Ukrainian Journal of Remote Sensing
journal homepage: www.ujrs.org.ua

https://doi.org/10.36023/ujrs.2024.11.4.275

UDK 631.559:528.8

Yield prediction at field level

O. O. Kryvoshein [1, 2,]*, O. A. Kryvobok [1, 2], D. O. Zhylchenko [1]

[1] *Earth Observing System Data Analytics Inc 800 W. El Camino Real, Suite 180, Mountain View, CA 94040, USA*
[2] *Ukrainian Hydrometeorological Institute, 37 Nauki av., 03028, Kyiv, Ukraine*

Yield prediction at the field level is crucial for optimizing agricultural productivity and ensuring food security. This study analyzes the yield variability of maize, sunflower, and winter wheat across 481 agricultural fields in two regions of Ukraine (Kyiv and Cherkasy) over a three-year period (2020–2022). The objective was to explore the influence of environmental factors on crop yield predictions using satellite and weather data, sowing dates, and field area as predictors in a machine learning model. The study employed Random Forest model. Satellite data from Sentinel-2, including NDVI and LAI values, were used to assess crop conditions during the growing season. For each investigated year during the April-September period, focusing solely on the NDVI and LAI values for each month. Weather data, especially precipitation, was also examined but found to have limited predictive power due to the coarser spatial resolution of the gridded data (6.5 km), which cannot fully account for the local variations within each grid cell. As a result, despite the strong correlation between precipitation and yield at a broader scale (regional), weather data alone were not sufficient to accurately predict yield variability at the field level. The results showed that maize had the highest yield variability, while sunflower and winter wheat exhibited more stable yields. For maize, the model demonstrated strong predictive performance, with an R-squared of 0.8 and an RMSE of 1.5 t/ha. The most significant predictors were vegetation indices in August and sowing date. The normalized RMSE for maize was 20%. For sunflower, the model exhibited moderate accuracy, with an R-squared of 0.4 and an RMSE of 0.9 t/ha. Key predictors included the average LAI in May and July. However, the model's predictive power was limited, resulting in a normalized RMSE of 23%. Winter wheat showed similar performance to sunflower, with an R-squared of 0.35 and an RMSE of 0.9 t/ha. Due to higher average yields, the normalized RMSE for winter wheat was 15%. Overall, the study demonstrates varying levels of model accuracy across different crops, with maize achieving the best predictive performance. The results also emphasize the need for additional factors, such as soil properties, microclimates, and detailed field management practices, to improve predictive models at the field level.

Keywords: Yield, prediction, crops, field, satellite data, Machine Learning.

1. Introduction

Accurate yield prediction at the field level is pivotal for modern agriculture, as it influences key decisions related to resource management, market strategies, and food security. With the growing demand for efficient agricultural practices and the pressures of climate change, the need for precise and timely yield forecasts has never been greater. Yield prediction at the field level is critical for optimizing agricultural practices and managing resources. Accurate forecasting of crop yields enables farmers to make informed decisions about planting, fertilization, irrigation, and harvesting, thereby enhancing productivity and sustainability. Traditional yield prediction methods (Kryvobok, 2018), which often rely on historical yield data and agronomic models, have limitations in their ability to incorporate real-time environmental variability and spatial heterogeneity as well as the simple machine learning model, with just a single predictor, lacks the capability to deliver high-accuracy predictions (Kryvoshein, 2023).

Recent advances in remote sensing technology and machine learning offer new opportunities to enhance yield prediction accuracy. Liu and Zhang provide an overview of remote sensing methods used for crop yield estimation and forecasting, highlighting key technologies (Liu & Zhang, 2022). They also focus on the use of satellite data combined with machine learning models for wheat yield prediction, showcasing its practical applications (Liu, Zhang & Liu, 2021). Khan and Hameed review the role of machine learning approaches in crop yield prediction, identifying both progress and challenges in the field (Khan & Hameed, 2023). Miao, Liu and Wang explore deep learning techniques integrated with remote sensing for field-scale crop yield prediction (Miao, Liu & Wang, 2024). Miller and Scott discuss advancements in forecasting crop yields by combining satellite and ground-based data (Miller & Scott, 2023). Zhang and Li evaluate the effectiveness of multi-source data combined with machine learning techniques for improving crop yield prediction (Zhang & Li, 2021). Huang and Liu present a real-time crop yield prediction model using remote sensing and ensemble

