8-22-2024 10:00 AM

Winter Wheat Biomass and Yield Estimation using Unmanned Aerial Vehicle-based and VENμS Satellite Imagery with Machine Learning Techniques

Marco S. Chiu Mr., *Western University*

Supervisor: Wang, Jinfei, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Geography and Environment

Abstract

Monitoring crop productivity is crucial in precision agriculture, often using biomass and yield as metrics to measure crop health and growth status. This thesis aims to predict dry above-ground biomass using Unmanned Aerial Vehicle (UAV) multispectral imagery, derived vegetation indices (VI), plant height, leaf area index (LAI), and plant nutrient content ratios. Additionally, the thesis tests the viability of VENμS satellite data as an alternative to other popular multispectral satellite data for predicting winter wheat yield. Conducted in two winter wheat fields in southwestern Ontario, Canada, the study employed Random Forest (RF) and Support Vector Regression (SVR) machine learning models with various variable combinations. The results demonstrate that the approach in biomass estimation was accurate and provided valuable insights into the applicability of biochemical parameters. Furthermore, VENμS produced promising yield prediction results, proving to be a better satellite platform compared to other publicly available satellite data for yield prediction.

Keywords

Precision agriculture, remote sensing, Unmanned Aerial Vehicle (UAV), VENμS, biomass estimation, yield prediction, plat nutrient contents, vegetation index, winter wheat, machine learning, Random Forests, Support Vector Machine

Summary for Lay Audience

The main source of food for the world's population is agriculture. As the global population grows, the strong demand for food sources and food security has emphasized the need for efficient and sustainable agricultural practices. Advances in technology have led to the development of precision agriculture, which involves applying technology and agricultural principles to strategically manage resources in all aspects of agricultural production, aiming to maximize crop performance and maintain environmental sustainability.

Remote sensing is the science of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device from a distance. In precision agriculture, this often involves using satellite imagery or imagery captured by cameras mounted on unmanned aerial vehicles (UAVs). Specialized sensors installed on these platforms can detect light emissions and reflections from the Earth's surface that are beyond the human eye's visible spectrum. For example, while humans cannot see near-infrared light, the sensors can detect and record the amount of near-infrared radiation reflected by plants, providing insights into their health and vigor. The reflectance data collected by these sensors can be transformed into indices that convey specific information about the plants, using formulas known as vegetation indices.

Biomass and yield are common metrics for evaluating the performance and productivity of crops. Accurately forecasting these metrics allows farmers to respond early and effectively during the growing stages to maximize harvest output. This is especially crucial for farmers in southern Ontario, where there is only one growing season each year. In this thesis, prediction models based on machine learning algorithms were developed to identify which variables are most important for accurately predicting biomass and yield. The results provide farmers with key information on the factors that most significantly influence these metrics, enabling them to monitor and manage these variables to effectively manage their crop production.

Co-Authorship Statement

This thesis follows an integrated-article format. The work in this thesis was conducted by the author under supervision of Dr. Jinfei Wang. Ideas, guidance, resources, and revision of this thesis was generously provided by Dr. Wang. She is the co-author of the both the articles in:

Chapter 2 (published)

Chiu, M. S., & Wang, J. (2024). Evaluation of Machine Learning Regression Techniques for Estimating Winter Wheat Biomass Using Biophysical, Biochemical, and UAV Multispectral Data. *Drones*, *8*(7), Article 7. https://doi.org/10.3390/drones8070287

Chapter 3 (published)

Chiu, M. S., & Wang, J. (2024). Local Field-Scale Winter Wheat Yield Prediction Using VENµS Satellite Imagery and Machine Learning Techniques. *Remote Sensing*, *16*(17), Article 17. https://doi.org/10.3390/rs16173132

Acknowledgments

First, I would like to express my deepest gratitude to my supervisor, Dr. Jinfei Wang (Professor, University of Western Ontario), for her unwavering support, guidance, and advice throughout my undergraduate and Master's studies. Despite the challenges and changes in the direction of my thesis due to data availability and equipment issues, Dr. Wang consistently provided innovative ideas and opportunities to explore unknown research gaps, ultimately helping us to formulate an achievable plan.

I am also immensely grateful for the opportunity to pause my studies from 2023 to 2024 to work at A&L Canada Laboratories, thanks to the chance and referral provided by Dr. Wang. This experience was invaluable in expanding my knowledge in GIS and remote sensing while I continued to write my thesis. Dr. Wang first introduced me to remote sensing during my undergraduate years, enabling me to complete my undergraduate thesis and eventually work towards my Master's thesis under her continued support and encouragement. Having Dr. Wang as a mentor over the past five years has been irreplaceable, and I will always be thankful for her guidance and support.

I would like to express my sincere gratitude to the staff at A&L Canada, the Geographic Information Technology and Applications (GITA) Lab, and the undergraduate work-study students for their invaluable assistance with data collection and processing. Your support during fieldwork, especially under the challenging conditions of the hot summer and amidst the COVID-19 pandemic, was greatly appreciated. I am truly grateful for all the help, guidance, and constructive criticisms over the years, which have contributed significantly to my growth as a researcher. Additionally, I extend my thanks to Dr. Jed Long, Dr. Jinhyung Lee, and Dr. Ayan Sadhu for their valuable feedback during my defense. Your guidance and critiques have been instrumental in refining this thesis.

Last but not least, thank you to my parents, Spencer and Louisa Chiu, my girlfriend Jasmine, my friends, and all my loved ones for always being there when I needed you. I was blessed not to have to worry about the cost of food and living alone because of my parents, allowing me to fully focus on my studies. I never expected to reach this far in academia, especially after barely passing my first year of undergraduate studies, but here I am, thanks to all the support I have had along the journey.

Table of Contents

List of Tables

List of Figures

List of Appendices

Glossary

AGB (Above Ground Biomass): AGB refers to the total mass of living plants, excluding roots, present above the soil surface. It includes all vegetation such as trees, shrubs, and grasses, and is a crucial parameter in ecological and environmental studies for assessing carbon storage and ecosystem productivity.

Bands/Channels: Bands (or channels) refer to specific ranges of wavelengths in the electromagnetic spectrum that the camera can capture. Each band corresponds to a particular portion of the spectrum and can be used to gather detailed information about the target's physical and chemical properties.

Ensemble Learning: Ensemble learning is a machine learning technique that combines multiple models to improve the overall performance and accuracy of predictions. The idea is that by aggregating the predictions from several models, the ensemble model can achieve better results than any individual model alone.

Increasing Node Purity: Increasing node purity is a concept used in decision tree algorithms to enhance the quality of the splits made at each node in the tree. Node purity refers to how homogenous or uniform the data points in a node are with respect to the target variable. Higher node purity means that the data points in the node are more similar to each other regarding the target variable.

K-fold Cross Validation: K-fold cross-validation is a technique used to evaluate the performance and robustness of a machine learning model. It helps to ensure that the model's evaluation is not overly dependent on a particular subset of the data.

LAI: Leaf Area Index (LAI) is a dimensionless value that represents the total leaf area per unit ground surface area. It is an important parameter in agricultural and ecological studies as it provides a measure of the amount of leaf material in a given area, which is directly related to the processes of photosynthesis, transpiration, and energy exchange.

Machine Learning: Machine learning is a field within artificial intelligence that focuses on the development of algorithms and models enabling computers to learn from and make decisions based on data. The key idea is that, rather than being explicitly programmed to

perform tasks, these systems improve their performance over time by identifying patterns and relationships within data.

Multispectral Camera: A multispectral camera is a specialized imaging device that captures image data at specific wavelengths across the electromagnetic spectrum. Unlike traditional cameras that capture images in visible light (red, green, and blue), multispectral cameras can capture images in both visible and non-visible wavelengths (such as near-infrared, short-wave infrared, etc.). This capability allows for detailed analysis of various physical and biological properties that are not visible to the naked eye.

Orthomosaic: A multispectral camera is a specialized imaging device that captures image data at specific wavelengths across the electromagnetic spectrum. Unlike traditional cameras that capture images in visible light (red, green, and blue), multispectral cameras can capture images in both visible and non-visible wavelengths (such as near-infrared, short-wave infrared, etc.). This capability allows for detailed analysis of various physical and biological properties that are not visible to the naked eye.

Precision Agriculture: Precision agriculture is an approach to farm management that uses information technology and data analysis to ensure that crops and soil receive exactly what they need for optimum health and productivity.

Random Forest: Random Forest is an ensemble learning method used for classification, regression, and other tasks. It operates by constructing a multitude of decision trees during training time and outputting the class (classification) or mean prediction (regression) of the individual trees.

Regression: Regression models are a type of statistical method used to predict a dependent variable (also called the response or outcome variable) based on one or more independent variables (also called predictors or features). They are fundamental tools in both supervised learning in machine learning and in traditional statistical analysis.

Support Vector Regression: Support Vector Regression (SVR) is a type of Support Vector Machine (SVM) used for regression problems. It utilizes the same principles as the SVM for classification but adapted for predicting continuous values rather than discrete class labels.

UAV: A UAV (Unmanned Aerial Vehicle), commonly known as a drone, is an aircraft without a human pilot on board. UAVs can be controlled remotely by a human operator or autonomously by onboard computers. They are used in a variety of applications across different fields due to their ability to capture data from perspectives that would otherwise be difficult, dangerous, or expensive to obtain.

Yield: Yield refers to the amount of a particular crop that is harvested per unit of land area. It is a critical measure in agriculture, reflecting the productivity and efficiency of farming practices, and is typically expressed in terms of weight (e.g., tons per hectare) or volume.

Chapter 1

1 Introduction

1.1 Background

With the increasing global population, the demand for food sources and food security has necessitated the development of efficient and sustainable agricultural practices. In the modern era, the agriculture industry faces challenges such as rising global food demand, crop disease and pest outbreaks, limited cultivated areas, and climate change. The climate in southern Ontario is projected to get considerably warmer and potentially wetter over the course of the 21st century (Hewer & Brunette, 2020). However, farmers in the province are less concerned about climate change compared to those in areas with more frequent extreme weather events (Tan & Reynolds, 2003; Reid et al., 2007). Agriculture and the agri-food sector are crucial to the Canadian economy, contributing approximately 7% to Canada's gross domestic product (GDP) and accounting for one in every nine jobs in 2022 (Agriculture and Agri-Food Canada, 2024). While climate change may not be an immediate concern for the Canadian agricultural industry, it is prudent to prepare for future challenges.

1.2 Precision Agriculture

Precision agriculture (PA) is an agricultural management approach that leverages remote sensing technologies and principles to manage spatial and temporal data associated with all aspects of agricultural production (Pierce & Nowak, 1999). This approach aims to better understand the variability within the studied field, leading to benefits such as improved crop yield and enhanced environmental quality by understanding and managing crop diseases. As sensing technology advances, the field of PA has grown accordingly. Modern tools, such as advanced spectral sensors and unmanned aerial vehicles (UAVs), have facilitated extensive research in this area (Radoglou-Grammatikis et al., 2020).

PA utilizes advanced technologies and data analysis techniques to maximize crop output while optimizing input. This is achieved by assessing quantified spatial and in-situ plant

data to inform agricultural practices, such as the application of water, labor, and fuel, thus minimizing costs and avoiding excessive waste, such as pesticide overuse and nutrient loss. Excessive use of pesticides and irrigation can cause nutrient loss, potentially harming future cultivation. PA incorporates multiple types of spatial technologies, including geographic information systems (GIS), ground-based handheld data collection, and remote sensing via ground-based or aerial vehicles, to formulate and strategize efficient agricultural practices (Chlingaryan et al., 2018).

1.2.1 Remote Sensing in Techniques in Precision Agriculture

Remote sensing refers to the science and art of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device that is not in direct contact with the target of interest (Lillesand et al., 2015). In precision agriculture (PA), remote sensing encompasses nearly every aspect, including biomass estimation and yield prediction (Sishodia et al., 2020). Spectral imagery obtained from sensors on various remote sensing platforms can be transformed into vegetation indices (VIs), which are quantitative measurements that gauge the general vigor of plants. These indices are derived from the reflectance rates of plant surfaces at different wavelength bands. Plant spectral information in the visible and near-infrared (NIR) wavelengths has been shown to have a high correlation with crop growth (Zhang et al., 2018), leading to the widespread use of VIs in crop yield-related remote sensing studies. Vegetation indices such as the normalized difference vegetation index (NDVI) and the normalized difference red edge (NDRE) have been validated as reliable predictors of winter wheat yield. Additionally, indices like the modified soil-adjusted vegetation index (MSAVI) and the modified chlorophyll absorption ratio index (MCARI) are effective in predicting the leaf area index (LAI) of winter wheat, an indicator correlated with crop yield (Fu et al., 2020; Panek et al., 2020; Tian et al., 2015; Xie et al., 2014).

