Postprint: Remote Sensing Methods for Wheat Yield Estimation in Northern Kazakhstan

Authors: Yin Hanmin

Date: 2022-04-14T00:48:29+00:00

Abstract

Taking the rain-fed farming areas in northern Kazakhstan as the study target area, based on spring wheat yield statistical data and remote sensing spectral indices, we conducted analysis on the optimal prediction period for spring wheat yield estimation and vegetation indices. Using regression analysis, random forest, support vector machine, and bidirectional recurrent neural network models to estimate spring wheat yield, we comparatively analyzed the simulation accuracy of different models. The results show that: for North Kazakhstan Oblast, Akmola Oblast, and Kostanay Oblast, the optimal prediction period for spring wheat yield estimation from 2007 to 2016 was June 26 to August 5, which is the critical period for spring wheat yield formation. The optimal vegetation index for spring wheat yield estimation in North Kazakhstan Oblast was the Green Chlorophyll Index (CIgreen) on July 12; in Akmola Oblast, it was the Green Wide Dynamic Range Vegetation Index (WDRVIgreen) on August 5; and in Kostanay Oblast, it was WDRVIgreen on July 12. Comparative analysis of the accuracy of the four models in simulating spring wheat yield revealed that, under conditions of limited sample points, the bidirectional recurrent neural network model achieved higher accuracy compared to other models in estimating spring wheat yield in the three northern oblasts of Kazakhstan. Correlation analysis results between spring wheat yield and Net Primary Productivity (NPP) show that the area proportions with coefficient of determination R^2 above 0.50 in North Kazakhstan Oblast, Akmola Oblast, and Kostanay Oblast were 44%, 94%, and 77%, respectively, indicating that the above yield estimation models can be applied to spring wheat yield estimation in the three northern oblasts of Kazakhstan, especially in Akmola Oblast and Kostanay Oblast.

Full Text

Preamble

Wheat Yield Estimation with Remote Sensing in Northern Kazakhstan

YIN Hanmin[1,2], Guli JIAPAER[1,2,3], YU Tao[1,2], Jeanine UMUHOZA[1,2], LI Xu[1,2]

[1]State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, Xinjiang, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Research Center for Ecology and Environment of Central Asia, Chinese Academy of Sciences, Urumqi 830011, Xinjiang, China

Abstract: Kazakhstan ranks as the world's leading flour exporter and serves as the granary of Central Asia. Its northern regions—North Kazakhstan, Aqmola, and Qostanay—are globally important wheat and flour production areas, with wheat comprising 86% of the cropping structure. Since 2010, Kazakhstan has ranked 12th worldwide in wheat and barley production and 5th in export volume. However, the rain-fed agricultural system in this region, combined with frequent drought stress in the monsoon climate zone, often leads to large-scale yield reductions that severely impact food security in import-dependent nations. This study targets the rain-fed farming zone of northern Kazakhstan to analyze optimal prediction timing and vegetation indices for spring wheat yield estimation using statistical yield data and remote sensing spectral indices. We employed regression analysis, random forest, support vector machine, and bidirectional recurrent neural network models to estimate spring wheat yields, comparing their simulation accuracy. Results indicate that for North Kazakhstan, Aqmola, and Qostanay, the optimal prediction period falls between July 12 and August 5, coinciding with the critical yield formation stage. The optimal vegetation index is the green chlorophyll index (CIgreen) for North Kazakhstan, the green wide dynamic range vegetation index (WDRVIgreen) for Aqmola, and WDRVIgreen for Qostanay. Comparative analysis reveals that, with limited sample points, the bidirectional recurrent neural network achieves higher accuracy than other models for estimating spring wheat yields across the three northern states of Kazakhstan. Correlation analysis with net primary productivity shows determination coefficients (R^2) of 0.44, 0.94, and 0.77 for North Kazakhstan, Aqmola, and Qostanay respectively, demonstrating that these models can be applied to spring wheat yield estimation in northern Kazakhstan, particularly in Aqmola and Qostanay.

Keywords: rain-fed wheat farming area; remote sensing yield estimation; vegetation index; regression model; machine learning; northern Kazakhstan

1. Introduction

Kazakhstan' s flour export volume ranks first globally, earning it the designation as Central Asia' s granary. The northern region—including North Kazakhstan, Aqmola, and Qostanay—represents a major global wheat and flour export zone, with wheat accounting for 86% of the cropping structure. Since 2010, Kazakhstan has ranked 12th in world wheat and barley production and 5th in export volume. In contrast, other Central Asian countries, constrained by large economic crop proportions, irrational planting structures, and limited arable land, face severe restrictions on grain production capacity, requiring substantial annual wheat imports from Kazakhstan to meet domestic demand.

