CNWheatGPC-500: the First 500-meter High-resolution Long-term Winter Wheat Grain Protein Content Dataset for China (2008-2019) from Multi-source Data

Xiaobin Xu[1], Lili zhou[1], Raffaele Casa[2], James Taylor[3], Hao Yang[5], Guijun Yang[4], Wenjiang Huang[5], Stefano Pignatti[6], Giovanni Laneve[7], Zhenhai Li[1]

1: Shandong University of Science and Technology, China, People's Republic of; 2: DAFNE, Università della Tuscia, Via San Camillo de Lellis, 01100 Viterbo, Italy; 3: UMRITAP, Montpellier SupAgro, Irstea, Univ. Montpellier, Montpellier 34000, France; 4: Key Laboratory of Quantitative Remote Sensing in Ministry of Agriculture and Rural Affairs, Information Technology Research Center, Bejing Academy of Agriculture and Forestry Sciences, Bejing 100097, China; 5: Institute of Crop Sciences, Chinese Academy of Agricultural Sciences/Key Laboratory of Crop Physiology and Ecology, Ministry of Agriculture and Rural Affairs, Beiing 100081, China; 6: Institute of Methodologies for Environmental Analysis (IMAA), National Council of Research (CNR), C. da S. Loja, 85050 Tito Scalo, Italy; 7: School of Aerospace Engineering (SIA), University of Rome "La Sapienza", SIA, via Salaria, 851, 00138 Roma, Italy

Abstract

In China, the exigency for precise wheat Grain Protein Content (GPC) data rises with growing food consumption demands and global market competition. However, due to the lack of extensive, prolonged high-resolution benchmark data, previous GPC studies have primarily focused on experimental fields, small geographic units, and limited temporal scopes. Additionally, the diverse geographical terrain in China exacerbates the challenges of large-scale GPC estimation. To address this challenge and the data gap, the first 500-meter spatial resolution, long-term winter wheat dataset covering major planting regions in China (CNWheatGPC-500) was created by integrating multi-source data from ERA5 and MODIS. The results demonstrate that the GPC estimation model based on HLM significantly outperformed other conventional models. The validation dataset exhibited an R^2 of 0.45 and an RMSE of 0.96%. In cross-validation, the RMSE values ranged from 0.90% in Gansu to 1.32% in Anhui. For leave-one-year-out cross-validation, the RMSE values ranged from 0.77% to 1.11%. CNWheatGPC-500 offers valuable insights for enhancing wheat production, quality control, and agricultural decision-making.

Introduction

Wheat is a vital staple crop globally, providing essential dietary calories and protein. China, as the largest producer and consumer of wheat, holds significant influence in the wheat market. Winter wheat, which comprises around 85% of China's total grain output during the summer harvest, is particularly important due to its high yield, protein content, and adaptability. In the face of climate change and geopolitical conflicts, timely and comprehensive information on winter wheat production is crucial for ensuring food security. Recent studies have emphasized the importance of predicting grain quality, with grain protein content (GPC) being a key quality trait. GPC significantly affects the nutritional and economic value of wheat, making its accurate estimation economically and practically significant.

Spatial monitoring of GPC has evolved from labor-intensive methods to remote sensing (RS)-based monitoring, offering early prediction capabilities. Various sensors, from handheld devices to satellite platforms, have emerged as efficient data acquisition tools for monitoring crop growth characteristics. Different environments exert different influences on the response of RS factors to GPC. However, due to insufficient spatial details of environmental factors and the significant impact of environmental interference on remote sensing data, it is necessary to combine multi-source data for GPC estimation.

Three primary categories of methods are used for predicting GPC: empirical methods, physically-based process models, and semi-mechanistic models. Empirical methods offer transparent models but may lack precision, while physically-based process models simulate crop growth considering various factors but require dense data input and parameter calibration. Semi-mechanistic models combine fundamental equations with empirical parameters to address uncertainties, showing promise in estimating agricultural parameters.