1.2.2 Popular Platforms in Remote Sensing

1.2.2.1 Satellite

Remote sensing imagery has been used in agricultural applications for over 60 years. Satellite imagery collected from the Landsat series, MODIS, and Sentinel-2 has been

widely adopted in crop monitoring and yield estimation research (Liaqat et al., 2017; Hunt et al., 2019). Researchers have had varying degrees of success in developing accurate crop yield estimation models, though none have matched the accuracy of UAV-based imagery models. Previous research shows that Sentinel-2 multispectral data alone was able to produce a regression model representing 70% of the winter wheat crop yield variability under optimal conditions (Zhao et al., 2020). The primary limitations of using satellite imagery compared to UAV-based images are the lower spatial and temporal resolutions. UAV systems can produce data at sub-10 cm spatial resolution, whereas satellites typically operate at meter-level spatial resolutions (Harwin & Lucieer, 2012).

Landsat 8 Operational Land Imager (OLI), launched in 2013, has a spatial resolution ranging from 15 to 30 meters across its nine spectral bands and revisits the same geographical location every 16 days (United States Geological Survey (USGS), 2019). Sentinel-2, launched in 2015, offers varied spatial resolutions across its 13 multispectral bands (10 m, 20 m, and 60 m) with a revisit time of 5 days, thanks to its constellation of twin satellites (European Space Agency (ESA), 2015). However, satellites such as VENμS, despite their limited publicly available data, and commercial satellites like PlanetScope and WorldView-3, provide spatial resolutions of 5 meters or below with almost daily revisits, making them potential alternatives to UAV-based images.

The ease of access to satellite data offers a significant advantage over UAV-based remote sensing. Most satellite data are widely available, and some, such as Landsat and Sentinel-2 data, are free of charge, which can reduce research costs for those who do not own equipment for spectral data collection. However, optical satellite images cover large areas, resulting in lower quality compared to UAV-collected data in terms of spatial and temporal resolution. Nonetheless, the accessibility and cost-effectiveness of satellite data make it an invaluable resource for agricultural research, particularly when acquiring high-resolution UAV data is not feasible.

1.2.2.2 UAV

UAVs, or unmanned aerial vehicles, have gained popularity in recent decades, with common consumers purchasing them for recreational purposes. In remote sensing, UAVs

offer a relatively low-cost alternative in terms of time, money, and manpower (Ehsani & Maja, 2013). Modern drones or UAVs enable users to program a planned flight path in advance, allowing the UAV to operate fully automatically at the site of interest.

Depending on the type of UAV being used, some can carry heavier objects and consume more energy, while others are lightweight, designed to carry lighter objects, and provide a longer flight period. Users can select the appropriate type of UAV based on the sensor adopted in the research and the size of the study area. Compared to optical satellite images, UAVs can produce higher spatial resolution because they capture data at lower altitudes, which can be adjusted manually. However, for large-scale data collection, satellite data is often a superior choice since UAVs typically cover smaller areas. UAVs require multiple battery packs and significant time to cover extensive study areas, often taking hours to cover tens of hectares of fields. Additionally, the quality of UAV images may deviate during long flight periods due to changing sunlight conditions. In contrast, satellite images capture snapshots over large areas instantaneously, providing more consistent data for large study areas.

1.2.3 Crop Productivity Metrics

Biomass in precision agriculture refers to the quantity or weight of the organism in a given area, usually represented in units of weight per area. This parameter is frequently used to assess the health of crops, their nutrient supply, and the effectiveness of agricultural management practices, thereby enabling predictions of grain yield potential (Bendig et al., 2015; Fu et al., 2014). Yield, in contrast, is the direct measurement of agricultural production per unit of land area. It is a straightforward metric that allows the agricultural industry to manage field fertilization and, when predicted accurately, to prevent food shortages (Han et al., 2020). Accurate predictions of biomass and yield enable farmers to respond early and effectively during the growing stages to maximize harvest output. This is especially crucial for farmers in southern Ontario, where there is only one growing season each year.

1.3 Research Questions

Remote sensing techniques are widely adopted in precision agriculture, utilizing spectral data to analyze crop properties such as health and vigor with nutrient levels, and productivity with metrics like biomass and yield estimation (Atkinson Amorim et al., 2022; Han et al., 2020; Lee et al., 2020; Yu et al., 2022). This study attempts to address two gaps in the field of crop productivity studies. First, while there is extensive research on biomass estimation, few studies have explored the use of biochemical parameters, such as plant nutrient contents, as predictive variables. Biomass and yield estimation generally rely on a range of variables such as plant height, leaf area index (LAI), and specific vegetation indices (VIs), including the renormalized difference vegetation index (RDVI) and modified hyperspectral variants of the normalized difference vegetation index (NDVI-like) (Bendig et al., 2014; Tian et al., 2015; Xie et al., 2014). These variables measure the physical structures of plants and are indicative of plant biomass and yield. It is well established by Marschner (2001) that plant nutrients and biochemistry are intricately linked to plant structure, health, and condition, all of which are critical factors in implementing precision agriculture (PA) strategies. Exploring the inclusion of biochemical parameters could potentially enhance the accuracy of biomass and yield predictions by providing a more comprehensive view of plant health and productivity. This integration of biochemical data with traditional physical metrics and VIs could lead to more precise and actionable insights for farmers, helping to optimize inputs and improve crop management practices.

Secondly, researchers have traditionally faced challenges with optical satellite imagery due to its relatively lower spatial resolution compared to ground-collected data (Fu et al., 2020). This limitation has often restricted research to regional scales rather than local, field-scale studies. However, VENμS (Vegetation and Environment monitoring on a New Micro-Satellite) offers a potential solution by providing higher spatial resolution data compared to most other satellites. With frequent revisits every 2 days and a wide range of multispectral bands, VENμS data can enhance the precision and applicability of satellite-based studies at the field scale.

Thus, the research questions we attempt to answer in this thesis are:

i. Can machine learning models accurately estimate winter wheat above-ground biomass (AGB) using plant height, LAI, UAV-based MicaSense multispectral bands, VIs, and plant nutrient content levels and ratios?

ii. What is the importance of the relationship between the variables and AGB? Which machine learning model was the most accurate in estimating winter wheat AGB? Are plant nutrient content levels and ratios significant predictors of winter wheat AGB?

iii. Can machine learning models accurately predict winter wheat yield using VIs derived from VENμS satellite imagery at different growth stages?

iv. What is the importance of the relationship between the VENμS-derived VIs and yield? Which machine learning model was the most accurate in predicting winter wheat yield using VENμS satellite imagery at a local, field-scale? Is the prediction accuracy comparable to that of other publicly available satellite data?

1.4 Research Objectives

The focus of this thesis is to explore the capability of machine learning models in predicting crop productivity. This includes evaluating machine learning regression models in estimating winter wheat above-ground biomass (AGB) using biophysical, biochemical, and UAV multispectral data, and predicting winter wheat yield using VENμS satellite imagery. The objectives of this study are:

i. To build machine learning regression models to estimate winter wheat AGB using parameters such as plant height, LAI, UAV-based MicaSense multispectral bands, the derived VIs, and plant nutrient content levels and ratios.

ii. To determine the optimal combinations of dates (growth stages) and parameters for AGB estimation in a winter wheat field located in southern

Ontario. Interpret the ranked importance of variables to evaluate the quality of the variables as predictors of AGB in the best performing regression model.

iii. To build machine learning regression models to predict winter wheat yield using VIs derived from VENµS satellite imagery.

iv. To determine the optimal combinations of dates (growth stages) and important VIs for yield estimation in a winter wheat field located in southern Ontario, and whether it is a viable alternative to other popular and publicly available multispectral satellite data. Then, uncover insights in the ranked importance of the variables and produce a yield prediction map.

1.5 Thesis Structure

This thesis is written in the integrated article format, comprising an introduction, two academic journal papers, and a conclusion. Chapter 1 provides background information, including a review of precision agriculture and its associated remote sensing applications, as well as the research questions and objectives of the thesis. Then, chapters 2 and 3 serve the purpose of incorporating new data sources and exploring their capabilities in predicting crop productivity. Chapter 2 presents a published journal paper on the use of biophysical, biochemical, and UAV multispectral imagery to estimate winter wheat biomass. Chapter 3 features a journal paper on using VENµS satellite imagery to predict winter wheat yield at a local, field scale. Chapter 4 concludes the thesis by summarizing the completed objectives and offering suggestions for future research.

1.6 Study Areas

According to Agriculture and Agri-Food Canada (2024), southwestern Ontario is one of the primary agricultural regions in Canada. The major field crops in this area include winter wheat, corn, and soybeans. In this thesis, two different winter wheat fields were studied in 2020 and 2022, respectively (Figure 1-1). Both fields are located west of the city of London, Ontario. The wheat field studied in 2020 is situated east of the village of Mount Brydges, while the field studied in 2022 is located further south, closer to the community of Melbourne. This region is classified as having a warm-summer humid

continental climate (Dfb) according to the Köppen climate classification system, with growing seasons generally lasting from April to October.

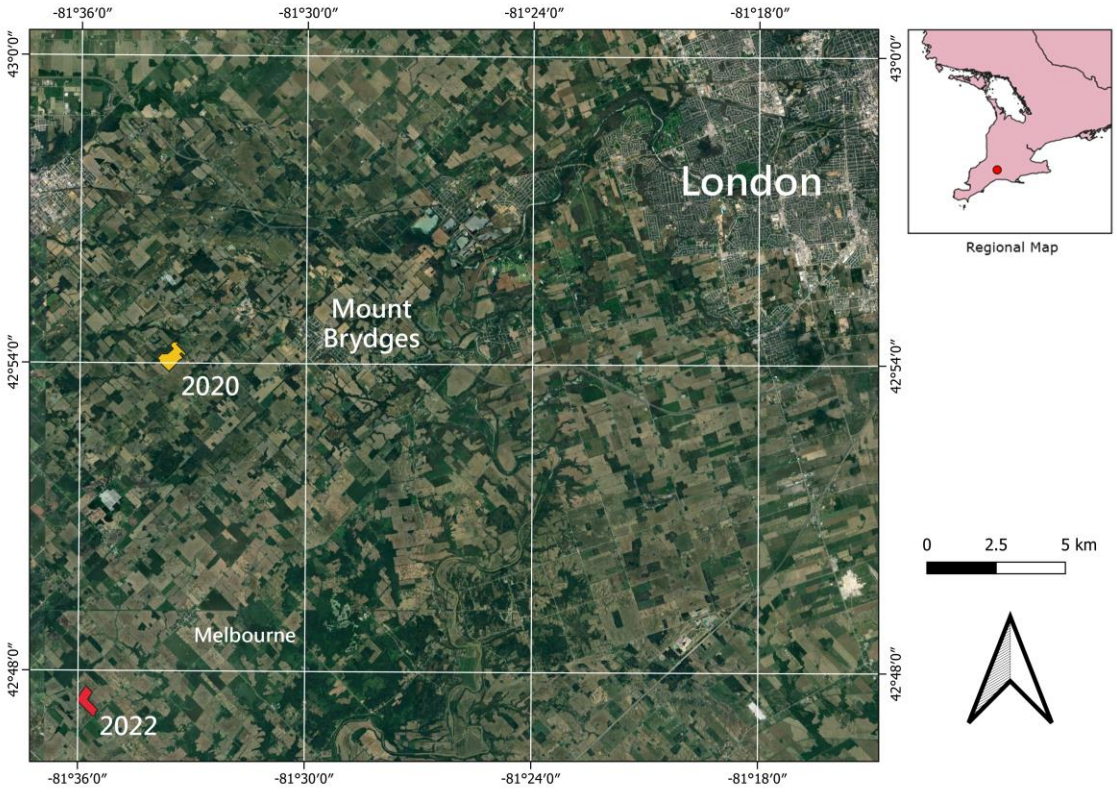


Figure 1-1. The study areas of the thesis. The names of the fields are denoted with their studied year respectively.

1.7 References

Agriculture and Agri-Food Canada. (2024, June 27). *Overview of Canada's agriculture and agri-food sector*. https://agriculture.canada.ca/en/sector/overview

Atkinson Amorim, J. G., Schreiber, L. V., de Souza, M. R. Q., Negreiros, M., Susin, A., Bredemeier, C., Trentin, C., Vian, A. L., de Oliveira Andrades-Filho, C., Doering, D., & Parraga, A. (2022). Biomass estimation of spring wheat with machine learning methods using UAV-based multispectral imaging. *International Journal of Remote Sensing*, *43*(13), 4758–4773. https://doi.org/10.1080/01431161.2022.2107882

Bendig, J., Bolten, A., Bennertz, S., Broscheit, J., Eichfuss, S., & Bareth, G. (2014). Estimating Biomass of Barley Using Crop Surface Models (CSMs) Derived from UAV-Based RGB Imaging. *Remote Sensing*, *6*(11), Article 11. https://doi.org/10.3390/rs61110395

Bendig, J., Yu, K., Aasen, H., Bolten, A., Bennertz, S., Broscheit, J., Gnyp, M. L., & Bareth, G. (2015). Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *International Journal of Applied Earth Observation and Geoinformation*, *39*, 79–87. https://doi.org/10.1016/j.jag.2015.02.012

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61–69. https://doi.org/10.1016/j.compag.2018.05.012

Ehsani, R., & Maja, J. M. (2013). The rise of small UAVs in precision agriculture. *Resource: Engineering and Technology for Sustainable World*, *20*, 18–19.