Crop yield estimation methods can be categorized into traditional and remote sensing approaches. Traditional methods typically involve regional manual surveys combined with agronomic and meteorological statistics to establish yield models. However, these approaches are time-consuming, labor-intensive, and unsuitable for dynamic spatiotemporal monitoring. Modern remote sensing technology provides effective tools for regional grain estimation and dynamic monitoring. Remote sensing-based methods 主要包括 three categories: empirical models, machine learning, and mechanistic models. Empirical models utilize electromagnetic wave information reflected from crop canopies to calculate vegetation indices that characterize crop conditions, establishing statistical relationships with actual yields to identify optimal indices for yield estimation. Machine learning models excel at handling high-dimensional variables and capturing complex linear and nonlinear relationships, making them increasingly valuable in geographical research. Mechanistic models simulate crop growth processes based on physiological characteristics, considering photosynthesis, respiration, and environmental factors such as temperature, precipitation, and soil fertility, then integrate remote sensing data with crop models for yield prediction.

Recent studies demonstrate the effectiveness of these approaches. Bolton and Friedl employed the two-band enhanced vegetation index (EVI2) and normalized difference water index (NDWI) to estimate corn and soybean yields in the central United States, finding highest correlations 65–75 days after crop green-up. Leroux combined MODIS NDVI, land surface temperature (LST), and SARRA crop model simulations to develop empirical statistical models for yield estimation in Africa' s Sahel region, showing that combined indices outperformed NDVI alone. Guo Rui used EVI to estimate winter wheat yields across various scales in Shandong Province, achieving accuracies no lower than 89.41%. An Qin compared multiple models in Changchun and found neural networks superior in stability and accuracy. Zeng Yan employed VTCI and LAI with support vector regression to estimate winter wheat yields in Guanzhong Plain, achieving R^2 values of 0.94. Huang Jianxi assimilated LAI and evapotranspiration data into the DSSAT crop model, improving accuracy by 8.2%.

Despite these advances, few studies compare multiple machine learning mod-

els with conventional methods for spring wheat yield estimation in northern Kazakhstan. This research addresses this gap by analyzing optimal prediction timing and vegetation indices using MODIS MOD09A1 and MOD16A2 products to calculate EVI and crop water stress index (CWSI). We employ linear regression, random forest, neural networks, and support vector machines to compare the yield estimation capabilities of eight vegetation indices, aiming to provide management guidance for local spring wheat production.

---

2. Study Area and Methods

2.1 Study Area Overview

Northern Kazakhstan includes North Kazakhstan, Qostanay, and Aqmola (Figure 1). Located between 49°09 -55°45 N and 61°30 -79°30 E, the region features a temperate continental climate with concentrated summer precipitation and long, cold winters. Winter average temperatures range from -15 to -20°C, dropping to -30°C, while summer averages reach 18-25°C. Annual precipitation varies by state: 353 mm in North Kazakhstan, 381 mm in Aqmola, and 407 mm in Qostanay. The Ishim and Tobol rivers flow through the region, which contains numerous freshwater lakes. Dominant land cover types include grassland (40%), cropland (35%), built-up areas, forest, bare land, and water bodies. The fertile chernozem and brown soils support flat terrain, making this a critical global wheat export region. However, rain-fed agriculture makes production highly vulnerable to precipitation variability.

2.2 Data Sources and Processing

Northern Kazakhstan lacks comprehensive meteorological and detailed soil data, limiting the application of crop growth models. Therefore, we employed empirical models based on remote sensing data. Cropland vector data for northern Kazakhstan were provided by the CASEarth project (http://data.casearth.cn/), produced by the remote sensing research team at the Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences. This dataset, derived from Landsat imagery using object-oriented classification with segmentation, decision tree classification, and change detection, achieves classification accuracies exceeding 90%.

We utilized MODIS MOD09A1 surface reflectance data (500 m resolution) to calculate vegetation indices. Yield estimation employed MOD15A2H LAI products for indirect validation. All MODIS data were processed on the Google Earth Engine platform, with quality control to extract clear pixels and minimize cloud contamination effects.

2.3 Vegetation Index Calculation

Eight frequently used vegetation indices for crop yield and biomass estimation were selected for analysis (Table 1). Spring wheat in northern Kazakhstan is sown in early May and harvested in early September. MOD09A1 data were downloaded for the critical growth period (May–September), and MOD15A2H LAI products were used for dynamic monitoring. The indices include: Normalized Difference Vegetation Index (NDVI), Two-band Enhanced Vegetation Index (EVI2), Wide Dynamic Range Vegetation Index (WDRVI), Saturation-adjusted NDVI (SANDVI), Green Wide Dynamic Range Vegetation Index (WDRVIgreen), Green Chlorophyll Index (CIgreen), Difference Vegetation Index (DVI), Optimized Soil-Adjusted Vegetation Index (OSAVI), and Normalized Multiband Drought Index (NMDI).

