

A STUDY ON OILSEED RAPE YIELD ESTIMATION BASED ON ENHANCED VEGETATION INDEX AT THE FLOWERING STAGE AND METEOROLOGICAL DATA

Qian Zhan¹

1. School of Earth Sciences and Resources, China University of Geosciences (Beijing)

ABSTRACT

Oilseed rape is a significant oilseed crop in China. Regional yield estimation is a highly effective method for ensuring the profitability of agriculture and addressing the issue of food security. This study focused on the principal oilseed rape regions in Zhejiang Province, China. A regression model based on remote sensing and meteorological data from 2015-2019 was proposed to estimate the yield of oilseed rape in Zhejiang Province by municipality. In the model, the average annual yield of oilseed rape depends on the mean enhanced vegetation index (EVI) at the flowering stage, the effective cumulative temperature in February and the total precipitation in May. The estimated data were cross-validated with yield data published by the Zhejiang Bureau of Statistics for the same period. The coefficient of determination (R^2) of the model is 0.839.

Index Terms— yield estimation, oilseed rape, flowering stage, EVI, meteorological data.

1. INTRODUCTION

As China's number one oilseed crop, rapeseed plays an important role in China's edible oil supply. Oilseed rape is one such commodity in high demand for both national and global consumptions. The development of remote sensing technology has brought innovations to agricultural yield estimation.

There are currently a large number of studies using remote sensing imagery for yield estimation. Among them, the most common method for estimating crop biomass is based on vegetation index. The basic principle is to analyze the correlation between different vegetation indices and crop biomass, then select the vegetation index with the highest accuracy and regression model by comparison, to construct a more appropriate empirical model for crop biomass estimation, and then make predictions of crop biomass [1].

Previous studies have shown that using vegetation indices at the crop flourishing or flowering stage gives optimal results for crop yield estimation [2, 3]. In recent years, more literature has shown that there is a strong relationship between SPAD (Soil and Plant Analyzer Development) chlorophyll values, leaf area index (LAI) and vegetation

indices. Some scientists use the vegetation indices to construct chlorophyll models for oilseed rape. Yin Zi et al. investigated the relationship between different vegetation indices and SPAD values of oilseed rape leaves at different fertility periods and established an estimation model. By comparing the correlation between the SPAD values of oilseed rape leaves at the whole fertility period and the vegetation indices, it was found that the model fit of oilseed rape at the flowering period was the best in the whole fertility period [4]. Therefore, using the vegetation index at flowering as an independent variable in the model helps to improve the accuracy compared to data from other periods.

In this study, we used the remote sensing and meteorological data at the flowering stage of oilseed rape to build the yield estimation model.

2. STUDY AREA AND DATA

2.1. Study area

The study area covers the major oilseed rape fields in Zhejiang Province between 27°02'N-31°11'N latitude and 118°01'E-123°10'E longitude, including the cities of Hangzhou, Huzhou, Jiaxing, Jinhua, Ningbo, Shaoxing, Wenzhou and Quzhou (Fig. 1). The region is one of the major production areas of oilseed rape in China, which has a subtropical monsoon climate, with average annual temperature between 15 °C and 18 °C, annual sunshine duration between 1100 and 2200 hours, and average annual precipitation between 1100 and 2000 mm [5].

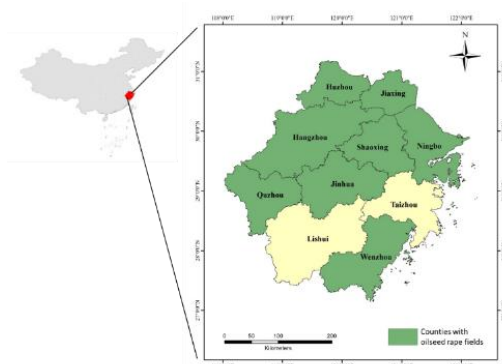


Fig.1 The location of the study area

Oilseed rape in the study area is a semi-wintering *Brassica napus*. It has four stages in the growing cycle of approximately 225 days: seedling stage, bolting stage, flowering stage, and pod stage. Table 1 shows the period of the four stages.

Table 1 Period of the four stages of oilseed rape in the study area

Stage	Seedling stage	Bolting stage	Flowering stage	Pod stage
Period	Oct.1- Jan.31	Feb.1- Mar.10	Mar.11- Apr.20	Apr.21- May.30

2.2. Data

We used three kinds of data from 2015 to 2019 to build the yield estimation model: remote sensing satellite images, climate characteristics, and yield data. All the data are at the flowering stage of the oilseed rape (Mar.11-Apr 20).

Remote sensing satellite images were used to collect vegetation indices of the oilseed rape fields. Landsat 8 OLI panchromatic and multi-spectral images came from *Geospatial Data Cloud* (<https://www.gscloud.cn>). After the atmospheric correction and radiometric calibration, the panchromatic band was fused with other bands to obtain a multi-spectral image with a resolution of 15 meters. Sentinel 2 data were a supplement when Landsat 8 data could not meet the requirements, coming from Copernicus Open Access Hub (<https://scihub.copernicus.eu>).