European Space Agency (ESA). (2015). *Sentinel-2 User Handbook* (pp. 1–64). ESA. https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook

Fu, Z., Jiang, J., Gao, Y., Krienke, B., Wang, M., Zhong, K., Cao, Q., Tian, Y., Zhu, Y., Cao, W., & Liu, X. (2020). Wheat Growth Monitoring and Yield Estimation based on Multi-Rotor Unmanned Aerial Vehicle. *Remote Sensing*, *12*(3), Article 3. https://doi.org/10.3390/rs12030508

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., & Zhang, J. (2020). Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China. *Remote Sensing*, *12*(2), Article 2. https://doi.org/10.3390/rs12020236

Harwin, S., & Lucieer, A. (2012). Assessing the Accuracy of Georeferenced Point Clouds Produced via Multi-View Stereopsis from Unmanned Aerial Vehicle (UAV) Imagery. *Remote Sensing*, *4*(6), Article 6. https://doi.org/10.3390/rs4061573

Hewer, M. J., & Brunette, M. (2020). Climate change impact assessment on grape and wine for Ontario, Canada's appellations of origin. *Regional Environmental Change*, *20*(3), 86. https://doi.org/10.1007/s10113-020-01673-y

Hunt, M. L., Blackburn, G. A., Carrasco, L., Redhead, J. W., & Rowland, C. S. (2019). High resolution wheat yield mapping using Sentinel-2. *Remote Sensing of Environment*, *233*(Complete). https://doi.org/10.1016/j.rse.2019.111410

Lee, H., Wang, J., & Leblon, B. (2020). Using Linear Regression, Random Forests, and Support Vector Machine with Unmanned Aerial Vehicle Multispectral Images to Predict Canopy Nitrogen Weight in Corn. *Remote Sensing*, *12*(13), Article 13. https://doi.org/10.3390/rs12132071

Liaqat, M. U., Cheema, M. J. M., Huang, W., Mahmood, T., Zaman, M., & Khan, M. M. (2017). Evaluation of MODIS and Landsat multiband vegetation indices used for wheat yield estimation in irrigated Indus Basin. *Computers and Electronics in Agriculture*, *138*, 39–47. https://doi.org/10.1016/j.compag.2017.04.006

Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). *Remote Sensing and Image Interpretation*. John Wiley & Sons.

Marschner, H. (2011). *Marschner's Mineral Nutrition of Higher Plants* (3rd ed.). Academic Press.

Panek, E., Gozdowski, D., Stępień, M., Samborski, S., Ruciński, D., & Buszke, B. (2020). Within-Field Relationships between Satellite-Derived Vegetation Indices, Grain Yield and Spike Number of Winter Wheat and Triticale. *Agronomy*, *10*(11), Article 11. https://doi.org/10.3390/agronomy10111842

Pierce, F. J., & Nowak, P. (1999). Aspects of Precision Agriculture. In D. L. Sparks (Ed.), *Advances in Agronomy* (Vol. 67, pp. 1–85). Academic Press. https://doi.org/10.1016/S0065-2113(08)60513-1

Radoglou-Grammatikis, P., Sarigiannidis, P., Lagkas, T., & Moscholios, I. (2020). A compilation of UAV applications for precision agriculture. *Computer Networks*, *172*, 107148. https://doi.org/10.1016/j.comnet.2020.107148

Reid, S., Smit, B., Caldwell, W., & Belliveau, S. (2007). Vulnerability and adaptation to climate risks in Ontario agriculture. *Mitigation and Adaptation Strategies for Global Change*, *12*(4), 609–637. https://doi.org/10.1007/s11027-006-9051-8

Sishodia, R. P., Ray, R. L., & Singh, S. K. (2020). Applications of Remote Sensing in Precision Agriculture: A Review. *Remote Sensing*, *12*(19), Article 19. https://doi.org/10.3390/rs12193136

Tan, C. S., & Reynolds, W. D. (2003). Impacts of Recent Climate Trends on Agriculture in Southwestern Ontario. *Canadian Water Resources Journal*, *28*(1), 87–97. https://doi.org/10.4296/cwrj2801087

Tian, J., Wang, S., Zhang, L., Wu, T., She, X., & Jiang, H. (2015). Evaluating different vegetation index for estimating lai of winter wheat using hyperspectral remote sensing data. *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 1–4. https://doi.org/10.1109/WHISPERS.2015.8075437

United States Geological Survey (USGS). (2019). *Landsat 8 (L8) Data Users Handbook* (pp. 1–93). USGS. https://d9-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/LSDS-1574_L8_Data_Users_Handbook-v5.0.pdf

Xie, Q., Huang, W., Liang, D., Chen, P., Wu, C., Yang, G., Zhang, J., Huang, L., & Zhang, D. (2014). Leaf Area Index Estimation Using Vegetation Indices Derived From Airborne Hyperspectral Images in Winter Wheat. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(8), 3586–3594. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. https://doi.org/10.1109/JSTARS.2014.2342291

Yu, J., Wang, J., Leblon, B., & Song, Y. (2022). Nitrogen Estimation for Wheat Using UAV-Based and Satellite Multispectral Imagery, Topographic Metrics, Leaf Area Index, Plant Height, Soil Moisture, and Machine Learning Methods. *Nitrogen*, *3*(1), Article 1. https://doi.org/10.3390/nitrogen3010001

Zhang, Y., Qin, Q., Ren, H., Sun, Y., Li, M., Zhang, T., & Ren, S. (2018). Optimal Hyperspectral Characteristics Determination for Winter Wheat Yield Prediction. *Remote Sensing*, *10*(12), Article 12. https://doi.org/10.3390/rs10122015

Zhao, Y., Potgieter, A. B., Zhang, M., Wu, B., & Hammer, G. L. (2020). Predicting Wheat Yield at the Field Scale by Combining High-Resolution Sentinel-2 Satellite Imagery and Crop Modelling. *Remote Sensing*, *12*(6), Article 6. https://doi.org/10.3390/rs12061024

Chapter 2

2 Evaluation of Machine Learning Regression Techniques for Estimating Winter Wheat Biomass Using Biophysical, Biochemical, and UAV Multispectral Data

2.1 Introduction

With the increasing growth of the global population, the strong demand for food sources and food security has highlighted the need to enhance the development of efficient and sustainable agricultural practices. In the modern era, challenges such as rising global food demand, crop diseases and pest outbreaks, limited cultivated areas, and climate change are affecting the entire agriculture industry. Tan and Reynolds found that in southern Ontario, water supply and demand are the major challenges for the agricultural industry (Tan & Reynolds, 2003). Notably, farmers in the province are less concerned about climate change compared to those in regions where extreme weather events are more prevalent (Reid et al., 2007). The agriculture and agri-food sectors contribute approximately 7% to Canada's gross domestic product (GDP), and one in every nine jobs in Canada was provided by this sector in 2022 (Agriculture and Agri-Food Canada, 2024). Although climate change is not an immediate challenge for the Canadian agricultural industry, it is prudent to be informed early and prepare for counteractions while we still have time to respond to unforeseen climate variations.

Precision agriculture (PA) utilizes advanced technologies and data analysis techniques to maximize crop output while minimizing inputs. This approach involves assessing quantified spatial and in situ plant data to guide agricultural practices, such as the application of water, labor, and fuel, thereby reducing costs and preventing excessive waste, like pesticide and nutrient loss. PA integrates various spatial technologies, including geographic information systems (GIS), handheld ground-based data collection devices, and remote sensing through ground-based or aerial vehicles, to develop and implement efficient agricultural strategies (Chlingaryan et al., 2018).

Above-ground biomass (AGB) is a frequently used parameter to indicate crop growth status and the effects of agricultural management practices, making AGB estimation one of the main applications in PA (Bendig et al., 2015; Li et al., 2015). In this study, we adopt a multivariate approach to estimate AGB using biophysical and biochemical parameters, utilizing in situ field data and high-resolution multispectral imagery collected by an unmanned aerial vehicle (UAV). Biophysical parameters included plant height and the leaf area index (LAI). In the context of using plant height as a predictor of AGB, the literature includes UAV-based height extraction methods that provide comprehensive coverage of the studied field, with multispectral cameras and LiDAR systems being common approaches (Bendig et al., 2014; Guo et al., 2024; Li et al., 2015). Research has yielded varying degrees of success in identifying plant height as an important factor correlating with AGB. Furthermore, the LAI has been proven to be a significant parameter for monitoring crop growth and estimating AGB. Liu et al. found a strong linear relationship between the LAI and AGB, though this relationship weakens after the crops' senescence (Liu et al., 2010). To explore the potential of variables that are strong predictors of AGB in the early growth stages of winter wheat, our study attempts to address the limitations of these variables in later growth stages by incorporating biochemical parameters.

While research related to biomass estimation is abundant, few studies have utilized biochemical parameters, such as plant nutrient contents, as predictors. Common variables in biomass and yield estimation included plant height, the LAI, and specific vegetation indices (VIs), like the renormalized difference vegetation index (RDVI) and the modified hyperspectral variant of the normalized difference vegetation index (NDVI-like) (Bendig et al., 2014; Tian et al., 2015; Xie et al., 2014). These variables measure the physical structures of plants and are indicative of plant biomass and yield, and it is well established that plant biochemistry is intricately linked to plant structure, health, and condition, all of which are critical factors in applying PA strategies (Cavender-Bares et al., 2020). According to Marschner (2001), macronutrients, micronutrients, and beneficial elements are essential classes of nutrients that promote plant health and growth through various mechanisms. For instance, nitrogen is a crucial macronutrient and is a major constituent of organic materials such as enzymes, chlorophyll, and compounds involved

in oxidation-reduction reactions. The nitrogen content in plant tissue can indicate yield potential and overall crop health. Micronutrients, including iron, manganese, copper, and zinc, along with beneficial elements, like sodium, boron, and aluminum, play vital roles in plant growth. These micronutrients are essential for redox reactions and other physiological processes. For example, iron is necessary for protein synthesis and increases ribosome abundance in leaf cells. Manganese and copper act as activators for various enzymes, including those involved in detoxifying superoxide radicals and synthesizing lignin. Zinc is important for maintaining membrane integrity, protein synthesis, and the production of the phytohormone indole-3-acetic acid (IAA). Although beneficial elements are essential only for certain plant types, they stimulate growth and enhance physiological functions. Sodium, for instance, facilitates the movement of substrates between the mesophyll and the bundle sheath and can partially substitute for potassium's role as an osmoticum. Boron contributes to cell wall stability by bridging polyuronides and promoting lignin synthesis. Understanding the significant roles of these nutrients underscores the potential of using plant nutrient contents as predictors of AGB. This approach could provide more comprehensive insights into crop health and productivity, thereby advancing the efficacy of PA practices.

The availability of plant nutrient data provides an opportunity to evaluate plant nutrient content ratios as predictors as well. Balanced nutrition is crucial for achieving high yields, and the overapplication of fertilizers can lead to reduced yields, soil and groundwater contamination, and harmful effects on human health and the environment (Bryant et al., 2000; Zhao et al., 2021). While few studies have explored using plant nutrient content ratios as predictors of AGB, ratios, such as nitrogen to phosphorus (N:P), have been used in crop fertilization as indicators of nutrient limitations, particularly when either nitrogen or phosphorus is the limiting factor for plant growth (Koerselman & Meuleman, 1996). The lack of research in this area, combined with the availability of relevant data, presents a promising opportunity to investigate the effectiveness of nutrient content ratios in estimating AGB.

Unmanned aerial vehicles (UAVs) are widely utilized in PA to capture timely, accurate, and cost-effective data on the earth's surface (Radoglou-Grammatikis et al., 2020).

