Postprint: Forecasting Grain Protein Content in Major Winter Wheat Production Regions Based on Remote Sensing and Meteorological Data

Authors: Wang Lin, Liang Jian, Meng Fanyu, Meng Yang, Zhang Yongtao, Li Zhenhai

Date: 2023-02-17T00:00:00+00:00

Abstract

Research on monitoring and forecasting wheat grain protein content is of great significance for guiding farmers in optimizing cultivation practices, enterprises in classification, storage, and procurement, wheat futures pricing, and import policy adjustments. This study, focusing on the main winter wheat producing regions (Henan, Shandong, Hebei, Anhui, and Jiangsu provinces), developed a multilayer linear prediction model for winter wheat grain protein content and generated forecasts for the 2019 season. To address biases in inter-annual and spatial scaling of prediction models, the protein content estimation model incorporated meteorological factors (temperature, precipitation, radiation), winter wheat gluten type, and heading-flowering stage Enhanced Vegetation Index (EVI). Results demonstrate that the model integrating three meteorological factors achieved higher accuracy in both the training set (R2 = 0.39, RMSE = 1.04%) and validation set (R2 = 0.43, RMSE = 0.94%) compared to models integrating two factors or a single factor. The protein content estimation model was applied to remote sensing data across the main winter wheat producing areas to generate a 2019 quality forecast map, which was further developed into a thematic map of winter wheat quality distribution for the Huang-Huai-Hai region. These results can provide data support for subsequent wheat planting zoning and the achievement of green, high-yield, high-quality, and efficient grain production.

Full Text

Estimating Grain Protein Content of Winter Wheat in Producing Areas Based on Remote Sensing and Meteorological Data

WANG Lin[1,2], **LIANG Jian**[3], **MENG Fanyu**[4], **MENG Yang**[1,2], **ZHANG Yongtao**[5], **LI Zhenhai**[1,2*]

[1]Key Laboratory of Quantitative Remote Sensing in Ministry of Agriculture and Rural Affairs/Beijing Research Center for Information Technology in Agriculture, Beijing 100097, China
[2]National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
[3]National Agro-tech Extension and Service Center, Beijing 100125, China
[4]Beijing Agriculture Technology Extension Station, Beijing 100029, China
[5]Jiangsu Nonidt Agricultural Science and Technology Co. Ltd, Nanjing 210001, China

Abstract: With increasing socioeconomic development and rising living standards, consumer demands for agricultural products have shifted from "whether they exist"and "whether there is enough"to "whether they are good"and "whether they are high-quality." However, Chinese agriculture still faces challenges of being large but not strong, and abundant but not optimal. In 2019, the Ministry of Agriculture and Rural Affairs, National Development and Reform Commission, Ministry of Science and Technology, and other agencies jointly released the *National Quality-Oriented Agriculture Strategy Plan (2018–2022)*, which proposed development goals such as "shifting agriculture from production-oriented to quality-oriented" and "pursuing a path of quality-driven agricultural development." Grain protein content (GPC) is a key indicator for wheat quality assessment, directly determining wheat purchase price, processing suitability, and utilization value. Therefore, monitoring and forecasting wheat GPC is crucial for guiding farmers in optimized cultivation, enterprises in classified harvesting and storage, futures pricing, and import policy adjustments.

This study focused on the main winter wheat producing areas (Henan, Shandong, Hebei, Anhui, and Jiangsu provinces) and constructed a hierarchical linear prediction model for winter wheat GPC, achieving pre-harvest GPC forecasts for the 2019 season. To address biases in interannual and spatial extrapolation of prediction models, the GPC estimation model incorporated meteorological factors (temperature, precipitation, radiation), wheat gluten type, and the Enhanced Vegetation Index (EVI) during the heading-flowering stage. Results demonstrated that the GPC estimation model integrating three meteorological factors achieved higher accuracy in both the calibration set (R^2 = 0.39, RMSE = 1.04%) and validation set (R^2 = 0.43, RMSE = 0.94%) compared with models incorporating two meteorological factors or a single factor alone. The GPC estimation model was then applied to regional remote sensing estimation across

the main winter wheat producing areas, generating a 2019 quality forecast map and a thematic map of wheat quality distribution in the Huang-Huai-Hai region. These results provide data support for subsequent wheat planting zoning and the achievement of green, high-yield, high-quality, and efficient grain production.