*E-mail: aleksandr.krivoshein@eosda.com

learning methods (Huang & Liu, 2022). Desloires et al. focus on predicting out-of-year corn yields by integrating Sentinel-2 satellite imagery with machine learning methods, offering insights into accurate field-level yield forecasting (Desloires et al., 2023). In (Marszalek, Körner & Schmidhalter, 2022) explore the use of satellite and climatological data for predicting multi-year winter wheat yields at the field level, demonstrating the value of combining remote sensing and climate information for improved yield predictions. Additionally, Hosseini et al. investigate soybean yield prediction at the field scale using Sentinel-1 data, showcasing the potential of radar satellite data in crop yield forecasting (Hosseini et al., 2020). These studies highlight the growing role of satellite data in enhancing the accuracy and reliability of yield predictions in agriculture.

Machine learning techniques have proven to be highly effective in processing and analyzing these complex datasets. Algorithms such as regression models, decision trees, and neural networks can identify patterns and relationships between various input factors, improving the precision of yield forecasts (Khan & Hameed, 2023). For example, Liu et al. (2021) demonstrated how integrating satellite data with machine learning models significantly enhanced wheat yield predictions. Additionally, deep learning approaches have shown promise in processing high-resolution satellite images to make real-time yield forecasts (Chen & Zhang, 2022; Zhou & Liu, 2024).

The combination of remote sensing data with machine learning techniques has led to the development of sophisticated predictive models that address the limitations of traditional methods. These models incorporate diverse data sources, such as weather data, soil properties, and crop-specific information, to generate accurate field-level yield predictions (Hernandez & Ritchie, 2023). Studies by Kumar and Verma (2023) and Shao & Chen (2023) highlight how integrating multiple data sources and employing ensemble models can improve prediction accuracy and reliability.

However, challenges remain in applying these advanced techniques universally across different crop types and geographic regions. In the subsequent sections, we will outline the methodology employed to build and validate our model, present the results of our field-level assessments, and discuss the implications of our findings for future research.

2. Materials and Methods

The yield data for this study were obtained from 481 agricultural fields–encompassing those with identical geometry but different crops cultivated in various years–across two regions of Ukraine (Fig. 1): Kyiv, which accounted for approximately 10% of the fields, and Cherkasy, comprising about 90%. The data covers a three-year period from 2020 to 2022 and includes crops such as wheat (146 fields), maize (223 fields), and sunflower (112 fields).

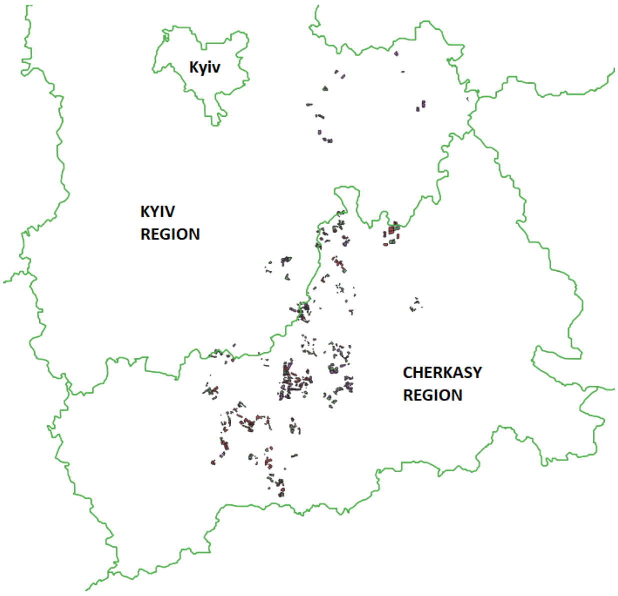

Fig. 1. Location of agricultural fields considering in this study (area of interest AOI)

Satellite data, sowing dates, and field area were primarily utilized as key predictors to develop a machine learning model for yield forecasting. While meteorological data were also examined during the modeling process, their contribution to predictive power was found to be limited at field level (please refer to the discussion section for further insights on this topic).

Given the relatively small number of agricultural fields and predictors in this study, advanced machine learning algorithms like neural networks were deemed less suitable. Instead, a simpler yet effective approach, Random Forest, was applied.

2.1. Input data

Satellite data were obtained from Sentinel-2 images for each investigated year during the April-September period, focusing solely on the NDVI (Normalized Difference Vegetation Index) and LAI (Leaf Area Index) values for each month. NDVI was calculated using red (Red) and near-infrared (NIR) bands based on following equation:

$$\mathrm{NDVI} = \frac{(\mathrm{NIR} - \mathrm{Red})}{(\mathrm{NIR} + \mathrm{Red})}. \qquad (1)$$

In this study, we estimated Leaf Area Index (LAI) using a Radiative Transfer Model (RTM) implemented in the Sentinel Application Platform (SNAP).

