and C.W. Davis. 1997. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. Int. J. Remote Sens. 18:1373–1379. doi:10.1080/014311697218485

Elvidge, C. D., Imhoff, M. L., Baugh, K. E., Hobson, V. R., Nelson, I., Safran, J., 2001. Night-time lights of the world: 1994–1995. ISPRS J. Photogramm. Remote Sens. 56: 81–99. doi:10.1016/S0924-2716(01)00040-5

Fisher, P.F., and M. Langford. 1995. Modelling the errors in areal interpolation between zonal system by Monte carlo simulation. Environ. Plann. A 27:211–224. doi:10.1068/a270211

Fritz, S., M. Massart, I. Savin, J. Gallego, and F. Rembold. 2008. The use of MODIS data to derive acreage estimations for larger fields: A case study in the south-western Rostov region of Russia. Int. J. Appl. Earth Obs. Geoinf. 10:453–466. doi:10.1016/j.jag.2007.12.004

Ghosh, T., R.L. Powell, C.D. Elvidge, K.E. Baugh, P.C. Sutton, and S. Anderson. 2010. Shedding light on the global distribution of economic activity. Open Geogr. J. 3:147–160. doi:10.2174/1874923201003010147

Hansen, P.M., and J.K. Schjoerring. 2003. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. Remote Sens. Environ. 86:542–553. doi:10.1016/S0034-4257(03)00131-7

Henderson, M., E.T. Yeh, P. Gong, C. Elvidge, and K. Baugh. 2003. Validation of urban boundaries derived from global night-time satellite imagery. Int. J. Remote Sens. 24:595–609. doi:10.1080/01431160304982

Huang, J., H. Wang, Q. Dai, and D. Han. 2014. Analysis of NDVI data for crop identification and yield estimation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7:4374–4384. doi:10.1109/JSTARS.2014.2334332

Huffman, G.J., R.F. Adler, D.T. Bolvin, G.J. Gu, E.J. Nelkin, K.P. Bowman, et al. 2007. The TRMM multi-satellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scale. J. Hydrometeorol. 8:38–55. doi:10.1175/JHM560.1

Ji, G.X., S.B. Liao, and Y.L. Yue. 2015. Spatial distribution of grain yield based on different sample scales and partitioning schemes and its error correction. (In Chinese.) Trans CSAE 31:272–278.

Khan, M.R., C.A.J.M.D. Bie, H.V. Keulen, E.M.A. Smaling, and R. Real. 2010. Disaggregating and mapping crop statistics using hypertemporal remote sensing. Int. J. Appl. Earth Obs. Geoinf. 12:36–46. doi:10.1016/j.jag.2009.09.010

Lei, Z., B. Wu, Z. Liang, and W. Peng. 2012. Patterns and driving forces of cropland changes in the Three Gorges Area, China. Reg. Environ. Change 12:765–776. doi:10.1007/s10113-012-0291-8

Li, X., L.L. Ge, and X.L. Chen. 2013. Detecting Zimbabwe's decadal economic decline using nighttime light imagery. Remote Sens. (Basel) 5:4554–4570. doi:10.3390/rs5094551

Liu, X.H., P.C. Kyriakidis, and M.F. Goodchild. 2008. Population-density estimation using regression and area-to-point residual kriging. Int. J. Geogr. Inf. Sci. 22:431–447. doi:10.1080/13658810701492225

Liu, Z., and B.G. Li. 2012. Spatial distribution of China's grain output based on land use and population density. (In Chinese.) Trans CSAE 28:1–7.

Lobell, D.B., G.P. Asner, J.I. Ortiz-Monasterio, and T.L. Benning. 2003. Remote sensing of regional crop production in the Yaqui Valley, Mexico: Estimates and uncertainties. Agric. Ecosyst. Environ. 94:205–220. doi:10.1016/S0167-8809(02)00021-X

Lobell, D.B. 2013. The use of satellite data for crop yield gap analysis. Field Crops Res. 143:56–64. doi:10.1016/j.fcr.2012.08.008

Mashaba, Z., G. Chirima, J.O. Botai, L. Combrinck, C. Munghemezulu, and E. Dube. 2017. Forecasting winter wheat yields using MODIS NDVI data for the Central Free State region. S. Afr. J. Sci. 113:11–12. doi:10.17159/sajs.2017/20160201

Matlock, R.B., Jr., J.B. Welch, F.D. Parker. 1996. Estimation population, density per unit area from mark, release, recapture data. Ecol. Appl. 6:1241–1253. doi:10.2307/2269604