Keywords: winter wheat; grain protein content (GPC); remote sensing; hierarchical linear model (HLM); meteorological data

1. Introduction

Research on remote sensing monitoring and forecasting of crop grain protein content (GPC) has been explored and preliminarily applied, with methods generally categorized into four types: (1) Empirical models based on the "remote sensing information–GPC" approach, which directly construct statistical models by analyzing remote sensing information during critical growth periods (sensitive bands, vegetation indices, red-edge parameters, etc.) [3–7]; (2) Quantitative models based on the "remote sensing information–agronomic parameters–GPC" approach, which build GPC prediction models through quantitative relationships between remote sensing data and key agronomic parameters, and between those parameters and GPC [8–10]; (3) Semi-mechanistic models integrating remote sensing data and ecological factors, which consider crop nitrogen translocation mechanisms and ecological factor effects on GPC, improving model interannual and spatial transferability [11–14]; and (4) Mechanistic models combining remote sensing information with crop growth models, which comprehensively consider various ecological factors during grain protein formation and use assimilation methods to minimize errors between simulated and observed variables, thereby adjusting model initial parameters and state variables to predict GPC [14,15].

Among these approaches, the first two categories have been most studied due to their simplicity and ease of implementation. However, they lack mechanistic interpretability and exhibit large biases when applied across regions and years. Assimilation models, while comprehensive, involve excessive input variables, complex algorithms, localization challenges, and high computational demands, limiting their regional applicability [16]. Li et al. [12] and Xu et al. [13] addressed environmental and varietal factors affecting quality forecasting by introducing Hierarchical Linear Models (HLM) to interpret the nested relationships among GPC, remote sensing, and environmental variables, constructing heading-stage prediction models that effectively reduced interannual and spatial extrapolation biases. Therefore, developing a semi-mechanistic GPC prediction method combining remote sensing information (reflecting crop nutrient status) with environmental factors (reflecting spatiotemporal variation) is feasible for building universally applicable, spatiotemporally scalable models.

Building on previous hierarchical remote sensing models for winter wheat quality, this study attempted to construct a GPC model across the main national winter wheat producing areas (Henan, Shandong, Hebei, Anhui, and Jiangsu provinces) to enable pre-harvest quality prediction.

2. Materials and Methods

2.1 Study Area Data collection was conducted across five major winter wheat producing provinces (Henan, Shandong, Hebei, Anhui, and Jiangsu) during the 2008, 2009, and 2019 growing seasons. The dataset comprised 898 samples, which reduced to 864 after outlier removal (200 samples in 2008, 283 in 2009, and 381 in 2019). Sample distribution is shown in [Figure 1: see original paper]. The dataset included 86 strong-gluten samples, 249 mixed strong-medium gluten samples, 380 medium-gluten samples, 2 mixed medium-weak gluten samples, and 147 weak-gluten samples.

2.2 Data Collection

2.2.1 Grain Protein Content Measurement At winter wheat maturity, samples were collected from farmers' fields with recorded coordinates. Three 1 m^2 sampling points were harvested per farmer, then naturally dried, weighed, and averaged. GPC (14% moisture basis) was measured using a FOSS Infratec™ 1241 near-infrared grain analyzer (Tecator, Höganäs, Sweden).

2.2.3 EOS/MODIS Remote Sensing Imagery Using the heading-flowering stage as the prediction window, this study employed the Enhanced Vegetation Index (EVI) derived from Moderate-resolution Imaging Spectroradiometer (MODIS) imagery for model development [13]. EVI was calculated as:

$$EVI = 2.5 \times \frac{R_{Nir} - R_{Red}}{R_{Nir} + 6 \times R_{Red} - 7.5 \times R_{Blue}} \tag{1}$$

where R_{Nir}, R_{Red}, and R_{Blue} represent reflectance in the near-infrared, red, and blue bands, respectively. To eliminate phenological differences and cloud cover effects, MODIS data from April–May in 2019 and 2020 were acquired, with maximum pixel values extracted to represent the heading-flowering stage.

