IMPROVE STATE-LEVEL WHEAT YIELD FORECASTS IN KAZAKHSTAN ON GEOGLAM'S EO DATA BY LEVERAGING A SIMPLE SPATIAL-AWARE TECHNIQUE

Anh Nhat Nhu
Department of Computer Science,
University of Maryland
NASA Harvest
College Park, MD 20742, USA
{anhu}@terpmail.umd.edu

Ritvik Sahapalj, Christina Justice, Inbal Becker-Reshef
Department of Geographical Science,
University of Maryland
NASA Harvest
College Park, MD 20742, USA
{ritvik,justicec,ireshef}@umd.edu

ABSTRACT

Accurate yield forecasting is essential for making informed policies and long-term decisions for food security. Earth Observation (EO) data and machine learning algorithms play a key role in providing a comprehensive and timely view of crop conditions from field to national scales. However, machine learning algorithms' prediction accuracy is often harmed by spatial heterogeneity caused by exogenous factors not reflected in remote sensing data, such as differences in crop management strategies. In this paper, we propose and investigate a simple technique called state-wise additive bias to explicitly address the cross-region yield heterogeneity in Kazakhstan. Compared to baseline machine learning models (Random Forest, CatBoost, XGBoost), our method reduces the overall RMSE by 8.9% and the highest state-wise RMSE by 28.37%. The effectiveness of state-wise additive bias indicates machine learning's performance can be significantly improved by explicitly addressing the spatial heterogeneity, motivating future work on spatial-aware machine learning algorithms for yield forecasts as well as for general geospatial forecasting problems.

1 INTRODUCTION

Accurate crop yield forecasts can benefit governments, policymakers, and individual farmers by providing better insights into various exogenous drivers that impact the agricultural markets. These insights can lead to earlier responses and better-informed decisions to improve food security at both regional and international scales (Becker-Reshef et al., 2022). Recently, machine learning algorithms have been applied on Earth Observation (EO) data and have shown a great potential to improve the reliability of these forecasts (Basso & Liu, 2019).

In this paper, we consider the use of EO data collected from the GEOGLAM Crop Monitor Ag-Met System (https://cropmonitor.org) and tree-based algorithms to directly forecast wheat yields in Kazakhstan, the 10^{th} largest wheat exporter in the world (FAO). A prominent challenge negatively impacting Machine Learning models' performance in forecasting yields is the spatial yield heterogeneity due to exogenous factors like local farming practices or crop varietals that are not reflected in remote sensing data. Lee et al. (2022) proposed to train a separate model for each province, successfully reducing the state-wise prediction errors. However, in our dataset, due to a very small amount of yield data available for each province (typically less than 20 data points), this approach results in highly unreliable and overfit models with error rates far exceeding those of baseline models, as shown in Figure 2. To improve upon this issue, we focus on reducing the errors, especially in provinces with the least accurate yield predictions, by using state-wise additive bias. First, we followed the methodologies in Sahajpal et al. (2020) to create features from EO data and investigate the performance of various baseline tree-based models, including XGBoost, CatBoost, and Random Forest, in forecasting wheat yields at the state level. Next, each state-wise additive bias was separately added to the model's predictions in each province to obtain the final yield forecast. This approach shows a remarkable increase in overall performance, with the most significant benefits

being seen in the province with the highest baseline yield errors (Almatinskaya). Furthermore, since state-wise bias adds no computational overhead during the inference process, this technique can be efficiently applied to improve yield predictions in other datasets.

2 DATA AND METHODS

2.1 COLLECTING AND EXTRACTING EO DATA

We use multiple EO predictors (https://cropmonitor.org/tools/agmet/) including crop phenological information derived from the MODIS NDVI that provides a proxy for crop vigor and phenology, MODIS Leaf Area Index (LAI), temperature, precipitation, SMAP soil moisture, and evaporative stress index (ESI). These inputs are subsets to cropped areas using a wheat crop mask for Kazakhstan. The EO products used here are complementary and capture different facets of crop response to abiotic factors (temperature, precipitation, solar radiation) and its variation by phenological growth stage and geography.