2.4 Model Development and Validation

To determine optimal vegetation indices and prediction timing, we calculated indices for the three states and fitted them to actual spring wheat yields using univariate linear regression. The coefficient of determination (R^2) and root mean square error (RMSE) evaluated model performance:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

where y_i and $\hat{y}_i$ are observed and predicted yields, and N is the number of years.

Given limited sample sizes, we employed Bootstrap resampling—randomly selecting N samples with replacement to create training sets, using unsampled data for validation. This process was repeated 1000 times to optimize model parameters. The selected optimal indices and timing served as inputs for four models: linear regression, random forest, support vector machine (linear kernel), and bidirectional recurrent neural network. For random forest, we optimized the number of regression trees rather than variable count (single variable input). For SVM, we trained the cost function. For the neural network, we adjusted neuron numbers using validation dataset RMSE to determine optimal parameters.

3. Results

3.1 Optimal Vegetation Indices and Prediction Timing

Figure 2 shows temporal variation in R^2 between vegetation indices and spring wheat yield. The normalized multi-band drought index (NMDI) performed poorly across all three states. In North Kazakhstan, R^2 values initially increased then decreased, peaking on July 12 with CIgreen ($R^2 = 0.51$, $P < 0.05$). In Aqmola, R^2 increased rapidly after July 12, peaking on August 5 with OSAVI ($R^2 = 0.56$, $P < 0.05$). In Qostanay, most indices showed strong correlations ($R^2 \geq 0.5$, $P < 0.05$) around July 12, with WDRVIgreen performing best ($R^2 = 0.53$, $P < 0.05$).

Table 2 summarizes the optimal indices and dates: North Kazakhstan—CIgreen on July 12 ($R^2 = 0.51$, RMSE = 131.8 $kg \cdot hm^{-2}$); Aqmola—WDRVIgreen on August 5 ($R^2 = 0.56$, RMSE = 143.0 $kg \cdot hm^{-2}$); Qostanay—WDRVIgreen on July 12 ($R^2 = 0.53$, RMSE = 135.5 $kg \cdot hm^{-2}$). These periods correspond to the critical yield formation stage from heading to maturity.

3.2 Model Performance and Spatial Distribution

Using the optimal indices and timing, we estimated spring wheat yields across the three states (Figure 3). Natural breaks classification in ArcGIS revealed similar spatial patterns among models. In North Kazakhstan, linear regression, SVM, and neural networks produced comparable distributions, while random forest showed more dispersed high-yield areas and weaker clustering. All models indicated lower yields in the southeast and higher yields in the north-central region.

In Aqmola, SVM exhibited fragmented patterns in southern areas with minimal regional yield differences. The other three models showed similar distributions, with higher yields in the north and lower yields in the east and south. In Qostanay, all four models demonstrated high similarity in spatial distribution, with higher yields in the north and lower yields in the south. SVM again showed smaller regional differences.

Overall, despite performance variations, all models consistently identified high- and low-yield zones. Random forest and SVM exhibited limitations: random forest tended to overestimate or underestimate yield ranges, while SVM produced smaller spatial yield variations due to its reliance on support vectors for hyperplane construction.

3.3 Yield Estimation Accuracy Assessment

We validated results using MOD17A3HGF NPP products, which serve as important biomass indicators. Correlation analysis between estimated yields and NPP (Figure 4) showed varying performance (Figure 5). In North Kazakhstan, regression, SVM, and neural networks performed better than random forest,

which showed only 29% significant correlation ($R^2 > 0.5$) and 18% high correlation ($R^2 > 0.8$). Significant correlations were concentrated in western and southeastern areas, while low correlations dominated the north—a high-yield region where MODIS data quality may be compromised by high rainfall or mixed cropping patterns (wheat accounts for only 35% of crops in North Kazakhstan versus >50% in other states).

In Aqmola, all models except SVM showed strong validation results, with neural networks performing best (94% significant/high correlation), followed by regression (90%) and random forest (78%). SVM' s smaller yield differences reduced correlation strength. In Qostanay, regression, SVM, and neural networks achieved high accuracy, while random forest performed poorly (31% weak correlation), particularly in southern areas.

These results demonstrate that with limited sample sizes, bidirectional recurrent neural networks and regression models offer greater reliability. While regression saves parameter optimization time and maintains robustness, its transferability is limited. Neural networks slightly outperform other machine learning models overall.