We used the minimum, average, and maximum indices of LAI and NDVI for each month of the growing season as satellite-based predictors.

Weather data were used in the form of gridded data (6.5 km resolution), generated from measurements collected by ground-based meteorological stations.

As previously mentioned, field area and sowing date were additional input parameters for the ML model. Overall, the average field sizes are approximately 70 hectares for maize, 80 hectares for sunflower, and 70 hectares for wheat. Table 1 presents the average field area for each crop by year, along with the mean sowing dates.

Table 1. Averaged field area (ha) and sowing date for each crop by year

Year	Crop	Field area	Sowing date	Number of fields
2020	Maize	70.6	15-apr	110
	Sunflower	74	7-apr	37
	Winter Wheat	65.1	5-oct (2019)	46
2021	Maize	73	20-apr	63
	Sunflower	92.7	5-may	34
	Winter Wheat	62.1	25-sep (2020)	50
2022	Maize	62.6	3-may	50
	Sunflower	70.4	1-may	41
	Winter Wheat	79.5	28-sep (2021)	50

The actual yield data, analyzed as the target variable, was examined based on distribution for each crop. Maize showed the widest distribution, with a mean of 6.5 t/ha and a standard deviation of 3.5 t/ha. In contrast, the distributions for sunflower and winter wheat were narrower, with sunflower having a mean of 3.3 t/ha and a standard deviation of 1.2 t/ha, and winter wheat a mean of 5.9 t/ha and a standard deviation of 1.1 t/ha (Fig. 2). To achieve accurate predictions, we needed to identify predictors that could best describe the variability of our target variable. In Table 2, you can see the list of all predictors used to train our ML model.

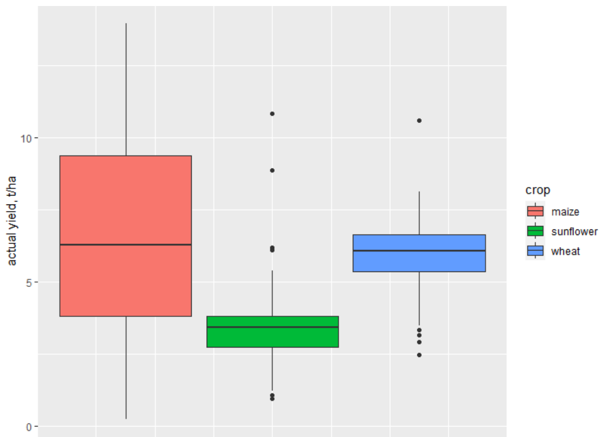

Fig. 2. Boxplots of actual yield data distribution per crops for 2020–2022

Table 2. List of predictors used to train the ML model

Type of predictors	Predictors
Satellite-based	a) Monthly-averaged LAI and NDVI for May, June, July, August, and September; b) Monthly maximum LAI and NDVI for the same months c) Monthly minimum LAI and NDVI for the same months
Meteorological	a) Sum of temperature for March, April, May, June, July, August, and September; b) Sum of precipitation for the same months
Other	a) Date of sowing; b) Field area

2.2. Machine learning algorithms

In this study, the caret package in R was utilized for yield prediction, employing a systematic approach. The process began with data preparation, including cleaning and feature engineering, followed by splitting the dataset into training and testing subsets. A control framework was established using the “trainControl()” function to implement cross-validation (10-fold) for model evaluation. The model was trained using the “train()” function with the method set to 'rf' (which stands for Random Forest), allowing for the specification of various algorithms and hyperparameter tuning through a grid search. Model performance was assessed on the test set using metrics such as RMSE and R^2. Additionally, feature importance was analyzed to understand the contributions of different variables to yield prediction. The model has been created for each crop, separately. The process culminated in model deployment for future predictive applications, demonstrating the effectiveness of our approach in building robust machine learning models for agricultural yield forecasting.

3. Results

3.1. Maize

Based on actual maize yield data, we observe significant variation across years (Fig. 3). In 2020, the yield reached a minimum value of approximately 3.8 tons per hectare on averaged, which is unusually low for this crop. One possible reason for this low yield is the occurrence of incorrect agronomic practices, coupled with insufficient rainfall during a critical stage of maize development. In July 2020, the Cherkasy region (where nearly all the tested maize fields are located) received only 34 mm of precipitation, which is significantly lower compared to 75 mm and 56 mm in 2021 and 2022, respectively.