2.2 DATA PREPROCESSING

The EO dataset has daily data spanning from 2001 to 2020. We subset this data to the crop growth season (May - September). We use EO data (NDVI, growing degree days, daily minimum and maximum temperature, soil moisture, evaporative stress index, and precipitation) to as input features for training and evaluating machine learning models and to compute state-wise bias. We also include information on the previous season's yield and the average yield from the last 5 years as additional variables in the model. Overall, we have 75 samples for each province (15 years x 5 months in the growing season).

2.3 MODEL TRAINING AND EVALUATION

We trained and evaluated the effectiveness of the state-wise bias by applying this bias to the baseline tree models (XGBoost, CatBoost, and Random Forest) to forecast wheat yields at the state level in Kazakhstan. The state-wise bias is automatically calculated during the training process of each model. Algorithm 1 presents the complete training pipeline to train models and compute state-wise bias. We leave one year for testing, as suggested by Meroni et al. (2021), and split the remaining data into training (10 years) and validation sets (4 years) for model optimization. To maximize the amount of data used in the training process and increase the robustness of state-wise bias to unseen data, we employed the k-fold cross-validation method each test year (Dinh & Aires, 2022). In each fold, the error of each state was sampled using the corresponding validation set of the fold. The final state-wise bias of each state is the average of the recorded validation errors in all k folds.

The fundamental motivation for computing state-wise bias is that we observed baseline models are often biased toward values close to the mean yields, underestimating high yields in provinces with high productions, as discussed in Section 3.1. These high yields can be caused by factors typically not covered in satellite data, such as political and economical forces that allow some provinces to be the main wheat producer of the country. Although we have incorporated the regional information as categorical data in baseline models, the models still suffer from this bias. Therefore, state-wise bias is proposed as a simple yet effective technique to alleviate this spatial heterogeneity problem, resulting in a significant decrease in both MAPE and RMSE, as shown in Section 3

3 RESULTS AND ANALYSIS

3.1 MODEL PERFORMANCE

Overall, the MAPE and RMSE of XGBoost complemented with state-wise bias are **22.5%** and **0.095 Mg/ha**, respectively. Our model explains 57% of the yield variation in our dataset. Based on the scatter plot, we observed that the model performs well when the yields are average or low, but it consistently underestimates the yield by a large margin when yields are much higher (≥ 1.75 Mg/ha). Those high yields are often from provinces with the highest wheat production, such as Almatinskaya, or in exceptionally good years.

Algorithm 1 Model training and state-wise bias computation

Input Input features $\boldsymbol{X}$, targets $\boldsymbol{y}$
Output model f, state-wise bias b

$(\boldsymbol{X}_{train/val}, \boldsymbol{y}_{train/val}), (\boldsymbol{X}_{test}, \boldsymbol{y}_{test}) =$ split $\boldsymbol{X}, \boldsymbol{y}$

for each $(\boldsymbol{X}_{train}, \boldsymbol{y}_{train}), (\boldsymbol{X}_{val}, \boldsymbol{y}_{val}) \in$ k-fold split **do**
    Initialize model f
    Fit $f(\boldsymbol{X}_{train})$, $\boldsymbol{y}_{train}$
    Evaluate on $f(\boldsymbol{X}_{val})$, $\boldsymbol{y}_{val}$
    Done training model f ▷ End training model for current fold

    $\hat{y}_{val} = f(\boldsymbol{X}_{val})$
    for each state **do**
        state_bias = mean($\boldsymbol{y}_{val}\,[state] - \hat{\boldsymbol{y}}_{val}\,[state]$) ▷ state-wise bias for current fold
        b[state].append(state_bias)
    end for
end for ▷ End training k models on k folds

for each state **do** ▷ Final state-wise bias for each state
    b[state] = mean(b[state])
end for