Meteorological characteristics, including precipitation, cumulative temperature, negative cumulative temperature, and hours of sunlight duration, came from the China Greenhouse Data Platform (<http://data.sheshiyuanyi.com/>).



Yield data came from the Statistical Yearbook of the Zhejiang Province (<http://tjj.zj.gov.cn/col/>), with city-level yield.

3. METHODS

3.1. Establishment of oilseed rape mask

During the flowering stage, the color of the corolla - yellow or orange - differs from that of other crops at the same time. Based on the characteristics, we established the interpretation flags (Table 2) and then extracted the oilseed rape fields in the study area using maximum-likelihood method (MLM).

Table 2 Interpretation flags of oilseed rape

True Color Image	False Color Image	Texture
		Stripe Block

As the reflectance of flowering oilseed rape at 520-600 nm and 760-900 nm is higher than that of other vegetation,

non-oilseed rape samples can be excluded by using the range of *GNDVI* values [6].

$$GNDVI = \frac{NIR - GRE}{NIR + GRE}$$

NIR is the reflectance value of the near inferred band. *GRE* is that of the green band.

The distribution of oilseed rape in Zhejiang province was obtained and is illustrated in Fig 2. The dark red spots are the oilseed rape areas.

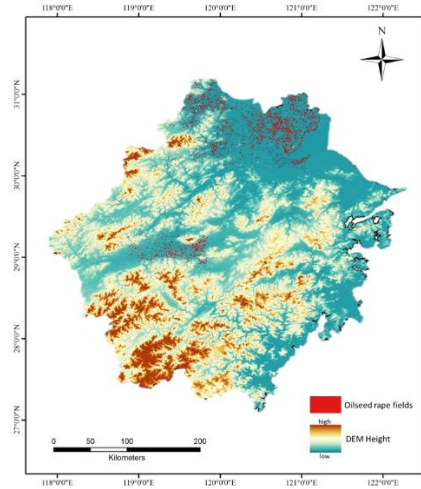


Fig. 2 Zhejiang Province oilseed rape distribution

3.2. Regression model

A multiple linear regression model was developed to estimate the oilseed rape yield. We used backward stepwise regression to build the model. The oilseed rape yield was chosen as a dependent variable in the model, while the vegetation index and meteorological characteristics were chosen as independent variables. We considered one dependent and eight independent variables:

- Y —average annual oilseed rape yield estimation by municipality, kg/ha;
- x_1 —the mean *EVI* value at the flowering stage the mask of the municipality's oilseed rape field;

$$EVI = 2.5 \times \left(\frac{NIR - RED}{NIR + 6RED - 7.5BLU + 1} \right)$$

NIR is the reflectance value of the near inferred band. *RED* is that of the red band. *BLU* is that of the blue band.

- x_2 —effective cumulative temperature in February by municipality, °C;

$$x_2 = \sum_{i=1}^n (T_i - 3)$$

n is the number of days in February and T_i is the average temperature on day i in February. $T_i > 3$, when $T_i < 3$, $(T_i - 3)$ is 0

- x_3 —total precipitation in May by municipality, mm;
- x_4 —negative cumulative temperature in January by municipality, °C;

$$x_4 = \sum_{i=1}^n (3 - T_i)$$

n is the number of days in January and T_i is the daily minimum temperature on day i in January. $T_i < 3$, when $T_i > 3$, $(3 - T_i)$ is 0

- x_5 —total hours of sunshine in March-April by municipality, hour;
- x_6 —effective cumulative temperature in March-April by municipality, °C;

$$x_6 = \sum_{i=1}^n (T_i - 10)$$

n is the number of days in March-April and T_i is the average daily temperature on day i in March-April. $T_i < 10$, when $T_i > 10$, $(T_i - 10)$ is 0

- x_7 —active cumulative temperature in March-April by municipality, °C;

$$x_7 = \sum_{i=1}^n T_i$$

n is the number of days in March-April and T_i is the average daily temperature on day i in March-April. $T_i > 10$, when $T_i < 10$, T_i is 0

- x_8 —total precipitation in March-April by municipality, mm.

The multivariate regression model was constructed as follows:

$$Y = b_0 + \sum_{j=1}^8 b_j x_j$$

b_0 is a constant, b_j is bias regression coefficient of the variable x_j .

3.3 Correlation analysis

This paper examines the variable's correlations by Pearson's correlation coefficient matrix (Table 3).

Table 3 Correlation matrix of dependent and independent variables

	Y	x1	x2	x3	x4	x5	x6	x7	x8
Y	1	.476	-.84	-.483	.443	.552	-.417	-.619	-.496
x1	.476	1	-.26	-.156	.249	0.194	-.479	-.535	-.234
x2	-.840	-0.257	1	.265	-.543	-.662	0.188	.454	.556
x3	-.483	-.156	0.265	1	.318	-.246	.488	.469	0.278
x4	.443	0.249	-.543	.318	1	.481	-0.072	-0.311	-0.216
x5	.552	0.194	-.662	-.246	.481	1	-.042	-.276	-.377
x6	-.417	-.479	.188	.488	-.072	-.042	1	.924	0.131
x7	-.619	-.535	.454	.469	-.311	-.276	.924	1	0.232
x8	-.496	-.234	.556	0.278	-.216	-.377	0.131	0.232	1

The significance level ($p < 0.05$) are highlighted in bold font.