Passive sensors, such as multispectral or hyperspectral cameras, RGB cameras, and active sensors, such as LiDAR, are typically mounted on UAVs to collect data for remote sensing applications in PA. These sensors are adopted because they do not require physical or destructive contact with plants to gather information. With the spectral data collected from remote sensing imagery, vegetation index (VI) calculations are made possible. VIs are mathematical transformations of spectral bands widely used in agricultural research to determine specific plant properties, such as the LAI, chlorophyll content, and nutrient levels (Wu et al., 2008; Xie et al., 2014; Yu et al., 2022). Consequently, VIs are commonly adopted for crop growth and health monitoring, including biomass estimation, and research has demonstrated that VIs can be effective predictors of biomass (Silleos et al., 2006; Sishodia et al., 2020). For instance, vegetation indices that performed well in the study by Fu et al. were derived using the red absorption portion (550 nm–750 nm) of the spectrum (Fu et al., 2014). On multispectral cameras, this typically includes the red band and red-edge bands. VIs that were proven by them as reliable predictors, such as the normalized difference vegetation index (NDVI) and the soil-adjusted vegetation index (SAVI), utilize spectral information from the red absorption portion. Based on these findings, it is imperative to further explore the biomass estimation capabilities of a diverse range of vegetation indices.

Although crop monitoring has traditionally relied on satellite imagery, UAV-based imagery offers significant advantages in terms of spatial and temporal resolution (Gómez et al., 2019; Liao et al., 2022). UAV systems can produce data with spatial resolutions of less than 10 cm compared to the meter-level spatial resolutions of satellite imagery. For example, the Landsat 8 Operational Land Imager, launched in 2013, has a spatial resolution varying between 15 and 30 m across its nine spectral bands and revisits the same geographical location every 16 days. Similarly, Sentinel-2, launched in 2015, features 13 multispectral bands with spatial resolutions of 10 m, 20 m, and 60 m, and a revisit time of 5 days with its constellation of twin satellites. Studies have indicated that UAV-based spectral data collected over smaller sampling areas explain more variation in wheat grain yield than the best-performing Sentinel-2 data. In comparison, Sentinel-2 data have yielded unsatisfactory results due to cloud coverage and lower temporal resolution (Bukowiecki et al., 2021). This underscores the superior spatial and temporal

resolution advantages that UAVs have over satellites for crop monitoring. Winter wheat was selected for this study due to its prominence as one of the most widely cultivated crops in southern Ontario (Ontario Ministry of Agriculture, Food and Rural Affairs, 2023). In recent years, machine learning regression methods, such as Random Forest (RF) and Support Vector Regression (SVR), have been extensively explored in biomass and yield estimation studies (Atkinson Amorim et al., 2022; van Klompenburg et al., 2020; Wang et al., 2022). A significant advantage of these machine learning regression methods over linear regression is their applicability to a wide range of data, as they do not assume linear relationships. Given the diverse categories of variables involved, it is crucial to use a method suitable for capturing complex, non-linear relationships to ensure the validity of the results and reduce variability (Tausch, 1989).

To make a well-informed estimation of AGB, it is essential to incorporate a wide range of data, including both biophysical and biochemical parameters. The objective of this study is to (i) investigate the relationships between AGB and factors, such as plant height, LAI, multispectral bands, VIs, and plant nutrient content levels and ratios; (ii) evaluate the effectiveness of RF and SVR models in estimating AGB; (iii) determine the optimal combinations of dates (growth stages) for AGB estimation in a winter wheat field located in southern Ontario; and (iv) identify the ranked importance and optimal combinations of variables for AGB estimation.

2.2 Materials and Methods

2.2.1 Study Area and Data Collection

The study site is in Southwest Middlesex County, Ontario, Canada, near the community of Melbourne, which is about 40 km southwest of the urban center of London, Ontario (Figure 2-1). Fieldwork was conducted in June of 2022, during which the average temperature was recorded at 18.8 ˚C and the relative humidity averaged %. The climate in the area is classified as warm summer humid continental climate (Dfb) according to the Köppen climate classification system. The area is predominantly agricultural croplands, and its major field crops include winter wheat, corn, and soybeans. Winter wheat was selected as the focus of this study. A winter wheat field covering 35.5 hectares

(approximately 355,000 m^2) in this region was designated as the specific area for investigation.

Figure 2-1. Location of the studied wheat field near Melbourne, ON, Canada over an ArcGIS Pro Basemap Image.

The cultivar in the studied field was soft red winter wheat, which was planted in October 2021. In the region of southwestern Ontario, winter wheat typically commences shooting in late April and is harvested from early to mid-July. Data acquisition was performed during the start of inflorescence emergence and heading stage to the ripening stage of the winter wheat. Studies have also pointed out that as AGB increases with the advancement of the growth stages, VIs' correlation with AGB decreases after the flowering stage (Di Bella et al., 2004; Wang et al., 2022). This decline is attributed to the maturation and yellowing of the plant's leaves, underscoring the significance of incorporating more variety of data in assessing the efficacy of machine learning regression models in AGB estimation.

As outlined in Table 2-1, the field was revisited every six to seven days for ground sampling to align with the changing phenological stages of the winter wheat. Ground sampling was scheduled to align with UAV flights, whenever possible, to maintain consistency in data collection relative to the growth stages. However, optimal conditions for ground sampling and UAV flights were not always synchronized due to potential adverse weather conditions, such as strong winds or rapid weather changes. Given the field's longest edge exceeding 850 m, two separate UAV flights were necessitated during each visit. Within the map depicted in Figure 2-2, 2 sets of sampling points were established: 16 sample points in a 4 × 4 grid on the northwest side and 12 sample points in a 4 × 3 grid on the southeast side. This placement aimed to maximize area coverage while minimizing labor intensity. The sampling points were positioned approximately 60 m apart, both vertically and horizontally, to ensure a representative distribution of data. A minimum distance of 50 m from all roads and houses was maintained to mitigate potential outliers and minimize disturbance to local residents. A GPS device was employed to facilitate precise revisits of these sample points during subsequent fieldwork sessions.

Table 2-1. Number of sample points and dates of data collection season.

Fieldwork Dates	Field Sample Point Groups	# of Sample Points	UAV Flight Dates	Phenology (BBCH Scale [1])
June 4	Winter Wheat Field W4 and W5	12 in W4, 16 in W5	June 8	Inflorescence emergence, heading (high 50 s to low 60 s)
June 10			June 10	Flowering, anthesis (60 s)
June 17			June 19	Development of fruit (70 s)
June 23			June 24	Ripening (low 80 s)

[1] Biologische Bundesanstalt, Bundessortenamt and CHemical industry.

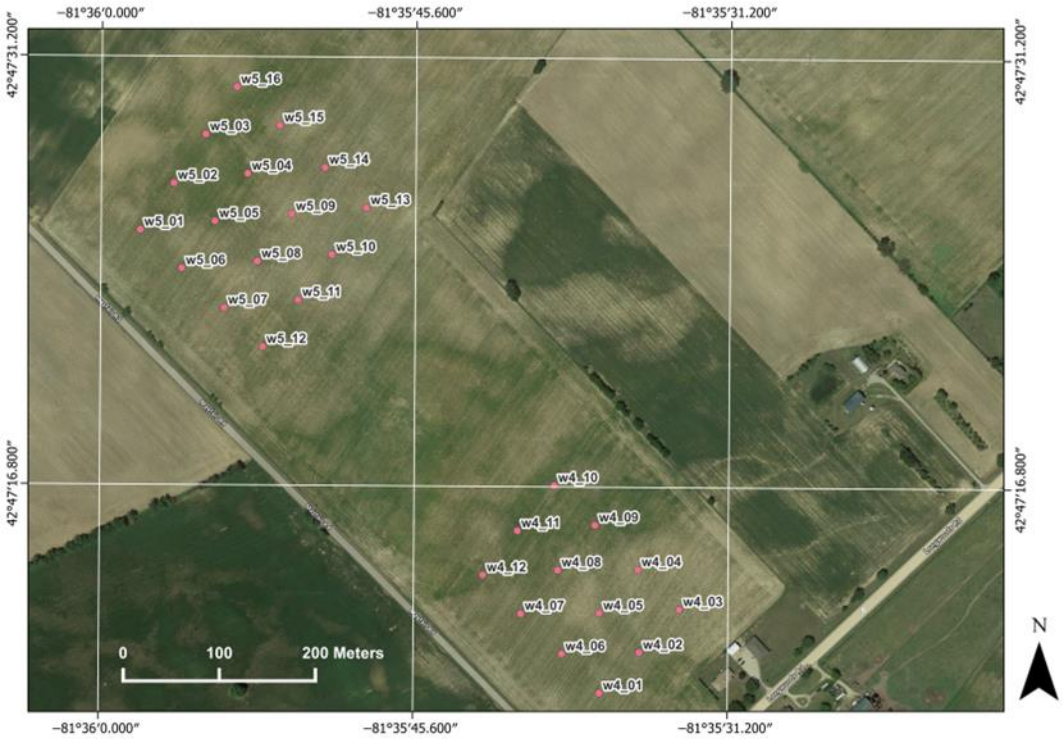


Figure 2-2. Location and distribution of the sample points.

Fresh AGB samples were destructively harvested in a 20 × 20 cm grid at each sample point and transported to A&L Canada Laboratories for immediate fresh weight determination on the day of collection. Subsequently, these samples were dried in an oven at 60 °C for 48–72 hours. After drying, the biomass was weighed, and the top leaves of the plants were analyzed to determine nutrient content levels using the A&L PT2 plant test. This test provided nutrient content levels expressed in percentages and parts per million (ppm), as well as both actual and expected nutrient content ratios for each sample. The expected ratio serves as a target value for farmers, aimed at enhancing plant quality; it remains consistent across the field but varies according to the growth stage. In contrast, the actual ratio, derived from the actual nutrient content percentages, varied from sample to sample. In this study, both nutrient content levels and the actual ratio were utilized as biochemical parameters for predicting AGB. For the purposes of this paper, the actual ratio will simply be referred to as “ratio,” as the expected ratio is not utilized in this study.

Additionally, the LAI and crop heights were measured at each sample point as physiological parameters for the machine learning models. A LI-COR LAI-2200C equipped with a 180° view cap was utilized to measure a single LAI value at each sample point as the canopy of the wheat field densified approaching crop maturity. At each sample point, a recording sequence was employed, consisting of four readings above the canopy and eight readings below the canopy near the plant roots, distributed evenly with four in one row and four in the adjacent row. Scattering corrections were applied as needed during the above canopy recording procedures, contingent on ambient lighting and sky conditions. Furthermore, six individual plant height values were measured within a 1 m radius of each sample point using a meter stick, from which an average height value was calculated for each point. During the height measurements, the plants were left undisturbed to maintain their natural posture.

2.2.2 UAV Imagery

The UAV employed in this study was the Da-Jiang Innovations (DJI) Matrice 100, equipped with a MicaSense RedEdge narrowband multispectral camera (MicaSense Inc., Seattle, WA, USA), which collected spectral information across various bands (Figure 2-3). All flights were scheduled between 10 a.m. and 2 p.m. under cloud-free or near cloud-free conditions to minimize illumination variability across the field. Flights were postponed and rescheduled to the nearest possible date if the weather conditions were suboptimal, ensuring alignment with ground data collection and the phenological stages of plant growth. Additionally, flights were conducted under the lowest possible wind conditions to reduce challenges in image mosaicking due to plant movement. The flight plan was designed using the Pix4Dcapture app, which allows the pilot to adjust flight settings dynamically. At both sections (W4 and W5) of the study field, flights were conducted at altitudes of 50–60 m above ground level, often at the upper limits of the UAV's manufacturer-recommended wind speeds. To preserve the data quality and flight efficiency, the UAV was set to fly at speeds between 3 to 4 m/s, depending on the windspeeds of the day. As the winter wheat matured and increased in height, the plants exhibited greater sway. Consequently, to ensure the accuracy of the resulting orthomosaics, all flights were performed with 85% front and side overlapping image

capture. The flight paths were executed in a zigzag pattern, aligning with the orientation of the crop rows to enhance the accuracy of the orthomosaics. The outputs were weekly generated MicaSense band orthomosaics with a spatial resolution of 4 × 4 cm. Unfortunately, the flight data collected in the first week experienced a four-day delay relative to the sampling date due to adverse flight conditions. On June 4th, thin clouds scattered across the sky led to the initial assessment of the flight data as inaccurate, followed by three days of intermittent showers or wind speeds too high for safe UAV operation.

Figure 2-3. Image of the MicaSense RedEdge narrowband multispectral camera.

2.2.3 UAV Image Processing

Pix4Dmapper (version 4.8.0) was employed to process the multispectral images collected, generating one orthomosaic image per band. Prior to each flight, the MicaSense camera was subjected to a radiometric calibration to ensure the accuracy of the reflectance data. This calibration involved positioning the camera above a MicaSense Calibrated Reflectance Panel to capture white reference images for each band, taking into account sensor influences and the scene's illumination conditions at the location. These white reference images, along with the reflectance values provided by the manufacturer's white board, were utilized in Pix4Dmapper for image calibration, enabling each of the five MicaSense bands to produce corrected reflectance data of the field. The process of

Structure from Motion (SfM), utilized by Pix4Dmapper, stitches together all the individual images captured by the camera. The high-overlapping image capture settings established for the flights facilitated this process, enhancing the accuracy of the results. The output comprised five orthomosaic images, each representing different reflectance values across the bands for both sections of the study area.