Despite advancements in RS technology and data availability in agriculture, large-scale and long-term datasets related to wheat GPC are still lacking. To address this gap, this study leveraged multi-source climate and RS data to develop a nationwide GPC estimation model based on the Hierarchical Linear Model (HLM). The model integrated meteorological datasets with maximum EVI and was rigorously assessed through extensive comparisons and cross-validation. The resulting CNWheatGPC-500 dataset provides valuable insights for sustainable agricultural development and food security.

Methods

➢ Wheat phenological data was paired with meteorological grids

First, wheat phenological data was paired with meteorological grids. This started by determining the day of year (DOY) corresponding to MA for each pixel within the meteorological grid. Subsequently, cumulative values for the respective meteorological variables were computed at 30-day intervals, over a total period of 90 days before MA, resulting in the generation of three effective cumulative values for each meteorological variable.

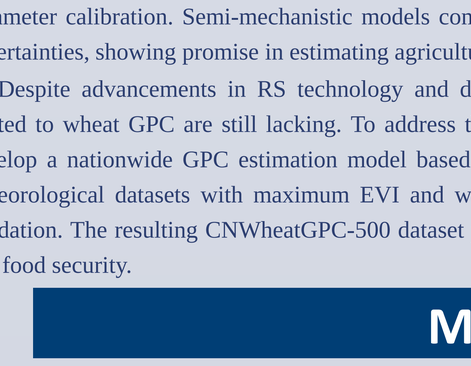

Figure 1. In response to distinct maturity (MA) observed among individual pixels, specific time durations were selected for synthesizing relevant features on a pixel-by-pixel basis.

➢ Development of hierarchical linear model (HLM)

HLM enables the partitioning of variability in nested data into two components: one arising from the individual level (i.e., how RS data responds to GPC), and the other stemming from the group level (i.e., how meteorological data influences the relationship between RS data and GPC). The general forms of the two-level relationships are presented as:

$$\text{Layer 1: } GPC_{ij} = \beta_{oj} + \beta_{1j} \times EVI + r_{ij},$$

where the GPC_{ij} is the grain protein content of an individual 1 within the population j, β_{0j} and β_{1j} represents intercept and slope, respectively. r_{ij} represents the random error. In this layer, the first linear structure of EVI response to wheat GPC is formed. The selected meteorological data has an impact on the relationship between wheat GPC and EVI, resulting in variations in slope and intercept:

$$\text{Layer 2: } \beta_{mj} = \gamma_{m0} + \sum_1^n(\gamma_{mn1}ET_n) + \sum_1^n(\gamma_{mn2}Tem_n) + \sum_1^n(\gamma_{mn3}Pre_n) + \sum_1^n(\gamma_{mn4}SR_n) + \mu_{mj},$$

where β_{mj} represents the β_0 and β_1 from the Level 1 model respectively, γ_{m0} is the intercept. γ_{mn1} to γ_{mn4} represent coefficient of each factor. The n values are 1, 2, and 3, representing the meteorological data synthesized for the n-th time interval (T1, T2 and T3). And μ_{mj} is the random effect of the Level-2, used to consider the correlation and variability between individuals within group.

➢ Intraclass Correlation Coefficient (ICC)

ICC serves as a crucial statistical metric for assessing the correlation among individual-level data within group-level data, the variability between different groups, and the effectiveness of hierarchical data. Its value ranges from 0 to 1, and a higher ICC value indicates a better fit to the characteristics of nested data, making it more suitable for HLM. The calculation is as follows:

$$ICC = \sigma_{\mu 0}^2 / \left(\sigma_{\mu 0}^2 + \sigma^2\right),$$

where σ^2 is the within-group variance, $\sigma_{\mu 0}{}^2$ is the between-group variance.