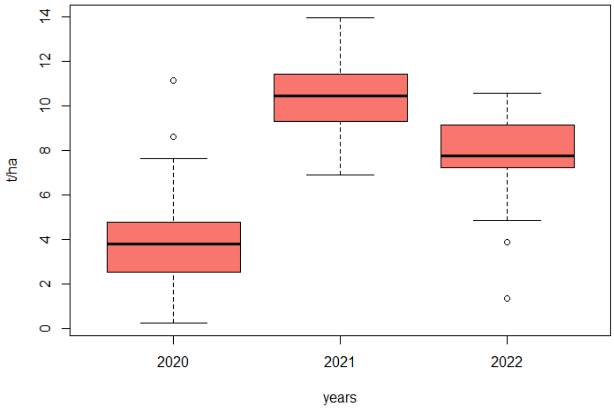

Fig. 3. Distribution of actual maize yield data by year

As mentioned in section 2.2, the Random Forest model was chosen for crop yield prediction. The trained model yielded satisfactory results on the holdout data, with an R-squared of approximately 0.8 and an RMSE of around 1.5 t/ha. The feature importance graph was used to identify the most significant predictors. As shown in Fig. 4, the top predictors for maize include vegetation indices in August, as well as the "datsow" variable, which refers to the sowing date (in day-of-the-year format). The normalized RMSE, calculated based on the average of the mean actual yields for each year (2020: 3.8 t/ha, 2021: 10.3 t/ha, 2022: 7.8 t/ha), is approximately 20%.

The feature importance graph in a Random Forest algorithm visually represents how important each feature (predictor) is in making predictions for the model.

Random Forest, an ensemble learning algorithm based on decision trees, calculates the importance of each feature by evaluating its contribution to reducing impurity (such as Gini impurity or entropy) at each decision node in the trees. Each line in the graph represents the importance of a specific feature in the model. The longer the bar, the more important the feature is for making accurate predictions (Breiman, 2001).

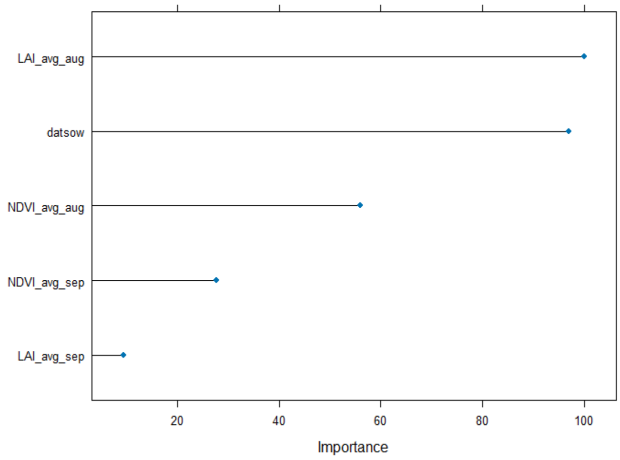

Fig. 4. Feature importance graph for maize prediction (avg – average; aug – August; sep – September; datsow – date of sowing)

3.2. Sunflower

The actual sunflower yield data shows little variation across the years in the testing period (see Fig. 5). In 2020 and 2022, the average yield was around 3 t/ha, while in 2021, it was 4 t/ha. The most likely explanation for this difference is the more favorable weather conditions in 2021 (good precipitation regime within AOI).

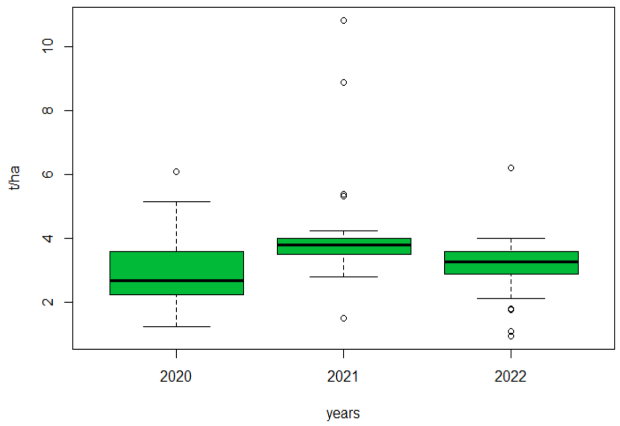

Fig. 5. Distribution of actual sunflower yield data by year

Additionally, as shown in Fig. 5, there are some extreme values for actual yield in certain fields (especially in 2021) that were excluded by the algorithm during model training. The cross-validation process shows that the model, on average, demonstrated moderate accuracy with an R-squared of 0.4 and an RMSE of approximately 0.9 t/ha. The top predictors were the average LAI for May and July (see Fig. 6). However, the selected predictors did not provide sufficient predictive power, as evidenced by a normalized RMSE of approximately 23%.