2.2.4 Meteorological Data Meteorological raster data for the study area were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) (http://www.ecmwf.int/). Three meteorological factors were selected: monthly accumulated temperature (°C · day), monthly solar radiation (MJ/m^2), and monthly total precipitation (mm) for April. These factors were chosen because the period from jointing to flowering (approximately one month) encompasses critical growth stages (jointing, flag leaf, heading, and flowering) that determine canopy establishment, grain number, and nutrient accumulation. Data processing was performed using Python 3.7 (Python Software Foundation, Portland, USA) and MATLAB 2014 (MathWorks Inc., Natick, USA).

2.3 Winter Wheat Quality Forecasting Model During sample collection, cultivar information was recorded and used to determine strong/medium/weak gluten characteristics. For regional forecasting, considering cultivar diversity and complexity, we referenced the *China Wheat Quality Zoning Scheme* released by the Ministry of Agriculture in 2001. Using cultivar promotion data compiled by the National Agro-tech Extension and Service Center and county-level main cultivar information, we determined the gluten type characteristics for each county, ultimately creating a county-scale winter wheat quality zoning map.

The Hierarchical Linear Model (HLM) is a least-squares regression analysis that accounts for nested data structures (e.g., students nested within classes). Unlike ordinary least squares regression, HLM addresses data non-independence by stratifying datasets and analyzing relationships within (Level 1) and between (Level 2) layers [17]. For regional and interannual GPC modeling, environmental factors across regions and years (Level 2) affect the relationship between remote sensing information and GPC (Level 1), making HLM appropriate [12,13].

The two-level model structure was as follows:

Level 1 Model (based on GPC, vegetation index, and gluten type):

$$GPC_{ij} = \beta_{0j} + \beta_{1j} \cdot EVI_i + \beta_{2j} \cdot Glu + r_{ij} \tag{2}$$

where GPC_{ij} is grain protein content, Glu is the gluten type value (strong = 1, medium-strong = 2, weak = 3, with mixed zones assigned 1.5 and 2.5) based on Li et al. [12], EVI_{ij} is the Enhanced Vegetation Index, β_{0j}, β_{1j}, and β_{2j} are intercept, EVI coefficient, and gluten type coefficient, respectively, and r_{ij} is random error.

Level 2 Model (relating Level 1 coefficients to normalized meteorological data):

$$\beta_{nj} = \gamma_{n0} + \gamma_{n1} \cdot nRad + \gamma_{n2} \cdot nTem + \gamma_{n3} \cdot nPre + \mu_{nj} \tag{3}$$

where $nRad$, $nTem$, and $nPre$ are normalized monthly solar radiation, accumulated temperature, and precipitation, respectively; β_{nj} corresponds to Level 1 intercept, EVI coefficient, and gluten type coefficient; γ_{n0} is the Level 2 intercept; γ_{n1}, γ_{n2}, and γ_{n3} are coefficients for radiation, temperature, and precipitation; and μ_{nj} is random error. Normalization followed equation (4) (using radiation as example):

$$nRad = \frac{Radi - Rad_{min}}{Rad_{max} - Rad_{min}} \tag{4}$$

where $Radi$, Rad_{max}, and Rad_{min} are monthly solar radiation, maximum radiation, and minimum radiation (MJ/m^2) for each pixel. Different meteorological factor combinations were tested to determine the optimal set.

2.4 Statistical Analysis After outlier removal, 864 samples from 2008, 2009, and 2019 were randomly split, with 80% (691 samples) used for model calibration and 20% (173 samples) for validation. MATLAB 2014 was used for statistical analysis between predicted and measured GPC. Model accuracy and reliability were evaluated using coefficient of determination (R^2) and root mean square error (RMSE), where R^2 approaching 1 indicates better fit and lower RMSE indicates stronger predictive ability and stability. Dominance Analysis [18] was employed to quantify the relative contribution of each independent variable to GPC prediction.

3. Results

3.1 Winter Wheat Quality Zoning Analysis Referencing the *China Wheat Quality Zoning Scheme* and considering main cultivar distribution in the five provinces, we created a 2019 winter wheat quality zoning map ([Figure 2: see original paper]). Strong-gluten wheat was primarily cultivated in northern regions, including central-southern Hebei, northern Henan, and Linyi and Binzhou in Shandong. Medium-gluten wheat had the largest planting area in Shandong, with additional distribution in southern Henan, northern Jiangsu, and northern Anhui. Weak-gluten wheat was mainly distributed in southeastern coastal Jiangsu, central-southern Anhui, and southern Henan.