2.2.4 Vegetation Indices

The orthomosaic images were used to calculate vegetation indices (VIs) in QGIS. In order to minimize GPS error in the weekly visit to the field sample points, the VI values were averaged within a 1 m radius of the sample point. Details of the camera bands are listed in Table 2-2.

Table 2-2. Spectral bands of the MicaSense multispectral camera.

Band	Name	Band Range (nm)	Center Wavelength (nm)	Bandwidth (nm)
1	Blue	465–485	475	20
2	Green	550–570	560	20
3	Red	663–673	668	10
4	Red Edge	712–722	717	10
5	NIR	820–860	840	40

A total of 13 Vis were calculated, as listed in Table 2-3. Indices such as the NDVI and SAVI have been previously validated as reliable predictors of winter wheat biomass (Fu et al., 2014). Additionally, several of the VIs make use of spectral information in the red edge and near-infrared wavelengths, which have been demonstrated to correlate strongly with crop growth, health, yield, and the LAI (Xie et al., 2014; Zhang et al., 2018). The chlorophyll index red edge (CL_RE) has been established as an effective VI for predicting crop nitrogen content, which serves as an indicator of plant vigor and productivity.

Table 2-3. Vegetation indices to be tested in this study.

VI [1]	Formula [2]	Authors
ARVI	$\frac{NIR - [Red - 1\times(Red - Blue)]}{NIR+[Red - 1\times(Red - Blue)]}$	Kaufman and Tanre, 1992
Cl_RE	$NIR \div RE - 1$	Gitelson et al., 2003

EVI	$\frac{2.5\times(NIR-Red)}{NIR+6\times Red-7.5\times Blue+1}$	Huete et al., 2002
GCVI	$NIR\div Green-1$	Gitelson et al., 2003
ISR	$Red\div NIR$	Fernades et al., 2003
MCARI	$[(RE-Red)-0.2\times(RE-Green)]\times RE\div Red$	Daughtry et al., 2000
MSAVI	$[2\times NIR+1-\sqrt{(2\times NIR+1)^2-8\times(NIR-Red)}]\div 2$	Qi et al., 1994
NDRE	$(NIR-RE)\div(NIR+RE)$	Gitelson and Merzyak, 1994
NDVI	$(NIR-Red)\div(NIR+Red)$	Rouse et al., 1974
OSAVI	$[1.16\times(NIR-Red)]\div(NIR+Red+0.16)$	Rondeaux et al., 1996
RDVI	$(NIR-Red)\div(\sqrt{NIR+Red})$	Roujean and Breon, 1995
RVI	$NIR\div Red$	Jordan, 1969
SAVI	$[1.5\times(NIR-Red)]\div(NIR+Red+0.5)$	Huete, 1988

[1] ARVI, atmospherically resistant vegetation index; Cl_RE, chlorophyll index red edge; EVI, enhanced vegetation index; GCVI, green chlorophyll vegetation index; ISR, infrared simple ratio; MCARI, modified chlorophyll absorption in reflectance index; MSAVI, modified soil-adjusted vegetation index; NDRE, normalized difference red edge; NDVI, normalized difference vegetation index; OSAVI, optimized soil-adjusted vegetation index; RDVI, renormalized difference vegetation index; RVI, ratio vegetation index; SAVI, soil-adjusted vegetation index.

[2] Blue, blue reflectance; green, green reflectance; red, red reflectance; RE, red edge reflectance; NIR, near-infrared reflectance.

2.2.5 Biochemical Parameters

In this study, 14 nutrient content levels and 8 derived nutrient content ratios were analyzed. Existing research has identified nitrogen and phosphorus as essential for protein synthesis, enzyme activities, and chlorophyll formation in plants (Novoa & Loomis, 1981; Shi et al., 2020). Additionally, potassium is crucial in mitigating stress from drought, cold temperatures, salinity, and biotic factors, such as diseases and pests (Oosterhuis et al., 2014). For example, sufficient potassium levels can enhance photosynthetic efficiency, improve water usage, and stabilize plant metabolism under drought conditions. Additionally, nutrient content ratios, such as nitrogen to sulfur (N:S) in plant leaves, are significant indicators of crop health and nutrient deficiency (Blake-Kalff et al., 2000; Pagani & Echeverría, 2011). This framework provided the basis for testing both individual nutrient content levels and ratios. The 14 nutrients tested included nitrogen (N), phosphorus (P), potassium (K), magnesium (Mg), calcium (Ca), sodium (Na), sulfur (S), boron (B), zinc (Zn), manganese (Mn), iron (Fe), copper (Cu), aluminum

(Al), and nitrate-N. The 8 nutrient content ratios evaluated were N:S, N:K, P:S, P:Zn, K:Mg, K:Mn, Ca:B, and Fe:Mn.

2.2.6 Machine Learning Regression Modeling

In the context of machine learning, regression models are used to predict continuous outcomes based on input variables. Two prominent techniques within this domain are Random Forest (RF) Regression and Support Vector Regression (SVR), both of which offer robust solutions to complex regression problems.

RF is an ensemble learning method that operates by constructing multiple decision trees during the calibration phase and outputting the mean prediction of the individual trees. This method capitalizes on the power of multiple decision trees to reduce overfitting, which is common in models relying on a single decision tree. Each tree in the forest is built from a random sample of the calibration data, and at each node, a subset of features is randomly chosen to decide the split. This randomness helps in making the model more resilient to noise in the dataset. Moreover, Random Forest can handle large datasets with higher dimensionality and can estimate which variables are important in the underlying relationships being modeled.

SVR, on the other hand, extends the concepts of Support Vector Machines (SVMs) from classification to regression. Unlike traditional methods that minimize the error between predicted and actual values, SVR attempts to fit the error within a certain threshold. It involves the creation of a hyperplane in a multidimensional space where the distance between the data points and the hyperplane is minimized, ensuring that errors do not exceed a defined threshold. This makes SVR particularly useful in cases where a margin of tolerance is specified in the predictions. SVR is highly effective in handling non-linear relationships through the use of kernel functions, which map input data into higher-dimensional spaces (Chang & Lin, 2011).

Both RF and SVR provide distinct advantages depending on the nature of the data and the specific requirements of the regression task (van Klompenburg et al., 2020). RF is generally preferred for problems with high-dimensional spaces and large datasets,

offering interpretations in terms of feature importance. SVR is advantageous when dealing with datasets where the prediction needs to stay within a certain range and is effective in capturing complex relationships through its kernel trick. When employed thoughtfully, both methods can yield highly accurate predictive models in a wide range of scientific and industrial applications.

Figure 2-4 displays the workflow of the methodology. The modeling was written in R programming language using R Studio by utilizing packages such as "randomForest" and "e1071" for RF and SVR, respectively. In both models, the independent variables were the VIs, MicaSense bands, plant physiological parameters, plant nutrient levels, and plant nutrient content ratios. Data collected over the four weeks were randomly divided into a 70% calibration set and a 30% validation set. In the RF models, using the default settings of 500 decision trees and the square-rooted number of variables considered at each split (mtry) provided the most stable results. For the SVM models, the Radial Basis Function (RBF) kernel was used with default parameters, which worked best in this study. The strength of the prediction model was assessed using the coefficient of determination (R^2) and root mean square error (RMSE). To ensure the model's strength and stability, we validated the results by creating random splits of the calibration and validation sets 100 times. The R^2 and RMSE values reported are the average values obtained from these splits. The equations for both metrics are as follows:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (1)$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and $\bar{y}_i$ is the mean of the observed values, and

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

where $\hat{y}_i$ represents the predicted AGB (g/m^2), y_i denotes the observed AGB (g/m^2), n is the total number of observations, and i serves as the summation index, incrementing by one.

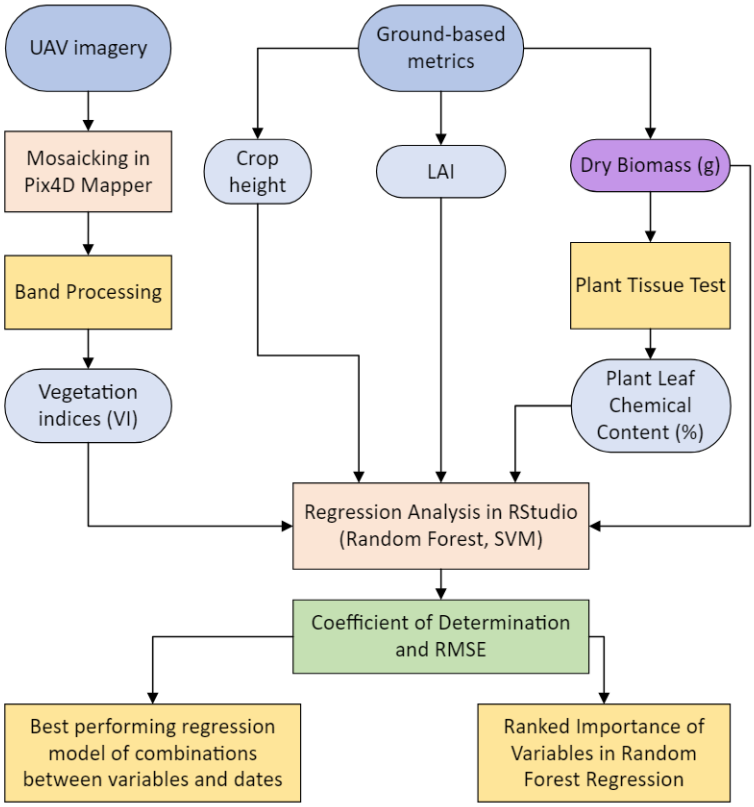


Figure 2-4. Methodology flowchart of this study.

2.3 Results

2.3.1 Biomass Data

AGB was destructively collected at each sample point. The dry weight of the sampled biomass progressively increased across different growth stages, as illustrated in Figure 2-5. Initially, AGB exhibited a modest increase during the first two weeks of fieldwork, followed by a significant acceleration in growth thereafter.

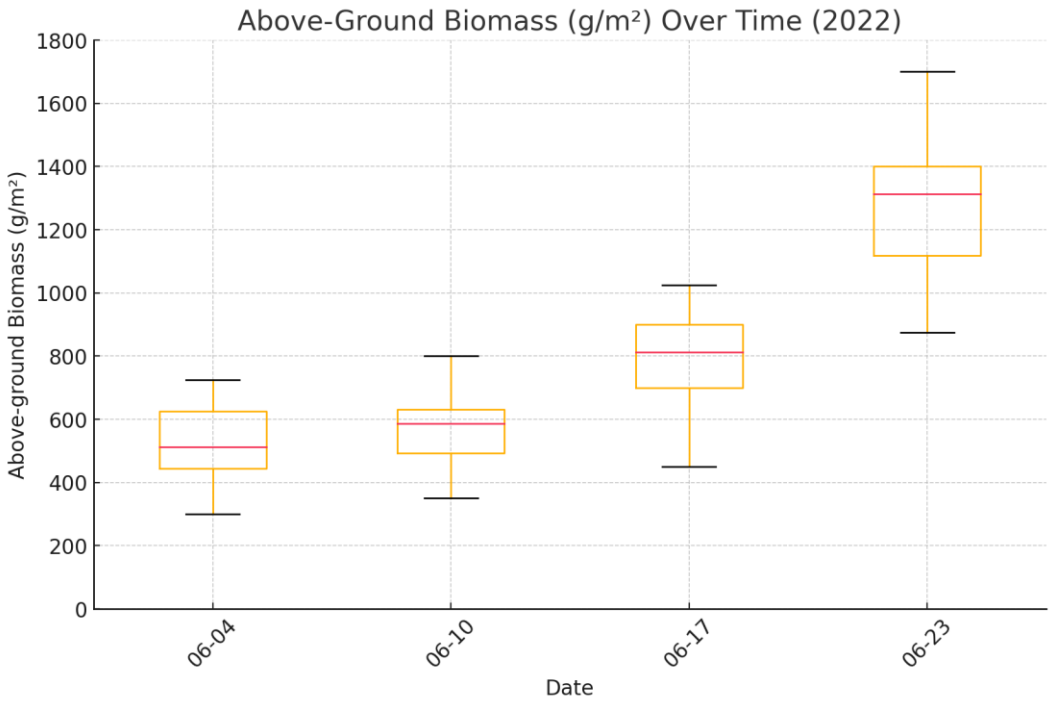


Figure 2-5. Distribution of above-ground biomass data throughout the four-week study period during the June 2022 growing season.