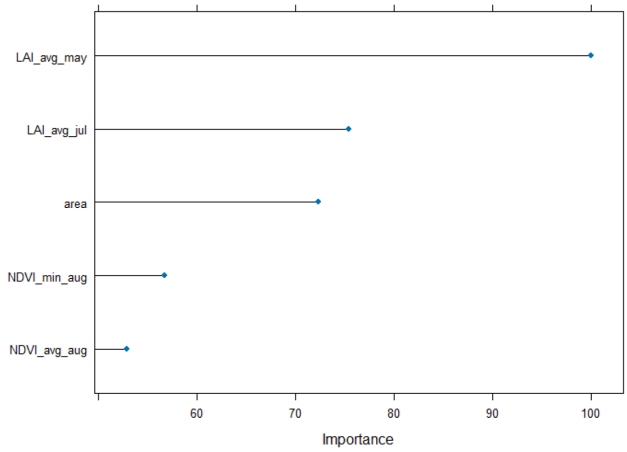

Fig. 6. Feature importance graph for sunflower prediction (avg – average; may – May; jul – July; aug – August; min – minimum)

3.3. Winter wheat

The trend in actual winter wheat yields over the years is similar to that of sunflower yields. The yield was nearly at 5.5 t/ha in both 2020 and 2022, while in 2021, it reached 6.5 t/ha (Fig. 7).

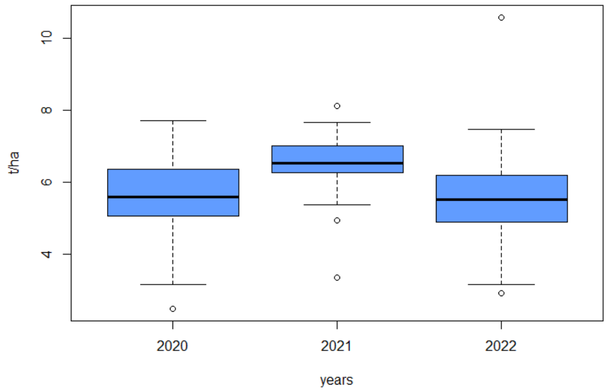

Fig. 7. Distribution of actual winter wheat yield data by year

The relatively consistent yield pattern for winter wheat makes it more difficult to identify suitable predictors that can capture all the fluctuations in its yield. The lack of extreme yield variation means that finding robust predictors to explain this variability is challenging. This is reflected in the moderate R-squared value (0.35) in yield prediction models, indicating that while some predictors can explain part of the yield variation (Fig. 8), the models cannot fully account for all of it.

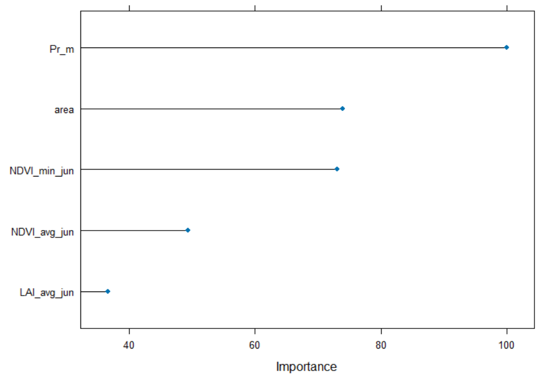

Fig. 8. Feature importance graph for winter wheat prediction (Pr_m – precipitation in March; avg – average; jun – June; aug – August; min – minimum)

In the Fig. 8 the graph reveals that the sum of March precipitation (Pr_m) is the most significant predictor (precipitation during this period is critical for the recovery of active plant growth after winter), followed by the area of the field. These two features have the highest importance scores, indicating that March rainfall and field size are the most influential factors in predicting the winter wheat yield. Next, the graph shows NDVI and LAI for June, but their importance is slightly lower than that of March precipitation and field area.