Return model f, state-wise bias b

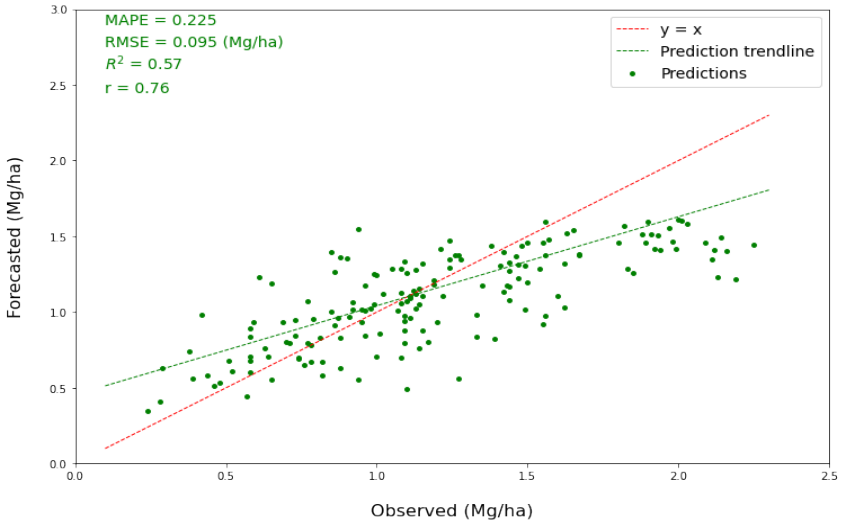

Figure 1: Scatter plot showing relationships between multi-year predicted and ground-truth yields of different provinces using leave-one-out year testing strategy.

3.1.1 Comparison to baseline models

To investigate the effect of state-wise bias, we test various models on different out-of-fold test years and compare the performance with and without state-wise bias. Our comparison involves both full dataset evaluation (national level) and evaluation by each province (regional level). Table 1 shows that the RMSE at the national level is decreased by **8.1%** to **9.76%**, resulting in an overall improvement over baseline models. The most significant improvements are observed in Almatinskaya (**24.04%** to **28.37%**) and Yujno-Kazachstanskaya (**6.95%** to **8.84%**) provinces, two provinces with the highest multi-year wheat yields and highest forecasting errors. Specifically, the average multi-year wheat yields of Almatinskaya and Yujno-Kazachstanskaya are 1.793 and 1.601 Mg/ha, respectively, while the national average yield is only 1.103 Mg/ha.

Table 1: Percentage change in RMSE of state-wise bias compared to the vanilla model (negative values represent improvements). The RMSE of each state is computed using all cross-year predictions for that state. We computed the RMSE's percentage change by subtracting the RMSE of vanilla models from the RMSE of state-wise bias, then dividing the result by the RMSE of vanilla models.

Province	XGBoost	CatBoost	Random Forest
Akmolinskaya	**-0.47%**	+1.13%	**-1.74%**
Aktyubinskaya	+1.54%	+1.35%	**-5.60%**
Almatinskaya	**-28.37%**	**-24.26%**	**-24.04%**
Jambylslkaya	+1.35%	+2.49%	+0.08%
Karagandinskaya	**-1.37%**	+0.55%	**-0.72%**
Kustanayskaya	**-3.85%**	**-2.64%**	**-3.55%**
Pavlodarskaya	**-0.65%**	+2.86%	**-2.18%**
Severo-Kazachstanskaya	+2.38%	+26.86%	**-15.42%**
Vostochno-Kazachstanskaya	**-4.48%**	**-4.11%**	**-1.17%**
Yujno-Kazachstanskaya	**-8.84%**	**-6.95%**	**-7.69%**
Zapadno-Kazachstanskaya	**-1.38%**	+0.14%	**-0.62%**
National	**-8.90%**	**-8.10%**	**-9.76%**

3.1.2 Comparison to region-specific models

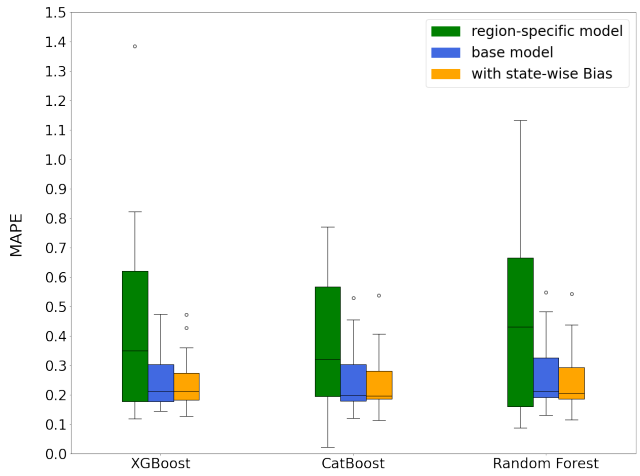

Figure 2: Comparison between wheat yield forecast errors of different models. Each MAPE in each boxplot represents the performance of each leave-one-out test year, spanning from 2006 to 2020.