2.3.2 Regression Models with All Variables

A total of 42 variables were utilized as predictors for AGB, including plant height, LAI, MicaSense bands, VIs, and levels and ratios of plant nutrient content. The datasets were categorized into single-date and multi-date groups to assess the temporal impact on the models and to identify the most effective date or combination of dates for estimating AGB. These variables were incorporated into the calibration and validation of the RF and SVR models, as detailed in Table 2-4. Overall, the RF models exhibited slightly superior performance compared to the SVR models, with multi-date RF models outperforming those based on single dates. The best-performing RF model, which utilized all variables across all four dates, achieved an R^2 of 0.93 and an RMSE of 90.98 g/m^2 in its calibration set, and an R^2 of 0.80 with an RMSE of 152.71 g/m^2 in its validation set. RF models that incorporated data from three dates also demonstrated high performance. Similarly, SVR models showed improved performance with multi-date data compared to single-date models. The optimal SVR model, employing all variables and data from June 10, 17, and

23, yielded an R^2 of 0.90 and an RMSE of 108.79 g/m^2 in its calibration set, and an R^2 of 0.77 with an RMSE of 156.61 g/m^2 in its validation set. Although this model was only marginally superior to its counterpart, which utilized data from all four dates, it featured a lower RMSE. It is noteworthy that almost all models based on single-date data were not significant, a result anticipated due to the high number of variables relative to the modest dataset size of 28 entries.

Table 2-4. Calibration and validation statistics: analysis by date and modeling approach (RF and SVR) using 42 variables, including plant height, the LAI, MicaSense bands, vegetation indices, and plant nutrient content levels and ratios [1]. *n* is the number of data entries.

Date	Model	(*n*)	Calibration		Validation		
			R^2	RMSE (g/m^2)	R^2	*p*-Value	RMSE (g/m^2)
June 4	RF	28	0.95	41.90	0.21	NS	137.37
	SVR	28	0.93	40.11	0.47	<0.05	132.17
June 10	RF	28	0.96	54.64	−0.13	NS	86.19
	SVR	28	0.85	58.74	−0.14	NS	99.63
June 17	RF	28	0.93	52.62	−0.02	NS	162.74
	SVR	28	0.76	76.20	−0.14	NS	132.08
June 23	RF	28	0.95	80.82	−0.14	NS	245.32
	SVR	28	0.70	113.80	−0.14	NS	238.80
June 4, 10	RF	56	0.95	44.03	−0.06	NS	134.58
	SVR	56	0.75	63.90	0.09	NS	130.63
June 4, 17	RF	56	0.94	60.62	0.57	<0.001	138.11
	SVR	56	0.85	77.20	0.47	0.001	155.47
June 4, 23	RF	56	0.97	75.91	0.68	<0.001	237.24
	SVR	56	0.96	83.87	0.59	<0.001	257.60
June 10, 17	RF	56	0.92	59.01	0.40	<0.01	123.23
	SVR	56	0.84	71.79	0.46	0.001	120.43
June 10, 23	RF	56	0.97	71.46	0.66	<0.001	212.96
	SVR	56	0.96	83.51	0.68	<0.001	201.40
June 17, 23	RF	56	0.94	86.27	0.23	<0.05	252.59
	SVR	56	0.90	103.30	0.22	<0.05	249.10
June 4, 10, 17	RF	84	0.94	58.61	0.41	<0.001	131.64
	SVR	84	0.84	73.80	0.38	<0.001	130.17
June 4, 10, 23	RF	84	0.96	82.96	0.67	<0.001	207.08
	SVR	84	0.94	98.08	0.71	<0.001	184.32
June 4, 17, 23	RF	84	0.94	91.83	0.76	<0.001	177.09
	SVR	84	0.90	112.88	0.72	<0.001	187.79
June 10, 17, 23	RF	84	0.94	89.50	0.72	<0.001	177.91

June 4, 10, 17, 23	SVR	84	0.90	108.79	0.77	<0.001	156.61
	RF	112	0.93	90.98	0.80	<0.001	152.71
	SVR	112	0.89	113.81	0.77	<0.001	165.71

[1] All calibration models are significant at *p*-value < 0.001.

It is crucial to note that the overall best-performing model is not necessarily the one with the highest R^2 value in either the calibration or validation sets. For example, the RF model using data from June 10 and 23 demonstrated a high R^2 of 0.97 and a low RMSE of 75.91 g/m^2 in the calibration set. However, the same model exhibited significantly weaker performance in the validation set, with an R^2 of 0.66 and a high RMSE of 212.96 g/m^2. This discrepancy suggests potential overfitting, indicating that while the model predicts the calibration data exceptionally well, it does not generalize effectively to new, unseen data.

2.3.3 Variable Importance Plot

RF modeling, which involves the use of numerous decision trees, was employed to generate a variable importance plot in R Studio using the "varImpPlot()" function. The plot displays increasing node purity (IncNodePurity) on the x-axis, which indicates the importance of each explanatory variable in predicting dry AGB on the decision trees. A higher IncNodePurity value signifies that the variable is more critical as a predictor. This method was employed to visualize the variable rankings in both the RF and SVR models, aiding in the identification of key predictors in the models.

The RF model incorporating all 42 variables demonstrated optimal performance when applied to the full four-date dataset. Analysis of the variable importance plot revealed that the NDVI was the most critical predictor, as depicted in Figure 2-6. Among the top ten most influential variables, the composition included five of the thirteen VIs utilized, two out of fourteen nutrient content levels, two of the eight nutrient content ratios, and plant height. Notably, the NDVI, ISR, ARVI, and RVI exhibited significantly higher IncNodePurity values compared to the remaining variables. These VIs are commonly associated with vegetation monitoring in agriculture and biomass estimation. Macronutrients such as N and K were also ranked in the top 10. N and K are crucial macronutrients that regulate enzymes and the synthesis of organic compounds.

Additionally, K plays a vital role in cell growth and the regulation of photosynthesis, both of which are responsible for plant development (Marschner, 2011; Oosterhuis et al., 2014).

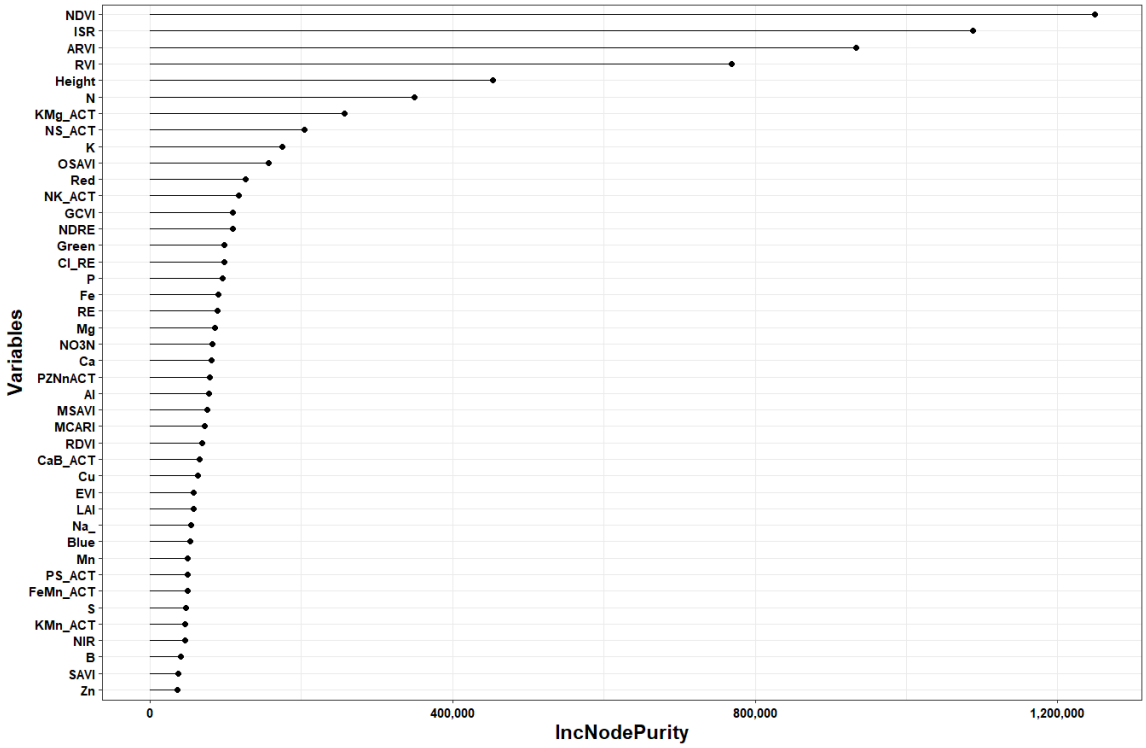


Figure 2-6. Variable importance plot produced with all 42 variables from all four dates. A higher IncNodePurity value indicates a higher impact on AGB estimation. Refer to Table 2-3 for the full names of vegetation indices. Al, aluminum; B, boron; Ca, calcium; CaB_ACT, calcium boron actual ratio; Cu, copper; Fe, iron; FeMn_ACT, iron manganese actual ratio; K, potassium; KMg_ACT, potassium magnesium actual ratio; KMn_ACT, potassium manganese actual ratio; Mg, magnesium; Mn, manganese; N, nitrogen; Na_, sodium; NK_ACT; nitrogen potassium actual ratio; NO3N, nitrate nitrogen; NS_ACT, nitrogen sulfur actual ratio; P, phosphorus; PS_ACT, phosphorus sulfur actual ratio; PZn_ACT, phosphorus zinc actual ratio; S, sulfur; Zn, zinc.

The SVR model that incorporated all 42 variables yielded the best results using data collected on June 10, 17, and 23. A variable importance plot generated from these three

dates identified the ISR as the most crucial predictor of AGB, with a slightly higher ranking than the NDVI, as shown in Figure 2-7. The top ten most important variables included five of the thirteen VIs utilized, three of the fourteen nutrient content levels, and two of the five MicaSense bands. Indices such as the NDVI, ISR, ARVI, and RVI displayed significantly higher IncNodePurity values than the other variables. Consistent with the results from the four-date analysis, the primary predictors remained the ISR, NDVI, RVI, and ARVI, albeit in a different order. In this three-date plot, MicaSense red and green bands, along with P, saw an increase in their rankings, moving into the top ten, a shift from their positions in the four-date plot. P is an essential macronutrient similar to N and K. However, its relevance had only increased in the four-date plot, while the importance of N and K had decreased. P is responsible for various biochemical reactions within the plant, including nitrogen fixation and the synthesis of nucleic acids and phospholipids, making it essential for the genetic and structural components of plant cells (Marschner, 2011). Although KMg_ACT was the highest-ranked nutrient content ratio in both plots, none of the nutrient content ratios were among the top ten variables in the three-date analysis.

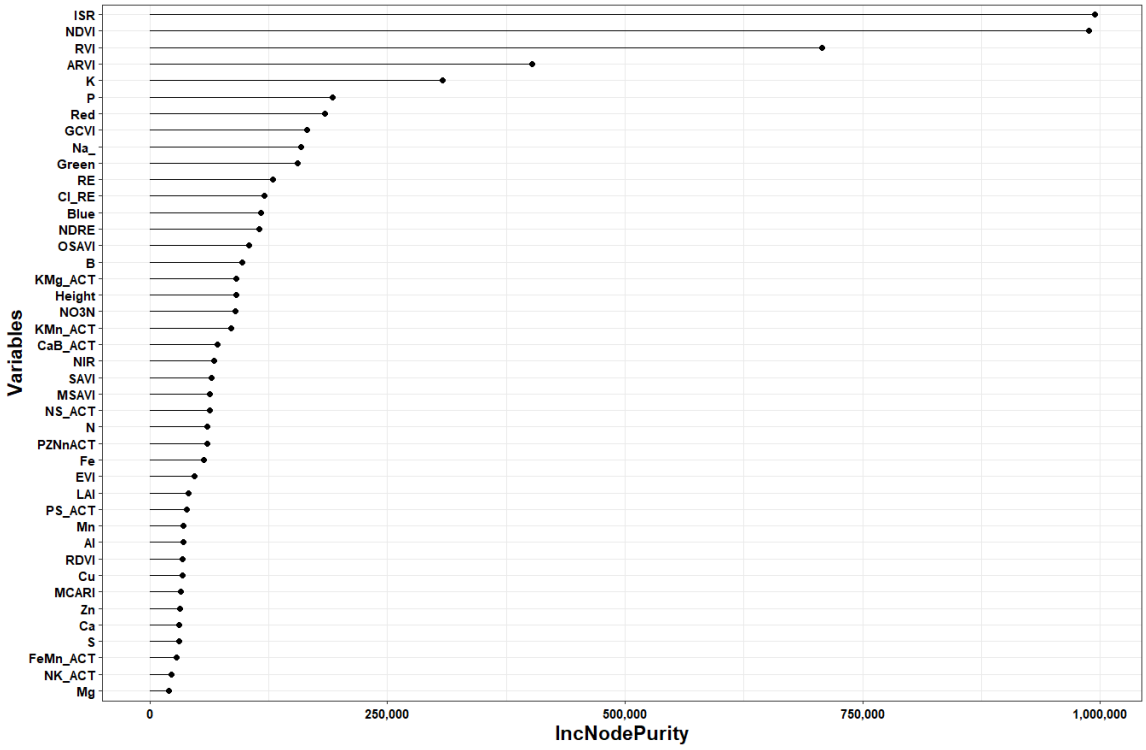


Figure 2-7. Variable importance plot produced with all 42 variables from June 10, 17, and 23. A higher IncNodePurity value indicates a higher impact on AGB estimation. Refer to Figure 2-6 for the full names of the variables.