The model achieves RMSE of 0.9 t/ha, reflecting a moderate level of prediction accuracy. Interestingly, this RMSE is the same as the RMSE for sunflower**,** indicating that the Random Forest model is equally effective in predicting the yield of both crops with a comparable level of precision. When normalized by the average yield of winter wheat, the normalized RMSE is approximately 15%.

4. Discussion

In agricultural research, one of the key objectives is to understand the variability of crop yields, as it plays a crucial role in optimizing production. Maize, sunflower, and winter wheat were the three crops tested in this study, each exhibiting varying levels of yield fluctuations. These variations are influenced by various environmental and climatic factors, and analyzing them can provide valuable insights for improving crop management strategies.

When comparing the yield variability of winter wheat, sunflower, and maize, distinct patterns emerge. Winter wheat and sunflower both exhibit relatively low yield variability, while maize shows significantly higher fluctuations. The standard deviation of winter wheat yield is 1.2 tons per hectare (t/ha), and sunflower has a slightly higher standard deviation of 1.3 t/ha. In contrast, maize experiences a much higher standard deviation of 3.4 t/ha. This indicates that winter wheat and sunflower yields are generally more stable, while maize yields are more sensitive to environmental fluctuations.

The lower variability in winter wheat and sunflower yields suggests that environmental factors influencing their growth are more predictable or consistent. This stability in yield variability makes these crops less sensitive to sudden changes in temperature, rainfall, or soil conditions. On the other hand, maize, with its higher yield variability, is more vulnerable to fluctuations in these environmental factors. This greater sensitivity to changes creates more uncertainty in predicting maize yields.

The relatively consistent yield patterns for winter wheat and sunflower make it more difficult to identify suitable predictors to capture the full range of yield variability. With less pronounced fluctuations, finding robust predictors that explain this variability becomes a challenge. As a result, models designed to predict yields for these crops tend to have a moderate R-squared value (0.35-0.4), suggesting that while some predictors can explain part of the yield variation, they cannot fully account for all of it.

In contrast, the higher yield variability of maize provides more opportunities to identify influential environmental factors, which could make predicting maize yields somewhat easier. However, this also means that maize yield prediction models tend to be less stable and may require more complex factors to improve their accuracy.

In our analysis, it is evident that weather, particularly precipitation, plays a significant role in determining the actual yield levels of crops. However, a key challenge arises from the fact that we only have access to gridded weather data with a spatial resolution of approximately 6.5 km. Each grid cell encompasses several fields, which may have varying yield outcomes due to differences in local conditions. This means that the variability in yield across individual fields within a single grid cell is not adequately captured by the average weather data for that grid.

While weather data, such as March precipitation, showed strong importance in predicting winter wheat yields in this study, its effectiveness at the field level is limited. The gridded weather data cannot fully account for the local variations within each grid cell. As a result, despite the strong correlation between precipitation and yield at a broader scale, weather data alone were not sufficient to accurately predict yield variability at the field level. Other factors, such as soil properties, field management practices, and microclimates, likely contribute to the yield differences within the same grid and need to be considered for more precise predictions.

It is important to note that the forecasting lead time for yield prediction, based on the top predictors, is typically around one month before harvest. For maize, usually harvested in September in Ukraine, the key predictors are vegetation indices from August. For sunflower, typically harvested in August, the main predictors are from May and July. For winter wheat, harvested at the end of June or early July, the primary predictors are from March and June.

The normalized RMSE (Root Mean Square Error) measures the accuracy of the model's predictions relative to the average yield of each crop. For maize, the normalized RMSE is approximately 20%, reflecting a relatively high level of accuracy (80%) in yield predictions. For sunflower, the normalized RMSE is slightly higher at 23%, suggesting comparable predictive performance (77%). For winter wheat, the normalized RMSE is around 15%, indicating a lower prediction error relative to its average yield.

5. Conclusion

This study explored the variability in crop yields of maize, sunflower, and winter wheat, emphasizing the role of environmental and climatic factors through vegetation indices, weather and some agromanagement data. The results highlight significant yield fluctuations in maize, with lower yields in 2020 due to adverse weather and probably agronomic practices, while sunflower and winter wheat displayed more consistent yield patterns over the years.

The Random Forest model, employed for yield prediction, demonstrated varying degrees of accuracy across crops. For maize, the model achieved a strong R-squared of 0.8 and an RMSE of 1.5 t/ha, with vegetation indices and sowing dates as key predictors. In contrast, sunflower and winter wheat had lower prediction accuracy, with RMSE values of 0.9 t/ha for both crops and moderate R-squared values, reflecting the challenges of predicting yields for crops with more stable yields and less pronounced fluctuations.