Quality zoning is influenced not only by ecological factors (meteorology and soil) but also by cultivar promotion and government policies. For example, in 2019, medium-gluten cultivars (Jimai 22, Luyuan 502, Shannong 28) accounted for 95.3% of total wheat area in Shandong, despite the region’ s suitability for strong-gluten wheat. Similarly, Hebei’ s *Strong-Gluten Wheat Industry Quality and Efficiency Improvement Plan (2019–2022)* designated specific counties for strong-gluten wheat promotion. Our zoning map was developed through quality zoning analysis and county-level main cultivar surveys, though incomplete county-level information necessitated using municipal-level data in many cases. Future research requires more precise cultivar distribution data to support accurate gluten type mapping for quality prediction models.

3.2 Meteorological Factor Selection for GPC Forecasting Model Various meteorological factor combinations were evaluated in the HLM framework (). Using a single factor, solar radiation ($nRad$) yielded the highest accuracy (R^2 = 0.31, RMSE = 1.12%), outperforming temperature ($nTem$) and precipitation ($nPre$) alone. Two-factor combinations improved accuracy marginally, with $nRad + nPre$ performing best (R^2 = 0.38, RMSE = 1.05%). The three-factor model achieved optimal performance (R^2 = 0.39, RMSE = 1.04%), confirming that GPC formation is jointly influenced by radiation, temperature, and precipitation, which exhibit substantial spatial variability across regions. Further research is needed on meteorological indicator selection and correlation analysis at different scales, particularly regarding spatial variation of meteorological factors.

Dominance Analysis revealed variable contributions to GPC prediction (). At the regional scale, gluten type (*Glu*) showed the highest relative importance (75.31%), as genetic characteristics fundamentally determine quality. Meteorological factors contributed in descending order: radiation (10.05%), temperature (7.61%), and precipitation (6.54%). Remote sensing information (EVI) contributed least. However, at the local scale (e.g., Hebei Province), where cultivar and meteorological variation decreased, *Glu* importance dropped to 48.22% while EVI importance increased to 22.65%, indicating that crop growth and nutrient status become more influential when environmental and genetic variability are reduced.

3.3 Winter Wheat GPC Forecasting Model The three-meteorological-factor model was selected as the regional forecasting model. [Figure 3: see original paper] shows the relationship between measured and estimated GPC for calibration and validation sets. The calibration set achieved $R^2 = 0.39$ (RMSE = 1.04%), with more overestimation than underestimation, particularly for some 2009 samples from Beijing and Jiangsu, though most points clustered near the 1:1 line. The validation set yielded $R^2 = 0.43$ (RMSE = 0.94%). Combined data showed $R^2 = 0.40$ (RMSE = 1.03%). These results demonstrate the feasibility and stability of the hierarchical linear model combining remote sensing and meteorological data for regional GPC prediction.

3.4 Regional GPC Forecasting Applying the GPC model regionally produced the 2019 winter wheat GPC estimation map ([Figure 4: see original paper]), showing a general north-to-south quality gradient. Highest GPC (>14.6%) occurred in Hebei, northern Shandong, and central-northern Henan. Moderate GPC (12.3%-13.6%) characterized western Shandong, eastern and southwestern Henan, and northern Anhui/Jiangsu. Lowest GPC (<12.3%) appeared in southeastern coastal Jiangsu and southeastern Henan.

4. Discussion and Conclusion

Based on 864 winter wheat samples (2008, 2009, 2019), this study developed a hierarchical linear GPC estimation model for the main producing areas (Henan, Shandong, Hebei, Anhui, Jiangsu). The two-level model (Level 1: GPC, EVI, gluten type; Level 2: normalized meteorological data) achieved calibration accuracy of $R^2 = 0.39$ (RMSE = 1.04%) and validation accuracy of $R^2 = 0.43$ (RMSE = 0.94%). Regional application generated the 2019 quality forecast map, demonstrating good universality and spatiotemporal scalability.