Besides baseline models, we also compare our approach with region-specific models, an approach that has been used in several works to forecast crop yields Lee et al. (2022). Figure 2 shows that when a separate model was trained on each province, the MAPE has unusually high variance (green boxplot), ranging from **5%** to **110%** and having a median of **40%**. This is due to limited data available for each province (75 rows), causing a highly unstable training and serious overfitting issue for this approach. Therefore, although training a region-specific model achieved excellent performance when there are many data available, this approach is not suitable for our case. On the contrary, the performance consistently improves in all baseline models when state-wise bias is introduced (orange boxplot). Although the median MAPE was not improved by a remarkable margin (from **22%** to **21%**), the maximum and Q3 MAPEs are significantly decreased compared to those of the baseline models. Specifically, the maximum MAPE decreases from **48%** to **42%** for Random Forest, **46%** to **41%** for CatBoost, and **47%** to **37%** for XGBoost. These observations indicate that the proposed state-wise bias is most effective in difficult cases (cases with the highest errors) while having positive yet small impacts on easier predictions.

4 Conclusion and Future Work

Machine Learning models are frequently biased toward average yield in the dataset, resulting in higher errors for provinces with crop yields far from the mean, as shown in Figure 1 This issue is exacerbated by the spatial heterogeneity between different provinces/states. Our simple state-wise bias approach can alleviate the margin of errors in such cases by "debiasing" the errors computed separately for each province. This results in significant prediction error reduction, especially in provinces with high multi-year yields. Based upon these observations, we aim to further explore other more effective spatial-aware algorithms, such as unsupervised spatial clustering, that are robust to geospatial variations.

Acknowledgments

The authors acknowledge USAID grant 720BHA21IO00261 for funding this work, as well as the efforts of our partners in FAO.

References

Faostat: Fao statistical databases 2022.

Bruno Basso and Lin Liu. Seasonal crop yield forecast: Methods, applications, and accuracies. *advances in agronomy*, 154:201–255, 2019.

Inbal Becker-Reshef, Varaprasad Bandaru, Brian Barker, Sylvain Coutu, Jillian M Deines, Bradley Doorn, Gary Eilerts, Belen Franch, Antonio Sanchez Galvez, Mehdi Hosseini, et al. The nasa harvest harvest program on agriculture agriculture and food security food security harvest agriculture. In *Remote Sensing of Agriculture and Land Cover/Land Use Changes in South and Southeast Asian Countries*, pp. 53–80. Springer, 2022.

T. L. A. Dinh and F. Aires. Nested leave-two-out cross-validation for the optimal crop yield model selection. *Geoscientific Model Development*, 15(9):3519–3535, 2022. doi: 10.5194/gmd-15-3519-2022. URL https://gmd.copernicus.org/articles/15/3519/2022/.

Donghoon Lee, Frank Davenport, Shraddhanand Shukla, Greg Husak, Chris Funk, Laura Harrison, Amy McNally, James Rowland, Michael Budde, and James Verdin. Maize yield forecasts for sub-saharan africa using earth observation data and machine learning. *Global Food Security*, 33: 100643, 2022.

Michele Meroni, François Waldner, Lorenzo Seguini, Hervé Kerdiles, and Felix Rembold. Yield forecasting with machine learning and small data: What gains for grains? *Agricultural and Forest Meteorology*, 308-309:108555, 2021. ISSN 0168-1923. doi: https://doi.org/10.1016/j.agrformet.2021.108555. URL https://www.sciencedirect.com/science/article/pii/S0168192321002392.

Ritvik Sahajpal, Lucas Fontana, Pedro Lafluf, Guillermo Leale, Estefania Puricelli, Dan O'Neill, Mehdi Hosseini, Mauricio Varela, and Inbal Becker-Reshef. Using machine-learning models for field-scale crop yield and condition modeling in argentina. 2020. doi: 10.31223/x52595.