Farmland productivity estimation based on vegetation indexes from remote sensing data

Shaoshuai Li[1,*], Baipeng Li[2], and Wenjing Cao[3]

[1]Land Consolidation and Rehabilitation Center, Ministry of Natural Resources, 100035 Beijing, China
[2]Aerospace Information Research Institute, Chinese Academy of Science, 100094 Beijing China
[3]Shenzhen Development Research Center for Real Estate and Urban Construction, 518000 Shenzhen, China

Abstract. Ensuring food security is a long-term and arduous task. Timely and accurate grasp of grain production capacity information can provide favourable data support for the nation to formulate macroeconomic plans and food policies. With the development of remote sensing technology, it has been widely used in crop yield estimation models. In this paper, the yield of spring maize in Da'an of Jilin province was estimated based on vegetation indexes calculated from Landsat-8 images. The results have shown that the fitting degree and estimation accuracy of yield estimation models at tasselling stage are significantly better than those at milk stage. Among these vegetation indexes, the model based on GNDVI has better fitting degree and estimation accuracy. This paper can provide reference for the post construction evaluation of high standard farmland in China.

1 Introduction

Ensuring food security is a long-term and arduous task. Timely and accurate grasp of grain production capacity information can provide data support for the nation to formulate related policies. With the development of remote sensing technology, remote sensing has been widely used in crop yield estimation models. For example, Sakamoto et al. [1] used multi-temporal remote sensing data and crop phenology characteristics to establish a statistical relationship between crop yield and vegetation indexes to estimate crop yield, and high estimation accuracy was obtained. J.Q. Ren. et al. [2] took American maize as the research object and each state in USA as the yield estimation area, and selected the best model by developing a relationship between NDVI and estimated maize yield of each state in 2011, and predicted the maize yield per unit. The results showed that the relative error of maize yield was only 2.12%. L.Y. Liu et al. [3] carried out statistical analysis on the ground spectral data and wheat yield data of each growth period, and built the yield estimation model of each growth period by analyzing the correlation coefficient curve, which displayed a higher accuracy. L. Bai et al. [4] measured the reflectance of cotton canopy at different stages with hyperspectral remote sensing data, and analyzed the relationship between spectral reflectance and yield. With the continuous emergence of high temporal and spatial resolution remote sensing data, remote sensing shows more and more incomparable advantages in crop yield estimation. It has become an inevitable trend to combine remote sensing data with traditional statistical data, meteorological data and agronomic data to estimate productivity. In this paper, the yield of spring maize in Da'an City of Jilin Province was estimated by using vegetation indexes from Landsat 8 satellite remote sensing data, to provide reference for the post construction evaluation of high standard farmland in China.

2 Study area and research data

2.1 Study area

Da'an (44°57'~ 45°45' N, 123°8' ~124°21' E) (Fig. 1) is a world-famous golden maize zone and a national grain base, situated in hinterland of Songnen plain, the northwest of Jilin province. Selection of Da'an was motivated by its typicality and existing field data. Da'an is about 4879km2, characterized by a continental monsoon climate. In an average year, annual mean temperature is 4.3℃ and annual accumulated temperature is 2921.3℃, and 3012.8 hours of sunshine and a precipitation total of about 413.7mm can be expected across a year. According to the statistical yearbook of Da'an for the past five years, spring maize is the staple crop, accounting for 87% of the total planting area. In this study, maize yield was estimated.

* Corresponding author: liss715@163.com

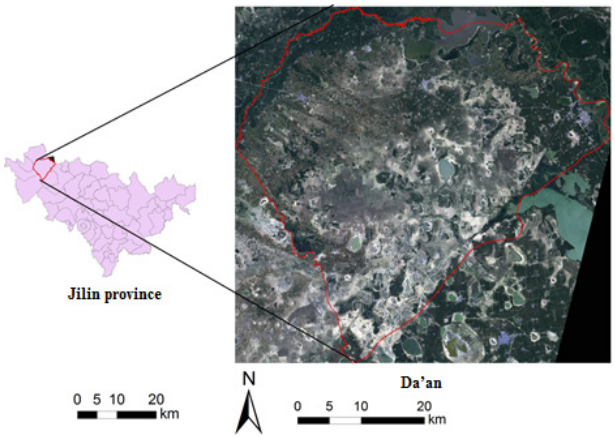


Fig. 1. Geographical location of Da'an

2.2 Data acquisition and processing

(1) Satellite remote sensing data. Spring maize is planted in late April and harvested in mid-September in Da'an. Studies have shown that (Guan K et al., 2017; Han Wenting et al., 2020; Zhao Wenliang et al., 2012; Zhu Wanxue et al., 2018) there are significant differences in the accuracy of crop yield estimation models based on vegetation index at different growth stages, with the highest accuracy at tasseling stage, followed by the milk stage. Thus, Landsat 8 data taken on July 21, 2017 (tasseling stage) and August 22, 2017 (milk stage) were selected in the study. After radiometric calibration, atmospheric correction with FLAASH and mosaicking, the remote sensing data were then geometrically corrected based on land use change survey data. Besides, clouds were removed with FMask cloud detection function of Envi, as there were several clouds on Lansat 8 images at milk stage.