Based on the above methods, the workflow framework of this study is as follows:

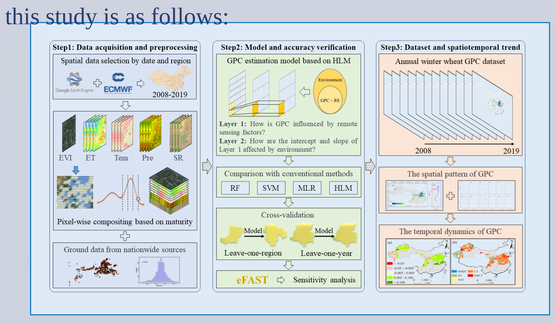

Figure 2. The workflow framework for generating CNWheatGPC-500.

Results

➢ **Inter-group variability in multisource data**

Intraclass Correlation Coefficient (ICC) was computed for various provinces and years(Figure 3). The results indicate significant differences across various years and provinces. Despite these discrepancies, the overall ICC values remained notably high, highlighting the multi-layered nature of spatial data, where a level of consistency exists within groups, but disparities between groups are pronounced.

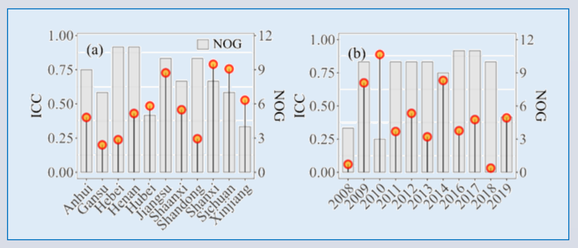

Figure 3. Regional variations in HLM: relationships between dependent and first-level independent variables with group effects.

➢ **Evaluating GPC estimation models**

The performance of the GPC estimation models on the validation dataset can be observed in Figure 4. The R2, RMSE, and nRMSE for the GPC estimated by RF compared to the measured GPC were 0.39, 0.99%, and 7.12%, respectively. Although it still delivered reasonable predictions, the performance of RF was comparatively diminished in the validation dataset. The performance of the SVM model has also declined, while the MLR model remains the poorest performing model. Notably, Figure 4 reveals the robust performance of the HLM in the validation dataset with an R2 of 0.45, RMSE of 0.96%, and nRMSE of 6.90%. Among all models, HLM stands out with the highest validation accuracy.

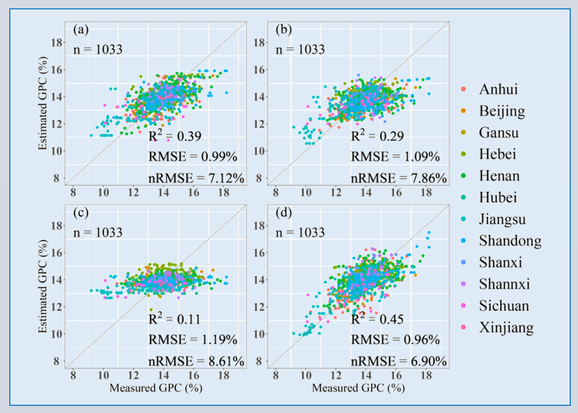

Figure 4. Performance comparison of winter wheat GPC estimation models based on RF (a), SVM (b), MLR (c), and HLM (d) in the validation dataset.

➢ Spatial pattern of winter wheat GPC in China

Figure 5 shows the spatial pattern of the annual average GPC value of wheat. Research illustrates substantial spatial heterogeneity in China‘s wheat GPC, with noticeably higher levels in northern regions compared to the southern ones, and a positive correlation between GPC and increasing latitude. However, an intriguing anomaly emerged in the southernmost region, encompassing select counties in Panzhihua City, situated outside the Sichuan Basin, where GPC levels were notably higher when compared to the more humid northern Sichuan Basin. Moving further north, GPC exhibited a gradual increase. Unlike the pronounced latitudinal variation observed in GPC, the variation in GPC with longitude did not display a discernible pattern. Additionally, GPC from various agricultural subregions underwent further statistical analysis with the boxplots of predicted GPC in shown Figure 5. Subregions B and C exhibited relatively higher GPC levels, while the southern subregions, D and E, displayed lower GPC levels.