The correlation coefficients with absolute values exceeding 0.6 are underlined in the table. The variables x_2 and x_5 ($\tau = -0.662$) are significantly correlated, while the variables x_6 and x_7 ($\tau = 0.924$) are highly correlated.

3.4. Validation

We used a leave-one-out (LOO) cross-validation, which leaves out one year at a time, permitting a comparison between the actual and estimated yield at that year.

To evaluate the accuracy of the estimation, we used the coefficient of determination (R^2), adjusted R^2 , root mean square error ($RMSE$), and mean absolute percentage error ($MAPE$) between estimated yield and actual yield. The $MAPE$ was expressed as a percentage, as follows:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

y_i is the actual value of the i th data, \hat{y}_i is the estimated value of the i th data, and n is the sample size.

4. RESULTS AND CONCLUSION

Variables x_5 and x_6 were excluded from the correlation analysis. Variables x_4 , x_7 , and x_8 were automatically excluded during stepwise model construction as insignificant indicators. Consequently, the multiple regression equation, which characterizes the dependence of oilseed rape yield in the study area on the variables included in the model, constructed according to the data of 2015–2018, has the following form:

$$Y = 1003.7x_1 - 6.216x_2 - 0.967x_3 + 2456.397$$

The model's coefficient of determination (R^2) is 0.839; the adjusted- R^2 is 0.823; the $RMSE$ is 0.1249 t/ha; the $MAPE$ is 4.174. All coefficients of this regression equation were found to be statistically significant ($p < 0.01$). See Table 4.

Table 4 The summary of regressions for the dependent variables

Source	Value	Std Dev	t	Pr > t
Intercept	2456.397	245.144	10.020	< 0.0001
x_3	-6.216	0.665	-9.347	< 0.0001
x_2	-0.967	0.279	-3.459	0.002
x_1	1003.700	291.802	3.440	0.002

As shown in the Fig. 3, the estimated values for most of the observed years are within the confidence interval ($\gamma = 0.95$) of the actual oilseed rape yield.

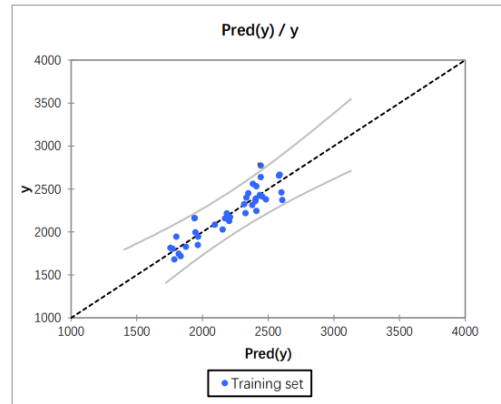


Fig.3 Actual and modelled estimates of oilseed rape yield in selected areas of Zhejiang Province, 2015-2019

After the LOO cross-validation, the results of the validation are presented in Table 5.

Table 5 Cross-validation results

Year	Adjusted_R ²	RMSE (kg/ha)	MAPE (%)
2015	0.803	133.474	4.393
2016	0.793	130.353	4.101
2017	0.806	136.288	4.544
2018	0.839	115.264	3.875
2019	0.855	114.011	3.667

The statistical significance of all validated models is less than 0.0001. The adjusted coefficient of determination of all models is above 0.8, with the exception of 2017, which exhibits the highest fit in 2019. All the MAPE are less than 5%, indicating that the regression models are accurate for estimating the year yield of oilseed rape in Zhejiang Province.

5. REFERENCES

- [1] Yuanbo W, Dejun F, Shujuan L, et al. Research progress on crop biomass estimation based on remote sensing information [J]. Remote sensing technology and applications, 2016, 31(3): 468-75.
- [2] Cai Y, Guan K, Lobell D, et al. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches [J]. Agricultural and Forest Meteorology, 2019, 274: 144-59.
- [3] Stepanov A, Dubrovin K, Sorokin A, et al. Predicting Soybean Yield at the Regional Scale Using Remote Sensing and Climatic Data [J]. Remote Sensing, 2020, 12(12): 1936
- [4] Zi Y, Qinrui C, Miao L, et al. SPAD estimation of oilseed rape leaves at different fertility stages based on spectral indices[J]. Journal of Northwest Agriculture and Forestry University (Natural Science Edition), 2017, 45(5): 66-72.
- [5] Zheng S. Physical Geography and Population of Zhejiang [J]. Policy Outlook, 2007, (01): 53-.
- [6] Dong W, Shenghui F, Zheng W. Research on oilseed rape extraction based on spectral features and colour features [J]. Agricultural Mechanics Journal, 2018, 49(3): 158-65.