(2) Vector data of land use change survey. This study obtained the cultivated land data of the land use change survey in Da'an in 2017, including dry land, irrigated land and paddy field. It should be noted that paddy fields are mainly planted with rice, while dry land and irrigated land are planted with maize. Therefore, the map spots of paddy field were excluded from the remote sensing images, and the vegetation indexes of dry land and irrigated land were then calculated. According to the investigation, there was no field interplanting in Da'an, and crop species can be distinguished based on the land change survey data (Genovese G et al., 2001). Therefore, the mixed pixel decomposition of spring maize was not considered.

(3) Survey data of spring maize yield. This paper adopted the method of household survey to obtain the spring maize yield data of 98 samples in 2017.

3 Study method

3.1 Vegetation index

Combining Landsat 8 images and vegetation characteristics of spring maize and referring to some literatures [5-7], six vegetation indexes closely related to crop yield were selected in the study, which were given in Table 1.

Table 1. Main vegetation indexes.

Vegetation Indexes	Equation
The Green Normalized Difference Vegetation Index (GNDVI)	$GNDVI = (NIR–G)/(NIR+G)$
The Normalized Difference Vegetation Index (NDVI)	$NDVI = (NIR–R)/(NIR+R)$
The Optimized Soil Adjusted Vegetation Index (OSAVI)	$OSAVI = (NIR–R)/(NIR+R+x)$ $(x=0.16)$
The Soil Adjusted Vegetation Index (SAVI)	$SAVI = (1+L)(NIR–R)/(NIR+R+L)$, $(L=0.5)$
The Structure Insensitive Pigment Index (SIPI)	$SIPI = (NIR–B)/(NIR+B)$
Enhanced Vegetation Index 2 (EVI2)	$EVI2 = 2.5(NIR–R)/(NIR+2.4R+1)$

R is red band reflectance, G is green band reflectance and NIR is near-infrared reflectance; L is a soil adjusted coefficient. Generally, an L=0.5 is the default value, which is used for correcting for the influence of soil brightness. x is an adjusted coefficient, and an x=0.16 is the default value, which can optimize L.

3.2 Spring maize yield estimation model based on remote sensing technologies

Linear regression model was adopted in the study, which applied least squares to establish a statistical relationship between spring maize yield and vegetation index. The regression line is:

$$Y = aVI+b \quad (1)$$

Where b is a constant, a is the regression coefficient, VI (vegetation index) is the value of the independent variable, and Y is the value of the dependent variable (yield). 69 samples (70%) were randomly chosen from the whole 98 samples using Geostatistical Analyst tool of ARCMAP, to develop a model, and 29 samples (30%) were used for validation.

3.3 Accuracy assessment

R2 (determination coefficient) and RMSE (root-mean-square error) were adopted for accuracy assessment. The higher R2 is, the better the model fitting degree is. The lower the RMSE, the higher the model accuracy is.

4 Results and analysis

4.1 Preliminary analysis of sample data

Among the 98 samples, the minimum yield was 2175 kg/ hm2 and the maximum yield was 7500 kg/ hm2. Fig. 2 showed the frequency histogram of sample yield with a normal distribution curve. Wholly speaking, the distribution was skewed. The spring maize yield was concentrated at 5,000~6,500 kg/hm2, which was representative to a certain extent.

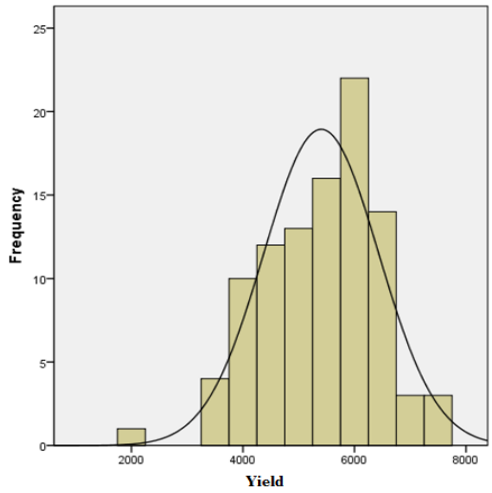
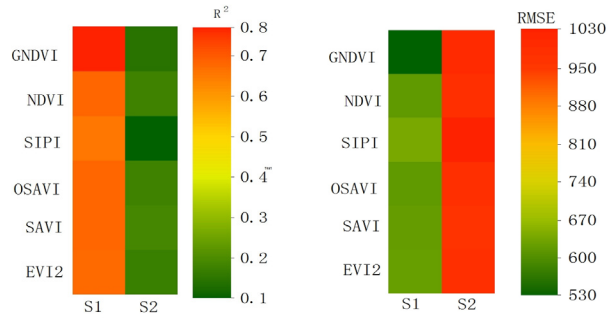