---

4. Discussion

Remote sensing yield estimation establishes model systems linking crop factors to production. While multi-factor models incorporating temperature, precipitation, and soil fertility can improve accuracy, they require large samples and risk multicollinearity issues. With small samples, multi-factor combinations may reduce precision and increase bias. Our single-index approach identifies optimal vegetation indices and timing for each state, revealing that correlations strengthen during the heading-to-maturity period (July 12–August 5), the critical yield formation stage.

Model comparisons show consistent spatial patterns for high- and low-yield zones across methods. However, random forest tends to overestimate yield ranges due to bootstrap sampling limitations with small datasets, where similar regression trees emerge and extreme yields may be undersampled. SVM' s structural risk minimization and reliance on support vectors can create hyperplanes that reduce spatial yield variability. Neural networks and regression models prove more stable and reliable under data constraints.

Validation using NPP products reveals data quality issues in high-rainfall northern areas of North Kazakhstan and Qostanay, where MODIS products may contain errors. Additionally, North Kazakhstan' s diverse cropping structure (only 35% wheat) complicates pure wheat yield estimation, unlike Aqmola and Qostanay where wheat exceeds 50% of plantings. This explains lower validation accuracies in North Kazakhstan.

---

5. Conclusion

This study analyzed optimal vegetation indices and prediction timing for spring wheat yield estimation in northern Kazakhstan. Except for NMDI, vegetation indices showed strong correlations with yield primarily between July 12 and August 5, coinciding with the critical yield formation period from heading to maturity. The optimal indices are CIgreen for North Kazakhstan, and WDRVI-green for both Aqmola and Qostanay.

Model comparisons indicate high spatial consistency in identifying yield zones. In North Kazakhstan, regression, SVM, and neural networks outperform random forest, though all models show relatively lower accuracy due to mixed cropping patterns and potential MODIS data quality issues in high-rainfall northern regions. In Aqmola and Qostanay, neural networks perform best, followed by regression models, while SVM and random forest show lower precision.

Overall, the bidirectional recurrent neural network provides the most accurate spring wheat yield estimation for northern Kazakhstan, particularly in Aqmola and Qostanay, offering valuable support for regional food security management.

References

[1] Wang Kaining, Wang Xiuxin. Research on winter wheat yield estimation with multiple remote sensing vegetation index combinations[J]. Journal of Arid Land Resources and Environment, 2017, 31(7): 44-49.

[2] Li Ning. Study on Kazakhstan' s wheat industry and its problems[J]. Grain Distribution Technology, 2013(2): 37-42.

[3] Ma Jun, Gong Xinshu. Research on food security in Central Asian countries[J]. World Agriculture, 2014(8): 22-26.

[4] Zhou Qingbo. Status and tendency for development in remote sensing of agriculture situation[J]. Journal of China Agricultural Resources and Regional Planning, 2004, 25(5): 12-17.

[5] Anup K P, Lim C, Ramesh P S, et al. Crop yield estimation model for Iowa using remote sensing and surface parameters[J]. International Journal of Applied Earth Observation and Geoinformation, 2006, 8(1): 26-33.

[6] Michael S R. Assessment of millet yields and production in northern Burkina Faso using integrated NDVI from the AVHRR[J]. International Journal of Remote Sensing, 1992, 13(18): 3871-3879.

[7] Michael S R. Operational yield forecast using AVHRR NDVI data: Reduction of environmental and inter-annual variability[J]. International Journal of Remote Sensing, 1997, 18(5): 1059-1077.

[8] Jiao Xianfeng, Yang Bangjie, Pei Zhiyuan, et al. Monitoring crop yield using NOAA/AVHRR based vegetation indices[J]. Transactions of the Chinese Society of Agricultural Engineering, 2005, 21(4): 104-108.

[9] Tuvdendorj B. Determination of appropriate remote sensing indices for spring wheat yield estimation in Mongolia[J]. Remote Sensing, 2019, 11(21): 2568-2570.

[10] Li Junling, Guo Qile, Peng Jiyong. Remote sensing estimation model of Henan Province winter wheat yield based on MODIS data[J]. Ecology and Environmental Sciences, 2012, 21(10): 1665-1669.

[11] Uno Y. Artificial neural networks to predict corn yield from compact airborne spectrographic imager data[J]. Computers and Electronics in Agriculture, 2005, 47(2): 149-161.

[12] Bolton D K, Friedl M A. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics[J]. Agricultural and Forest Meteorology, 2013, 173: 74-84.