While weather, particularly March precipitation, proved to be a significant predictor for winter wheat, the spatial resolution of the gridded weather data limited its ability to capture yield variability at the field level. This highlights the need for incorporating additional factors such as soil properties, field management practices, and microclimates to improve prediction models and better account for the variability within individual fields. The study underscores the complexity of crop yield prediction and the need for more precise data to enhance agricultural decision-making.

Author Contributions: Conceptualization, O. Kryvoshein and O. Kryvobok; methodology, O. Kryvoshein; formal analysis, O. Kryvoshein; investigation, O. Kryvoshein and O. Kryvobok; data curation, D. Zhylchenko and O. Kryvoshein; writing – original draft preparation, O. Kryvoshein; writing – review and editing, O. Kryvoshein, O. Kryvobok and D. Zhylchenko; visualization, O. Kryvoshein. All authors have read and agreed to the published version of the manuscript.
Funding: This research received no external funding.
Data Availability Statement: Not applicable.
Acknowledgments: The authors would like to express their sincere gratitude to the Earth Observing System Data Analytics company (eosda.com) for support. We are also grateful to reviewers and editors for their valuable comments, recommendations, and attention to the work.
Conflicts of Interest: The authors declare no conflict of interest.

Внесок авторів: Концептуалізація – О. Кривошеїн та О. Кривобок; методологія – О. Кривошеїн; формальний аналіз – О. Кривошеїн; дослідження – О. Кривошеїн та О. Кривобок; обробка даних – Д. Жильченко та О. Кривошеїн; написання – підготовка оригінального тексту – О. Кривошеїн; написання – рецензування та редагування – О. Кривошеїн, О. Кривобок та Д. Жильченко; візуалізація – О. Кривошеїн. Всі автори прочитали та погодилися з опублікованою версією рукопису.
Фінансування: Це дослідження не отримало зовнішнього фінансування.
Доступність даних: Не застосовується.
Подяки: Автори висловлюють щиру подяку компанії Earth Observing System Data Analytics (eosda.com) за підтримку. Також вдячні рецензентам та редакторам за цінні коментарі, рекомендації та увагу до роботи.
Конфлікти інтересів: Автори заявляють, що не мають конфлікту інтересів.

References

Kyvobok, O., Kryvoshein, O., Adamenko, T. (2018). Peculiarities of technological adaptation of the cgms system for agricultural crops monitoring in Ukraine. *Ukrainian hydrometeorological journal*, 22, 64–79.

Kryvoshein, O., Kryvobok, O., Kozhushko, O. (2023). Correlation of satellite-based LAI and actual crop yield. *Ukrainian journal of remote sensing*, 10(3), 21–27.

Liu, X., Zhang, Y. (2022). A review of remote sensing-based crop yield estimation and forecasting. *Remote Sensing*, 14(2), 368. doi:10.3390/rs14020368.

Liu, X., Zhang, S., Liu, L. (2021). Integration of satellite data and machine learning for crop yield prediction: A case study of wheat. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174, 89–101. doi:10.1016/j.isprsjprs.2020.12.007.

Khan, M. S., Hameed, M. A. (2023). Machine learning approaches for crop yield prediction: A comprehensive review. *Computers and Electronics in Agriculture*, 205, 107552. doi:10.1016/j.compag.2023.107552.

Miao, Y., Liu, Y., Wang, Z. (2024). Field-scale crop yield prediction using a combination of remote sensing and deep learning techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 62(3), 1234–1246. doi:10.1109/TGRS.2023.3123456.

Miller, J. R., Scott, T. (2023). Advancements in crop yield forecasting through integration of satellite and ground-based data. *Precision Agriculture*, 24(1), 85–101. doi:10.1007/s11119-022-09945-3.

Zhang, L., Li, J. (2021). Predicting crop yield using multi-source data and machine learning techniques: A comprehensive evaluation. *Journal of Agricultural and Food Chemistry*, 69(12), 3482–3492. doi:10.1021/acs.jafc.1c00234.

Huang, H., Liu, X. (2022). "Real-time crop yield prediction using remote sensing and ensemble learning methods. *International Journal of Applied Earth Observation and Geoinformation*, 109, 102649. doi:10.1016/j.jag.2022.102649.