Fig. 2. Frequency histogram of spring maize sample yield

4.2 Comparative analysis of yield estimation models based on data of tasselling and milk stages

Fig. 3 showed the R2 and RMSE of models built at tasseling stage and milk stage, S1 represented tasseling stage and S2 represented milk stage. The R2 at tasseling stage was greater than 0.6, with the highest value of 0.75. The R2 at the milk stage were less than 0.2. RMSE at tasseling stage was about 600 kg/ hm2, and RMSE at milk stage was between 900 kg/ hm2 to 1000 kg/ hm2. It can be seen that the fitting degree and evaluation accuracy of the yield estimation model at tasseling period are significantly better than that of models at the milk period.



R^2 RMSE

Fig. 3. Comparison of spring maize yield estimation models at different stages (tasselling and milk stages)

4.3 Analysis of yield estimation models based on different vegetation indexes

Regression equations were established after remote sensing data taken on tasseling period was further analyzed, and the result were shown on Table 2. In terms of model fitting degree, the model built on GNDVI provided the highest R2 (R2=0.756), while other models provided a slightly lower R2 (all are greater than 0.6), so the overall fitting was good. In terms of model estimation accuracy of training samples, the model built on GNDVI provided with the lowest RMSE (RMSE=531.74), while RMSE of the rest were above 600. In model estimation accuracy of testing samples aspect, the model developed on SIPI provided with the lowest RMSE, while the rest provided slightly higher RMSE. In general, models based on GNDVI was good at fitting degree and yield estimation accuracy. The estimated yield distribution diagrams of spring maize based on the GNDVI models were shown in Fig. 4.

Table 2. Regression equations developed between estimated yield and actual yield based on the vegetation indexes of spring maize at tasselling stage.

Vegetation index	Regression equation	R^2	Training samples RMSE	Testing samples RMSE
GNDVI	$y_{pre}=16023.321y_{act}-5172.023$	0.756	531.74	993.37
NDVI	$y_{pre}=10727.228y_{act}-1904.299$	0.672	616.68	985.51
SIPI	$y_{pre}=15160.722y_{act}-6162.174$	0.645	641.73	928.37
OSAVI	$y_{pre}=9247.964y_{act}-1904.28$	0.672	616.68	985.51
SAVI	$y_{pre}=14723.397y_{act}-6479.048$	0.668	620.86	974.32
EVI2	$y_{pre}=3731.373y_{act}+168.788$	0.664	624.26	995.38

y_{pre} is the estimated yield and y_{act} is the actual yield.

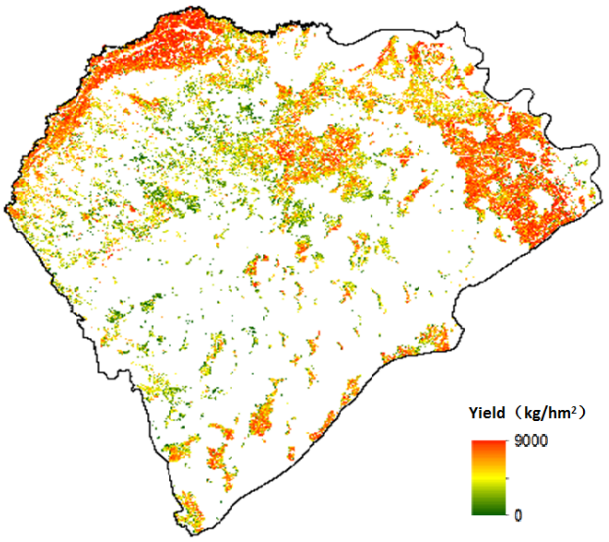


Fig. 4. Distribution diagram of estimated yield of spring maize based on GNDVI

Fig. 5 illustrated that when the actual yield at sampling site was less than 5500 kg/ hm2, the model may overestimate the yield; otherwise, the model may underestimate the yield.