2.3.4 Regression Models with Selected Variables

Variable selection is essential for reducing redundancy and complexity in regression models with a larger variety of variables. It enhances model performance and interpretability by focusing on the most relevant predictors, thereby reducing overfitting and computational complexity. This process also improves the model's ability to generalize well to new data, ensuring more robust and accurate predictions. As indicated by the variable importance plots (Figures 2-6 and 2-7), although the best-performing RF and SVR models utilized all 42 variables, the importance of the explanatory variables in estimating AGB varied considerably. For the RF models, based on the variable importance plot in Figure 2-6, we selected the top four, seven, ten, fourteen, twenty, and twenty-nine variables, determined by reductions in the increase in node purity, which helped establish a ranking threshold. A similar selection process was applied to the SVR

models, grouping variables into rankings of the top five, seven, ten, fourteen, twenty, and twenty-eight based on the variable importance plot in Figure 2-7.

Additionally, two distinct groups of variables were evaluated in both RF and SVR models. Research has substantiated that UAV multispectral data alone can effectively predict AGB, leading to the testing of a class consisting solely of MicaSense bands and VIs (Zhu et al., 2023). Moreover, plant nutrient content, though seldom used as a predictor for AGB, displayed high importance in some explanatory variables. Consequently, nutrient content levels and ratios were also explored as a separate class for testing.

Table 2-5 outlines the statistics for RF model calibration and validation sets using data from various dates and combinations of variables. The most effective date combination for RF utilized data from all four dates. For the calibration sets, the R^2 remained high across most groups, consistently above 0.9, except for the group containing only the top four variables. The RMSE values ranged from 89.19 to 119.81 g/m^2. Notably, the RMSE values were significantly elevated for groups comprising solely multispectral data and VIs, as well as those limited to plant nutrient content levels and ratios, with both exceeding 100 g/m^2. A decreasing trend in the RMSE was observed as more variables were included in the calibration sets, continuing up to the top 20 variables. Although the calibration sets exhibited similar performance, the validation set using only multispectral data and VIs outperformed the set that included only plant nutrient content levels and ratios. Nevertheless, neither was the best performing among the RF models. The validation set with the top seven variables demonstrated higher performance, which continued to improve up to the model incorporating the top twenty variables. The validation sets exhibited R^2 values between 0.59 and 0.81, with RMSE values spanning from 149.95 to 213.49 g/m^2.

Table 2-5. Statistics of the RF models for above-ground biomass estimation with all dates (June 4, 10, 17, 23) and different combinations of variables (n = 112) [1].

Variables	Number of Variables	Calibration		Validation	
		R^2	RMSE (g/m^2)	R^2	RMSE (g/m^2)
All VIs + 5 MicaSense bands	18	0.91	102.19	0.73	175.63

All plant nutrient content + ratios	22	0.93	100.79	0.68	196.54
Top 4: NDVI, ISR, ARVI, RVI	4	0.87	119.81	0.59	213.49
Top 7: top 4 + height, N, KMg_ACT	7	0.91	100.90	0.76	167.32
Top 10: top 7 + NS_ACT, K, OSAVI	10	0.93	92.36	0.79	156.00
Top 14: Top 10 + red, NK_ACT, GCVI, NDRE	14	0.93	89.53	0.78	160.04
Top 20: Top 14 + green, Cl_RE, P, Fe, RE, Mg	20	0.93	89.19	0.81	149.95
Top 29: top 20 + NO3N, Ca, PZn_ACT, Al, MSAVI, MCARI, RDVI, CaB_ACT, Cu	29	0.94	89.41	0.81	151.52

[1] All models are significant at p-value < 0.001.

The overall best-performing RF model employed a combination of the top 20 variables, achieving an R^2 of 0.93 and an RMSE of 89.19 g/m^2 in the calibration set and an R^2 of 0.81 and an RMSE of 149.95 g/m^2 in the validation set. Analyses of the models with the top 29 variables and all 42 variables (as detailed in Table 2-4) indicated a decline in model performance with the addition of more variables. The R^2 values plateaued while the RMSE increased, suggesting that eliminating lower-ranked variables from the variable importance plot can enhance the performance of the RF models.

Table 2-6 presents the statistics for SVR models using data from June 10, 17, and 23 across various variable combinations. The calibration sets of the SVR models displayed a range of performance with R^2 values between 0.72 and 0.88 and RMSE values from 119.65 to 178.05 g/m^2. The highest performance was observed with the top twenty-eight variables and the lowest with the top five variables. There was a linear increase in R^2 and a corresponding decrease in the RMSE as the number of variables in the calibration sets increased. Unlike the calibration sets, the validation sets did not exhibit consistent trends, with R^2 values ranging from 0.62 to 0.77 and RMSE values between 154.36 and 206.40 g/m^2. In contrast to the RF models, the SVR model utilizing only multispectral data and VIs performed worse than the model focusing solely on plant nutrient content levels and ratios. However, neither of these models achieved the highest performance.

Table 2-6. Statistics of the SVR models for above-ground biomass estimation with the three dates (June 10, 17, 23) and different combinations of variables (n = 112) [1].

Variables	Number of	Calibration		Validation	
		R^2	RMSE	R^2	RMSE

	Variables		(g/m^2)		(g/m^2)
All VIs + 5 MicaSense bands	18	0.81	145.51	0.62	190.51
All nutrient content + ratios	22	0.85	136.75	0.69	206.40
Top 5: ISR, NDVI, RVI, ARVI, K	5	0.72	178.05	0.66	187.07
Top 7: Top 5 + P, Red	7	0.76	162.81	0.62	198.99
Top 10: top 7 + GCVI, Na, green	10	0.81	147.21	0.66	184.68
Top 14: top 10 + RE, Cl_RE, blue, NDRE	14	0.81	145.18	0.69	179.08
Top 20: top 14 + OSAVI, B, KMg_ACT, height, NO3N, KMn_ACT	20	0.86	128.30	0.73	165.47
Top 28: Top 20 + CaB_ACT, NIR, SAVI, MSAVI, NS_ACT, N, PZn_ACT, Fe	28	0.88	119.65	0.77	154.36

[1] All models are significant at p-value < 0.001.

The best overall performing SVR model utilized the top 28 variables, achieving an R^2 of 0.88 and an RMSE of 119.65 g/m^2 in the calibration set and an R^2 of 0.77 and an RMSE of 154.36 g/m^2 in the validation set. Comparing the performance of the model with the top 28 variables to that using all 42 variables (as detailed in Table 2-4), the former exhibited slightly better generalization capabilities, as indicated by a lower RMSE in the validation set. Therefore, the SVR model with 28 variables was considered superior due to its slightly better balance between training accuracy and validation error. Nonetheless, the differences in validation performance were minimal, suggesting that both models are relatively comparable in their ability to generalize.

2.4 Discussion

In this study, RF and SVM regression methods were used to predict the AGB of winter wheat, utilizing UAV multispectral MicaSense bands, associated VIs, plant biophysical parameters (plant height and the LAI), and plant biochemical parameters (nutrient content levels and ratios). During the first two weeks of sampling, the variation in AGB was low; however, it began to increase rapidly starting from the third sampling date, June 17. This rapid change aligns with the growth stages of winter wheat. On June 4 and 10, the plants were in their late heading and early flowering stages, respectively—a transition period marked by a slowdown in height increase due to shifts in developmental priorities and physiological changes. Initially, plant height and leaf area are major contributors to AGB as the stem elongates and leaves enlarge. By the heading stage, most stem elongation is complete, with culms extended and plant height largely established. What follows is

primarily the emergence of the inflorescence from the flag leaf's sheath, which does not significantly contribute to further height increase. As the plant transitions to the reproductive phase, its focus shifts from vegetative to reproductive growth, including the formation and maturation of the inflorescence. On June 17, the winter wheat entered the fruit development stage, channeling photosynthates from leaves and stems into the developing grains, which accumulate starch, proteins, and other nutrients, significantly increasing their weight and overall biomass. Finally, during the ripening stage on June 23, the grains transform from a watery, milky substance into a hard, dry state—a process marked by the continuous accumulation of dry matter, primarily starch, enhancing total AGB.

The RF and SVR models were initially calibrated using all 42 variables across single- and multi-date datasets. In the validation sets, models based on single dates generally performed poorly and were statistically non-significant. Our finding agrees with Atkinson Amorim et al.'s (2022) work that models using multiple date combinations performed better, especially if the combinations involved the last two dates. The best performance was observed in the RF model that utilized data from all four dates. According to the corresponding variable importance plot, the NDVI emerged as the most crucial variable, with other top-ranked variables primarily consisting of multispectral data. This finding aligns with Fu et al.'s findings that the NDVI and its narrowband-modified variations were effective predictors of biomass using partial least squares regression. Moreover, the top-ranked VIs in AGB estimation mainly comprised the NIR band, further supporting Fu et al.'s (2014) identification of NIR as a sensitive band region for AGB. In PA, employing UAVs equipped with spectral cameras is a common and cost-effective method to capture multispectral data and estimate AGB (Hassan et al., 2018; Wei et al., 2023; Zhang et al., 2018). Although previous studies have proposed new methods for estimating crop AGB using multispectral data, our study demonstrates that existing machine learning methods can also produce comparable results when increasing the variety of predictors, such as plant nutrient content levels and ratios. Furthermore, senescence might affect the use of multispectral data for crop biomass monitoring at post-flowering stages, potentially reducing model performance compared to pre-flowering

stages (Sharma et al., 2022). Our study did not encounter this issue, likely due to the inclusion of additional variables beyond multispectral data and derived VIs.

We successfully optimized the machine learning models by selecting the top-ranked variables from the variable importance plots. This approach is consistent with findings from other research, which has proven that machine learning is an effective method for predicting biomass in crops, such as wheat and oats (Atkinson Amorim et al., 2022; Sharma et al., 2022; Wang et al., 2022). While we were using a very similar camera setup installed on the UAV as Sharma et al. (2022), our findings proved that including biophysical and biochemical parameters in the analysis can significantly increase the RF and SVR models' accuracy. Similar to the results of Lu et al. (2019), the best-performing RF model proved more accurate than the top SVR model. We tested a total of 42 variables. Although the SVR model performed well, nearly matching the RF model, the latter was better in terms of performance and ease of use. In the RF model, the top twenty variables were selected, which included eight of the thirteen VIs, three of the five MicaSense bands, five of the fourteen nutrient content levels, three of the eight nutrient content ratios, and plant height as the model's performance started dropping with more variables being added. In comparison, Wang et al. (2022) reported higher RMSE values in their post-flowering stage analysis than ours in their linear regression, partial least squares regression (PLSR), and RF models. Our model performance was comparable with theirs, and the lower RMSE values in our models could be advantageous for making timely adjustments in fertilizer and water application recommendations. This is especially crucial in unstable climate conditions, where late-stage growth adjustments are necessary. The RF method, which utilizes multiple decision trees, seems particularly well-suited to handling a high number of variables and avoiding overfitting. As reported by Wang et al., RF outperformed linear regression and PLSR by having higher stability in predicting AGB during their study period. Our findings aligned with theirs, as RF proved to be the more stable model when handling a variety of data compared to SVR. Conversely, the SVR model required user hyper-tuning and used a kernel trick function to separate data into groups, relying on the radial distance between points to provide meaningful insights for the model. We believe that our research findings can be applied to North American wheat fields located at similar latitudes to southern Ontario, Canada. However, the model

may perform differently under varying agricultural practices and environmental conditions. Future applications of these findings should take these factors into account.

2.5 Conclusions

2.5.1 Contributions of Utilizing Multiple Categories of Variables in AGB Estimation

This study tested the effectiveness of multispectral data and biophysical and biochemical parameters in predicting the AGB of winter wheat using machine learning methods. Variables tested include UAV MicaSense bands, the derived VIs, plant height, the LAI, and plant nutrient content levels and ratios. The best result was obtained from the RF model with an R^2 of 0.81 and an RMSE of 149.95 g/m^2 using the top 20 variables, with a close to even split between spectral variables and nutrient content variables.