Desloires, J., Ienco, D., Botrel, A. (2023). Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods. *Computers and Electronics in Agriculture*, 209, 107807.

Marszalek, M., Körner, M., Schmidhalter, U. (2022). Prediction of multi-year winter wheat yields at the field level with satellite and climatological data. *Computers and Electronics in Agriculture*, 194, 106777.

Hosseini, M. et al. (2020). Soybean yield prediction at field level using Sentinel-1 Data. *AGU Fall Meeting Abstracts*, GC034-06.

Hernandez, M., Ritchie, J. (2023). Field-level yield prediction using a combination of ground truth data and remote sensing. *Precision Agriculture*, 24(3), 498–512. doi:10.1007/s11119-022-09956-2.

Zhou, Y., Liu, H. (2024). Machine learning for real-time crop yield prediction: A case study on maize using satellite and weather data. *Computers and Electronics in Agriculture*, 214, 106957. doi:10.1016/j.compag.2023.106957.

Kumar, A., Verma, R. (2023). Fusion of satellite imagery and machine learning for improving crop yield estimation accuracy. *Agricultural Systems*, 208, 103728. doi:10.1016/j.agsy.2023.103728.

Shao, J., Chen, X. (2023). Field-scale yield forecasting with deep learning and remote sensing: Applications and challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 189, 142–154. doi:10.1016/j.isprsjprs.2023.01.009.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.

ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ НА РІВНІ ПОЛЯ

Кривошеїн О. О.[1, 2], Кривобок О. А.[1, 2], Жильченко Д. О.[1]
[1] *Earth Observing System (EOS) Data Analytics, 01025, м. Київ, вул. Десятинна, 5*
[2] *Український гідрометеоролоічний інститут, 03028, м. Київ, проспект Науки, 37*

Прогнозування врожайності на рівні поля є критично важливим для оптимізації сільськогосподарської продуктивності та забезпечення продовольчої безпеки. В цьому дослідженні проведений аналіз варіабельності врожайності кукурудзи, соняшнику та озимої пшениці на 481 сільськогосподарському полі в двох регіонах України (Київська та Черкаська області) за три роки (2020–2022). Метою було дослідити вплив навколишніх факторів на прогнози врожайності культур за допомогою супутникових і погодних даних, дат сівби та площі полів як предикторів у моделі машинного навчання. У дослідженні було використано модель Random Forest. Супутникові дані з Sentinel-2, включаючи значення NDVI та LAI, використовувалися для оцінювання стану культур під час вегетаційного періоду. Для кожного року дослідження в період з квітня по вересень основна увага приділялася значенням NDVI та LAI для кожного місяця. Погодні дані, зокрема опади, також були проаналізовані, але їх прогностична здатність виявилася обмеженою через значне просторове розрізнення даних сітки грідів (6.5 км), що не дає змоги повною мірою врахувати місцеві варіації в межах кожного гріду. Як результат, незважаючи на сильну кореляцію між опадами та врожайністю на більш високому рівні (регіональному), погодні дані не були достатніми для точного прогнозування варіабельності врожайності на рівні поля. Результати показали, що кукурудза мала найвищу варіабельність врожайності, тоді як соняшник і озима пшениця показували більш стабільні врожаї. Для кукурудзи модель продемонструвала відносно високі прогностичні результати з R-квадратом 0,8 та RMSE 1,5 т/га. Найважливішими предикторами були вегетаційні індекси в серпні та дата сівби. Нормалізоване RMSE для кукурудзи становило 20%. Для соняшника модель показала помірну точність з R-квадратом 0,4 та RMSE 0,9 т/га. Ключовими предикторами були середнє значення LAI у травні та липні. Однак, прогностична здатність моделі була обмежена, що призвело до нормалізованого RMSE 23%. Озима пшениця показала подібні результати до соняшника з R-квадратом 0,35 та RMSE 0,9 т/га. Завдяки вищим середнім значенням врожайності нормалізоване RMSE для озимої пшениці становило 15%. Загалом дослідження демонструє різні рівні точності моделі для різних культур, при цьому кукурудза показала найкращу прогностичну ефективність. Результати також підкреслюють необхідність врахування додаткових факторів для покращення прогностичних моделей на рівні поля (таких як властивості ґрунту, мікроклімат і детальний агроменеджмент).

Ключові слова: врожайність, прогноз, культури, поле, супутникові дані, машинне навчання.

Рукопис статті отримано 04.12.2024
Надходження остаточної версії: 17.12.2024
Публікація статті: 30.12.2024