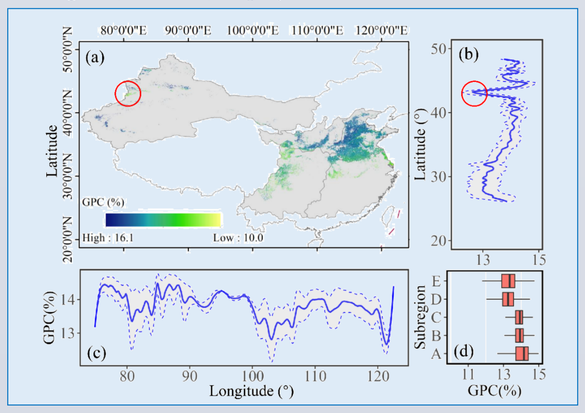

Figure 5. The spatial pattern (a) of the annual average GPC of winter wheat in China during 2008 to 2019 and its variation curve with latitude (b) and longitude (c). The curve is smoothed using Locally Estimated Scatterplot Smoothing (LOESS) with a smoothing window of 0.05. Panel (d) presents a box plot of predicted GPC by agricultural subareas, with values beyond the edges not displayed.

➢ **Temporal dynamics of GPC across diverse geographic regions**

The Sen's slope map in Figure 6 revealed a predominance of positive trends, with the Yangtze Middle-Lower Plain within subregion E exhibiting the most pronounced trend. The trend diagnosis map consistently illustrate significant spatial heterogeneity, showcasing the trend of higher GPC levels at higher latitudes and lower levels at lower latitudes. Unlike the pronounced latitudinal variation observed in GPC, the variation in GPC with longitude did not display a discernible pattern.

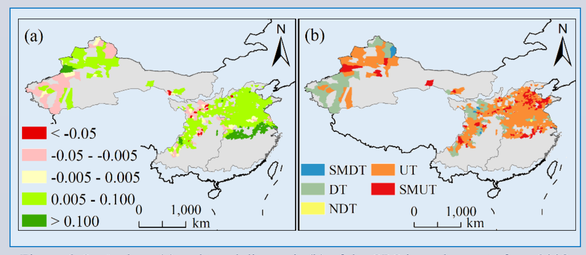

Figure 6. Sen’s slope (a) and trend diagnosis (b) of the GPC in each county from 2008 to 2019. SMUT represents significantly monotonic upward trend, UT represents upward trend, NDT represents no discernible trend, DT represents downward trend, and SMDT represents significantly monotonic downward trend.

Data Records

The first 500-meter spatial resolution, long-term winter wheat dataset covering major planting regions in China (CNWheatGPC-500) was created by integrating multi-source data from ERA5 and MODIS. Distributed under the Creative Commons Attribution 4.0 International license, the CNWheatGPC-500 dataset not only advances our understanding of winter wheat GPC in China but also facilitates research and analysis in an open and collaborative manner. The dataset is designated as YearCNWheatGPC-500, where "Year" represents the years spanning from 2008 to 2019. It encompasses a comprehensive 12-year dataset presented in TIF format. The CNWheatGPC-500 product generated in this study is available at https://doi.org/10.5281/zenodo.10066544. Kindly contact the authors for further inquiries and more detailed information.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (42271396), the Key Research and Development project of Shandong Province (LJNY202103) and the European Space Agency (ESA) and Ministry of Science and Technology of China (MOST) Dragon (57457).

Achievements

[1] Xiaobin Xu, Lili Zhou, James Taylor, et al. The first 500-meter, long-term winter wheat grain protein content dataset for China from multi-source data. Scientific Data, under review.

[2]李振海,许晓斌,许冀斌等.一种省域尺度的冬小麦籽粒蛋白质含量预测方法[P].山东省:CN117457066A,2024-01-26.