This study utilized MODIS imagery (250 m spatial resolution) for large-scale application in the Huang-Huai-Hai region. However, the mixed-pixel problem in winter wheat area extraction was not thoroughly addressed. Future research should explore high-resolution imagery (Landsat-TM, Sentinel-2, or Chinese GF series satellites) for wheat quality forecasting. Additionally, sample representativeness could be improved, as data were collected through demonstration plots

(2008–2009) and farmer assistance (2019), potentially limiting spatial representativeness. While April meteorological data were used because the jointing-flowering stage is critical for canopy development and nutrient accumulation, future studies should integrate regional phenology extraction to more accurately capture key growth period meteorological data.

References

[1] WANG J, LI C, LIU L, et al. Estimating grain protein content in winter wheat with multi-temporal hyperspectral measurements[J]. Scientia Agricultura Sinica, 2008, 41(9): 2633-2640.

[2] LI S, TAN H, WANG K, et al. Research progress in wheat grain protein content monitoring using remote sensing[J]. Transactions of the CSAE, 2009, 25(2): 304-309.

[3] JIN X, XU X, FENG H, et al. Estimation of grain protein content in winter wheat by using three methods with hyperspectral data[J]. International Journal of Agriculture & Biology, 2014, 16(3): 498-504.

[4] ONOYAMA H, RYU C, SUGURI M, et al. Estimation of rice protein content before harvest using ground-based hyperspectral imaging and region of interest analysis[J]. Precision Agriculture, 2017, 19: 721-734.

[5] RODRIGUES F A, BLASCH G, DEFOURNY P, et al. Multi-temporal and spectral analysis of high-resolution hyperspectral airborne imagery for precision agriculture: Assessment of wheat grain yield and grain protein content[J]. Remote Sensing, 2018, 10: ID 930.

[6] XU X, LI C, DONG Y, et al. Estimating grain protein content in winter wheat with multi-temporal hyperspectral measurements[J]. Sensor Letters, 2014, 12: 1-5.

[7] BARMEIER G, HOFER K, SCHMIDHALTER U. Mid-season prediction of grain yield and protein content of spring barley cultivars using high-throughput spectral sensing[J]. European Journal of Agronomy, 2017, 90: 108-116.

[8] CHEN P, WANG J, PAN P, et al. Remote detection of wheat grain protein content using nitrogen nutrition index[J]. Transactions of the CSAE, 2011, 27(9): 75-80.

[9] QU S, LI Z, QIU C, et al. Remote sensing prediction of winter wheat grain protein content based on nitrogen nutrition index at anthesis stage[J]. Transactions of the CSAE, 2017, 33(12): 186-193.

[10] ZHAO H, SONG X, YANG G, et al. Monitoring of nitrogen and grain protein content in winter wheat based on sentinel-2A data[J]. Remote Sensing, 2019, 11(14): ID 1724.

[11] LI Z, XU X, JIN X, et al. Remote sensing prediction of winter wheat protein content based on nitrogen translocation and GRA-PLS method[J]. Scientia Agricultura Sinica, 2014, 47(19): 3780-3790.

[12] LI Z, TAYLOR J, YANG H, et al. A hierarchical inter-annual wheat yield and grain protein prediction model using spectral vegetative indices and meteorological data[J]. Field Crops Research, 2020, 248: ID 107711.

[13] XU X, TENG C, ZHAO Y, et al. Prediction of wheat grain protein by

coupling multisource remote sensing imagery and ECMWF data[J]. Remote Sensing, 2020, 12(8): ID 1349.
[14] LI Z, JIN X, ZHAO C, et al. Estimating wheat yield and quality by coupling the DSSAT-CERES model and proximal remote sensing[J]. European Journal of Agronomy, 2015, 71: 53-62.
[15] LI Z, WANG J, XU X, et al. Assimilation of two variables derived from hyperspectral data into the DSSAT-CERES model for grain yield and quality estimation[J]. Remote Sensing, 2015, 7(9): 12400-12418.
[16] LI Z, YANG G, WANG J, et al. Remote sensing of grain protein content in cereal: A review[J]. China Agricultural Informatics, 2018, 30(1): 46-54.
[17] FELDSTAIN A, WOLTMAN H, MACKAY J C, et al. An introduction to hierarchical linear modeling[J]. Tutorials in Quantitative Methods for Psychology, 2012, 8(1): 62-69.
[18] YE D, NG Y K, LIAN Y. Culture and happiness[J]. Social Indicators Research, 2015, 123(2): 519-547.

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