[13] Leroux L. Crop monitoring using vegetation and thermal indices for yield estimates: Case study of a rainfed cereal in semi-arid West Africa[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2016, 9(1): 347-362.

[14] Guo Rui. Monitoring and forecasting method of winter wheat yield in Shandong Province[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(7): 156-163.

[15] An Qin. Research on methods of maize yield estimation by remote sensing in Changchun region[D]. Changchun: Jilin University, 2018.

[16] Zeng Yan, Wang Di, Zhao Xiaojuan, et al. Study on yield prediction of winter wheat in Guanzhong Plain based on SVR[J]. China Agricultural Informatics, 2019, 31(6): 10-20.

[17] Huang Jianxi, Ma Hongyuan, Tian Liyan, et al. Comparison of remote sensing yield estimation methods for winter wheat based on assimilating time sequence LAI and ET[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(4): 197-203.

[18] Tan Changwei. Comparison of the methods for predicting wheat yield based on satellite remote sensing data at anthesis[J]. Scientia Agricultura Sinica, 2017, 50(16): 3101-3109.

[19] Ma Hongyuan, Huang Jianxi, Huang Hai, et al. Ensemble forecasting of regional yield of winter wheat based on WOFOST model using historical meteorological dataset[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(9): 257-266.

[20] Yang Lianbing, Zheng Hongwei, Luo Geping, et al. Retrieval of soil salinity content based on BP neural network optimized by genetic algorithm[J]. Geography and Geo-Information Science, 2021, 37(2): 21-37.

[21] Gitelson A A. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation[J]. Journal of Plant Physiology, 2004, 161(2): 165-173.

[22] Gu Y. NDVI saturation adjustment: A new approach for improving cropland performance estimates in the Greater Platte River Basin, USA[J]. Ecological Indicators, 2013, 30: 1-6.

[23] Gitelson A A. Remote estimation of crop gross primary production with Landsat data[J]. Remote Sensing of Environment, 2012, 121: 197-207.

[24] Gitelson A A. Remote estimation of canopy chlorophyll content in crops[J]. Geophysical Research Letters, 2005, 32(8): 1-4.

[25] Mirchooli F. Spatial distribution dependency of soil organic carbon content to important environmental variables[J]. Ecological Indicators, 2020, 116: 1-5.

[26] Han Qifei, Luo Geping, Bai Jie, et al. Characteristics of land use and cover change in Central Asia in recent 30 years[J]. Arid Land Geography, 2012, 35(6): 909-918.

[27] Yao F. Estimation of maize yield by using a process based model and remote sensing data[J]. Physics and Chemistry of the Earth, 2015, 87: 142-152.

[28] Mkhabela M S, Mashinini N N. Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA AVHRR[J]. Agricultural and Forest Meteorology, 2005, 129(1): 1-9.

[29] Tian Yanjun, Shi Ying, Shuai Yanmin, et al. Land cover information retrieval from temporal features based on remote sensing images[J]. Arid Land Geography, 2021, 44(2): 450-459.

[30] Wang Hua, Yang Qianpeng, Tian Yunjie, et al. Vegetation coverage monitoring in Central Asian countries using multi-temporal Landsat images[J]. Arid Land Geography, 2020, 43(4): 1023-1032.

[31] Xu Xingang, Wu Bingfang, Meng Jihua, et al. Research advances in crop yield estimation models based on remote sensing[J]. Transactions of the Chinese Society of Agricultural Engineering, 2008(2): 290-298.

[32] Jiang Z. Development of a two-band enhanced vegetation index without a blue band[J]. Remote Sensing of Environment, 2008, 112(10): 3833-3845.

[33] Dara A. Mapping the timing of cropland abandonment and recultivation in northern Kazakhstan using annual Landsat time series[J]. Remote Sensing of Environment, 2018, 213: 49-60.

[34] Guo H. Determining variable weights for an optimal scaled drought condition index (OSDCI): Evaluation in Central Asia[J]. Remote Sensing of Environment, 2019, 231: 1-17.

[35] Yin H. Monitoring fire regimes and assessing their driving factors in Central Asia[J]. Journal of Arid Land, 2021, 13(5): 500-515.

[36] Gao Yukun. Aboveground forest biomass estimation based on machine learning algorithms and multi-source data in a typical subtropical region[D]. Hangzhou: Zhejiang Agriculture and Forestry University, 2018.

[37] Xu Yi, Dong Xuanyan, Wang Junjie. Use of remote multispectral imaging to monitor chlorophyll a in Taihu Lake: A comparison of machine learning models[J]. Journal of Hydroecology, 2019, 40(4): 48-57.

Note: Figure translations are in progress. See original paper for figures.