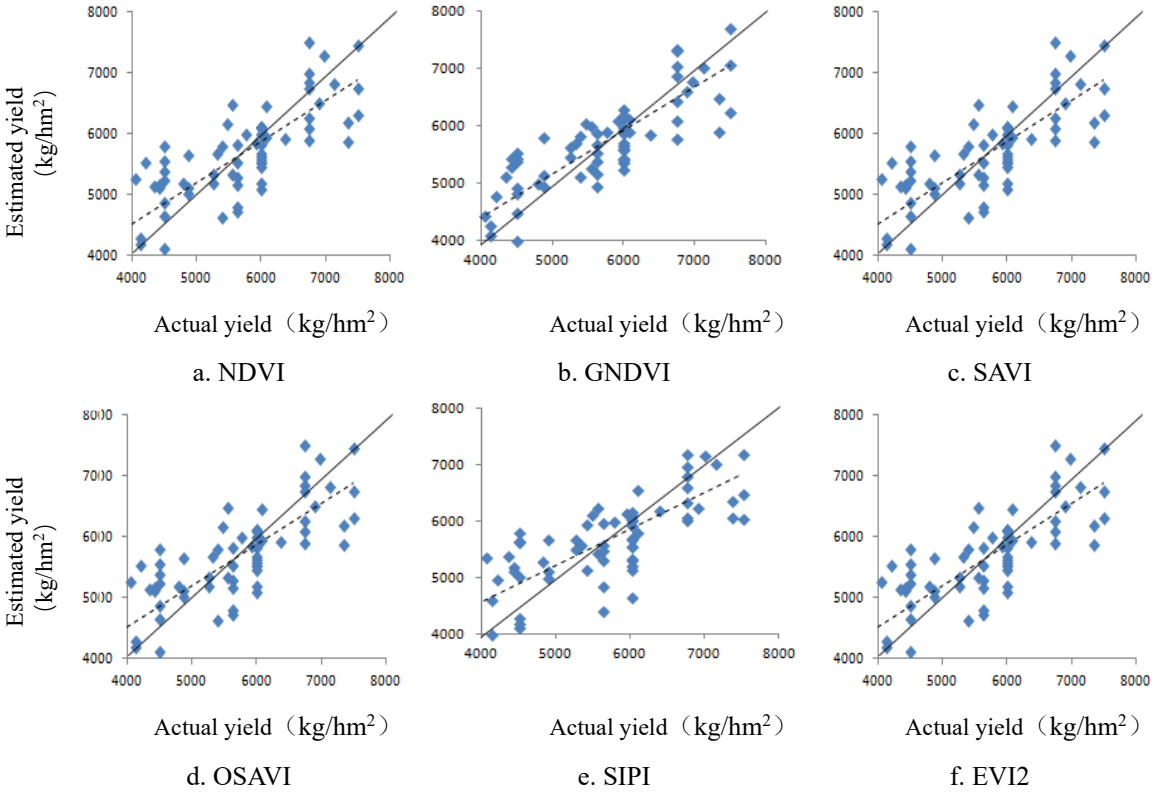


Fig. 5. Comparison between estimated yield and actual yield of spring maize at tasselling stage

5 Conclusions and discussions

5.1 Conclusions

In this paper, linear regression models were established to estimate the spring maize, and the performance of different vegetation indexes were compared. The result indicates that the fitting degree and evaluation accuracy of the model built on tasseling period data are better than that of the model built on milk period data. Among all vegetation indexes, the model based on GNDVI exhibits better performance on fitting degree and estimation accuracy. However, when the actual yield of samples is lower than 5500 kg/hm2, the model will overestimate the yield; otherwise, the model will underestimate the yield.

5.2 Discussions

This study attempted to develop multiple linear regression models using multiple vegetation indexes, but the regression model can not meet the statistical requirements due to collinearity among the indexes. Ongoing work can be focused on two aspects. First, in terms of model construction, the linear model is currently used to fit the complex relationship between spring maize yield and vegetation indexes, and the artificial neural network algorithm can be explored to improve the fitting accuracy of the model in the future. Second, in terms of selection of remote sensing data, GF-2 data with higher spatial resolution and GF-6 data carrying red-edge band can be adopted in the next step, which may improve the estimation accuracy.

References

1. T. Sakamoto, M. Yokozawa, H. Toritani, M. Shibayama, N. Ishitsuka, H. hno. Remote Sens Environ, **96(3)**: 366-374 (2005).
2. J.Q. Ren, Z.X. Chen, Q.B. Zhou, J. Liu, H.J. Tang. J. Remote Sens, **19 (4)** : 568- 577 (2015).
3. L.Y. Liu, J.H. Wang, W.J. Huang, C.J. Zhao, B. Zhang. Q.X. Tong. Trans. CSAE. **20(1)**: 172-175 (2004)
4. L. Bai. J. Wang. G.Y. Jiang.P. Yang.S.J. Sun., CAAS. **41(8):**2499-2505 (2008)
5. K. Guan. J. Wu. J.S. Kimball. Remote Sens Environ, **199(3)**: 333-349 (2017)
6. W.X. Zhu. S,J, Li, X.B. Zhang. Y. Li. Z.G. Sun. Trans. CSAE. **34(11)**: 78-86 (2018)
7. W.T. Han. X.S. Peng. L.Y. Zhang. Y.X. Niu. Trans. CSAM. **51(01)**: 148-155 (2020)