Mennis, J. 2003. Generating surface models of population using dasymetric mapping. Prof. Geogr. 55:31–42. doi:10.1111/0033-0124.10042

Mkhabela, M.S., P. Bullock, S. Raj, S. Wang, and Y. Yang. 2011. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agric. For. Meteorol. 151:385–393. doi:10.1016/j.agrformet.2010.11.012

Mkhabela, M.S., M.S. Mkhabela, and N.N. Mashinini. 2005. Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR. Agr Forest Meteorol. 129(1–2):1–9. doi:10.1016/j.agrformet.2004.12.006

Monfreda, C., N. Ramankutty, and J.A. Foley. 2008. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. Global Biogeochem. Cycles 22: GB1022. doi:10.1029/2007GB002947

Potgieter, A.B., K. Lawson, and A.R. Huete. 2013. Determining crop acreage estimates for specific winter crops using shape attributes from sequential MODIS imagery. Int. J. Appl. Earth Obs. Geoinf. 23:254–263. doi:10.1016/j.jag.2012.09.009

Potter, P., N. Ramankutty, E.M. Bennett, and S.D. Donner. 2010. Characterizing the spatial patterns of global fertilizer application and manure production. Earth Interact. 14:1–22. doi:10.1175/2009EI288.1

Reeves, M.C., M. Zhao, and S.W. Running. 2005. Usefulness and limits of MODIS GPP for estimating wheat yield. Int. J. Remote Sens. 26:1403–1421. doi:10.1080/01431160512331326567

Stahl, K., R.D. Moore, J.A. Floyer, M.G. Asplin, and I.G. Mckendry. 2006. Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. Agric. For. Meteorol. 139:224–236. doi:10.1016/j.agrformet.2006.07.004

Sutton, P., D. Roberts, C. Elvidge, and K. Baugh. 2001. Census from Heaven: An estimate of the global human population using night-time satellite imagery. Int. J. Remote Sens. 22:3061–3076. doi:10.1080/01431160010007015

Tobler, W.R., U. Deichmann, J. Gottsegen, and K. Malloy. 1995. The global demography project. National Centre for Geographic Information and Analysis, Univ. of California Santa Barbara, Santa Barbara, CA,

Tobler, W., Deichmann, U., Gottsegen, J., Maloy, K., 1997. World population in a grid of spherical quadrilaterals. Popul. Space Place 3:203–225. doi:10.1002/(SICI)1099-1220(199709)3:3<203::AID-IJPG68>3.0.CO;2-C

Vicente-Serrano, S., M. Saz-Sánchez, and J. Cuadrat. 2003. Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): Application to annual precipitation and temperature. Clim. Res. 24:161–180. doi:10.3354/cr024161

Wu, S., X. Qiu, and L. Wang. 2006. Using semi-variance image texture statistics to model population densities. Cartogr. Geogr. Inf. Sci. 33:127–140. doi:10.1559/152304006777681670

Yang, X., Y. Huang, P. Dong, D. Jiang, and H. Liu. 2009. An updating system for the gridded population database of China based on remote sensing, GIS and spatial database technologies. Sensors (Basel) 9:1128–1140. doi:10.3390/s90201128

You, L. Z., and S., Wood. 2005. Assessing the spatial distribution of crop areas using a cross-entropy method. Int. J. Appl. Earth Obs. Geoinf. 7:310–323. doi:10.1016/j.jag.2005.06.010

You L., Wood S., 2006. An entropy approach to spatial disaggregation of agricultural production. Agr Syst. 90(1–3):329–347. doi:10.1016/j.agsy.2006.01.008

You, L.Z., S. Wood, and U. Woodsichra. 2009. Generating plausible crop distribution maps for Sub-Saharan Africa using a spatially disaggregated data fusion and optimization approach. Agric. Syst. 99:126–140. doi:10.1016/j.agsy.2008.11.003

Yue, W.Z., J.B. Gao, and X.C. Yang. 2014. Estimation of gross domestic product using multi-sensor remote sensing data: A case study in Zhejiang province, East China. Remote Sens. (Basel) 6:7260–7275. doi:10.3390/rs6087260

Zhu, X., P. Shi, and Y. Pan. 2012. Development of a gridded dataset of annual irrigation water withdrawal in China. In: International Conference on Agro-Geoinformatics, Shanghai. 2–4 Aug. 2012. IEEE, Shanghai.