The inclusion of plant nutrient content levels and ratios as predictors in this study represents an advancement in the field of biomass estimation. Traditionally, these variables are not commonly utilized. The utilization of a lower-cost UAV multispectral camera setup, combined with biophysical and biochemical parameters, particularly at the later growth stages of winter wheat post-flowering, demonstrates a cost-effective method to predict AGB when urgent changes in late growth stages are needed to counteract unpredicted weather events, such as forest fire smoke and haze.

2.5.2 Limitations and Future Work

Though machine learning algorithms are capable of analyzing variables across categories, it is important to recognize the empirical nature of the machine learning models used. These models, by design, rely on existing datasets for validation and can only approximate the true AGB, which is only verifiable at harvest. This intrinsic limitation highlights the potential discrepancies between predicted and actual outcomes. Such limitations underscore the necessity for ongoing calibration and testing of these models under varied agricultural conditions and across different crop cycles to ensure their reliability and accuracy. The applications of these models can also be limited to dataset access, which is a common limitation in PA research because in situ measurements often

are required, and that comes with the associated intensive labor, costs, and conditions. Additionally, it is important to recognize that the variables tested are not the only associated factors that affect AGB. Variables such as weather conditions, soil properties, field topography, moisture supply, and more need to be considered to define the condition of the plants.

Due to constraints in data and time availability, this research could not be conducted earlier in the growth stages of the winter wheat. Future studies should consider extending the time span to investigate the models' effectiveness more comprehensively. Additionally, further exploration into the use of UAV-based spectral data for biomass estimation is suggested. This exploration should particularly focus on wavelengths or VIs that strongly correlate with plant nutrient levels and ratios. More importantly, it should emphasize hyperspectral bands, which have proven to be highly accurate in monitoring crop growth and estimating yield (Guo et al., 2023). Furthermore, high spatial and temporal resolution satellite imagery can serve as a viable alternative to UAV imagery, eliminating the need for UAVs. Examples such as Planetscope and VENμS both have frequent revisit periods and high spatial resolution for a local, field-scale study. We demonstrated the effectiveness of using plant nutrient content levels and ratios as parameters to estimate AGB in this study. Therefore, further research into non-destructive methods using remote sensing techniques to obtain these data is recommended for future biomass estimation studies. Since plant height has also been proven to be a reliable predictor in this study, integrating Real-Time Kinematic (RTK) UAVs or LiDAR-equipped UAVs could enhance the precision and quantity of height data collection across the entire field. Combining these technologies with biomass estimation models could lead to the development of highly accurate AGB estimation maps, providing a more detailed understanding of biomass and crop yield potential.

2.6 References

Atkinson Amorim, J. G., Schreiber, L. V., de Souza, M. R. Q., Negreiros, M., Susin, A., Bredemeier, C., Trentin, C., Vian, A. L., de Oliveira Andrades-Filho, C., Doering, D., & Parraga, A. (2022). Biomass estimation of spring wheat with machine learning methods using UAV-based multispectral imaging. *International Journal of Remote Sensing*, *43*(13), 4758–4773. https://doi.org/10.1080/01431161.2022.2107882

Bendig, J., Bolten, A., Bennertz, S., Broscheit, J., Eichfuss, S., & Bareth, G. (2014). Estimating Biomass of Barley Using Crop Surface Models (CSMs) Derived from UAV-Based RGB Imaging. *Remote Sensing*, *6*(11), Article 11. https://doi.org/10.3390/rs61110395

Bendig, J., Yu, K., Aasen, H., Bolten, A., Bennertz, S., Broscheit, J., Gnyp, M. L., & Bareth, G. (2015). Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *International Journal of Applied Earth Observation and Geoinformation*, *39*, 79–87. https://doi.org/10.1016/j.jag.2015.02.012

Blake-Kalff, M. M. A., Hawkesford, M. J., Zhao, F. J., & McGrath, S. P. (2000). Diagnosing sulfur deficiency in field-grown oilseed rape (Brassica napus L.) and wheat (Triticum aestivum L.). *Plant and Soil*, *225*(1–2), 95–107. https://doi.org/10.1023/A:1026503812267

Bryant, C. R., Smit, B., Brklacich, M., Johnston, T. R., Smithers, J., Chiotti, Q., & Singh, B. (2000). Adaptation in Canadian Agriculture to Climatic Variability and Change. In S. M. Kane & G. W. Yohe (Eds.), *Societal Adaptation to Climate Variability and Change* (pp. 181–201). Springer Netherlands. https://doi.org/10.1007/978-94-017-3010-5_10

Bukowiecki, J., Rose, T., & Kage, H. (2021). Sentinel-2 Data for Precision Agriculture?—A UAV-Based Assessment. *Sensors*, *21*(8), Article 8. https://doi.org/10.3390/s21082861

Cavender-Bares, J., Gamon, J. A., & Townsend, P. A. (Eds.). (2020). *Remote Sensing of Plant Biodiversity*. Springer Nature. https://doi.org/10.1007/978-3-030-33157-3

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27. https://doi.org/10.1145/1961189.1961199

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61–69. https://doi.org/10.1016/j.compag.2018.05.012

Daughtry, C. S. T., Walthall, C. L., Kim, M. S., de Colstoun, E. B., & McMurtrey, J. E. (2000). Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. *Remote Sensing of Environment*, *74*(2), 229–239. https://doi.org/10.1016/S0034-4257(00)00113-9

Di Bella, C. M., Paruelo, J. M., Becerra, J. E., Bacour, C., & Baret, F. (2004). Effect of senescent leaves on NDVI-based estimates of fAPAR: Experimental and modelling evidences. *International Journal of Remote Sensing*, *25*(23), 5415–5427. https://doi.org/10.1080/01431160412331269724

Fernandes, R., Butson, C., Leblanc, S., & Latifovic, R. (2003). Landsat-5 TM and Landsat-7 ETM+ based accuracy assessment of leaf area index products for Canada derived from SPOT-4 VEGETATION data. *Canadian Journal of Remote Sensing*, *29*(2), 241–258. https://doi.org/10.5589/m02-092

Fu, Y., Yang, G., Wang, J., Song, X., & Feng, H. (2014). Winter wheat biomass estimation based on spectral indices, band depth analysis and partial least squares regression using hyperspectral measurements. *Computers and Electronics in Agriculture*, *100*, 51–59. https://doi.org/10.1016/j.compag.2013.10.010

Gitelson, A. A., Gritz †, Y., & Merzlyak, M. N. (2003). Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal of Plant Physiology*, *160*(3), 271–282. https://doi.org/10.1078/0176-1617-00887

Gitelson, A. A., Viña, A., Arkebauer, T. J., Rundquist, D. C., Keydan, G., & Leavitt, B. (2003). Remote estimation of leaf area index and green leaf biomass in maize canopies. *Geophysical Research Letters*, *30*(5), 1248. https://doi.org/10.1029/2002GL016450

Gitelson, A., & Merzlyak, M. N. (1994). Quantitative estimation of chlorophyll-*a* using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology B: Biology*, *22*(3), 247–252. https://doi.org/10.1016/1011-1344(93)06963-4

Gómez, D., Salvador, P., Sanz, J., & Casanova, J. L. (2019). Potato Yield Prediction Using Machine Learning Techniques and Sentinel 2 Data. *Remote Sensing*, *11*(15), Article 15. https://doi.org/10.3390/rs11151745

Guo, Y., He, J., Zhang, H., Shi, Z., Wei, P., Jing, Y., Yang, X., Zhang, Y., Wang, L., & Zheng, G. (2024). Improvement of Winter Wheat Aboveground Biomass Estimation Using Digital Surface Model Information Extracted from Unmanned-Aerial-Vehicle-Based Multispectral Images. *Agriculture*, *14*(3), Article 3. https://doi.org/10.3390/agriculture14030378

Guo, Y., Xiao, Y., Hao, F., Zhang, X., Chen, J., de Beurs, K., He, Y., & Fu, Y. H. (2023). Comparison of different machine learning algorithms for predicting maize grain yield using UAV-based hyperspectral images. *International Journal of Applied Earth Observation and Geoinformation*, *124*, 103528. https://doi.org/10.1016/j.jag.2023.103528

Hassan, M. A., Yang, M., Rasheed, A., Jin, X., Xia, X., Xiao, Y., & He, Z. (2018). Time-Series Multispectral Indices from Unmanned Aerial Vehicle Imagery Reveal Senescence

Rate in Bread Wheat. *Remote Sensing*, *10*(6), Article 6. https://doi.org/10.3390/rs10060809

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1), 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, *25*(3), 295–309. https://doi.org/10.1016/0034-4257(88)90106-X

Jordan, C. F. (1969). Derivation of Leaf-Area Index from Quality of Light on the Forest Floor. *Ecology*, *50*(4), 663–666. https://doi.org/10.2307/1936256

Kaufman, Y. J., & Tanre, D. (1992). Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. *IEEE Transactions on Geoscience and Remote Sensing*, *30*(2), 261–270. IEEE Transactions on Geoscience and Remote Sensing. https://doi.org/10.1109/36.134076

Koerselman, W., & Meuleman, A. F. M. (1996). The Vegetation N:P Ratio: A New Tool to Detect the Nature of Nutrient Limitation. *Journal of Applied Ecology*, *33*(6), 1441–1450. https://doi.org/10.2307/2404783

Li, W., Niu, Z., Huang, N., Wang, C., Gao, S., & Wu, C. (2015). Airborne LiDAR technique for estimating biomass components of maize: A case study in Zhangye City, Northwest China. *Ecological Indicators*, *57*, 486–496. https://doi.org/10.1016/j.ecolind.2015.04.016

Liao, C., Wang, J., Shan, B., Song, Y., He, Y., & Dong, T. (2022). Near real-time yield forecasting of winter wheat using Sentinel-2 imagery at the early stages. *Precision Agriculture*, *24*(3), 807–829. https://doi.org/10.1007/s11119-022-09975-3

Liu, J., Pattey, E., Miller, J. R., McNairn, H., Smith, A., & Hu, B. (2010). Estimating crop stresses, aboveground dry biomass and yield of corn using multi-temporal optical data combined with a radiation use efficiency model. *Remote Sensing of Environment*, *114*(6), 1167–1177. https://doi.org/10.1016/j.rse.2010.01.004

Marschner, H. (2011). *Marschner's Mineral Nutrition of Higher Plants* (3rd ed.). Academic Press.

Novoa, R., & Loomis, R. S. (1981). Nitrogen and plant production. *Plant and Soil*, *58*(1), 177–204. https://doi.org/10.1007/BF02180053

Ontario Ministry of Agriculture, Food and Rural Affairs. (2023, November 8). *Census farm data collection*. Ontario Data Catalogue. https://data.ontario.ca/dataset/census-farm-data-collection

Oosterhuis, D. M., Loka, D. A., Kawakami, E. M., & Pettigrew, W. T. (2014). The Physiology of Potassium in Crop Production. In D. L. Sparks (Ed.), *Advances in Agronomy* (Vol. 126, pp. 203–233). Elsevier. https://doi.org/10.1016/B978-0-12-800132-5.00003-1

Pagani, A., & Echeverría, H. E. (2011). Performance of Sulfur Diagnostic Methods for Corn. *Agronomy Journal*, *103*(2), 413–421. https://doi.org/10.2134/agronj2010.0265

Qi, J., Chehbouni, A., Huete, A. R., Kerr, Y. H., & Sorooshian, S. (1994). A modified soil adjusted vegetation index. *Remote Sensing of Environment*, *48*(2), 119–126. https://doi.org/10.1016/0034-4257(94)90134-1

Radoglou-Grammatikis, P., Sarigiannidis, P., Lagkas, T., & Moscholios, I. (2020). A compilation of UAV applications for precision agriculture. *Computer Networks*, *172*, 107148. https://doi.org/10.1016/j.comnet.2020.107148

Reid, S., Smit, B., Caldwell, W., & Belliveau, S. (2007). Vulnerability and adaptation to climate risks in Ontario agriculture. *Mitigation and Adaptation Strategies for Global Change*, *12*(4), 609–637. https://doi.org/10.1007/s11027-006-9051-8

Rondeaux, G., Steven, M., & Baret, F. (1996). Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment*, *55*(2), 95–107. https://doi.org/10.1016/0034-4257(95)00186-7

Roujean, J.-L., & Breon, F.-M. (1995). Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. *Remote Sensing of Environment*, *51*(3), 375–384. https://doi.org/10.1016/0034-4257(94)00114-3

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. *NASA Special Publications*, *351*(1), 309.

Sharma, P., Leigh, L., Chang, J., Maimaitijiang, M., & Caffé, M. (2022). Above-Ground Biomass Estimation in Oats Using UAV Remote Sensing and Machine Learning. *Sensors*, *22*(2), Article 2. https://doi.org/10.3390/s22020601

Shi, Q., Pang, J., Yong, J. W. H., Bai, C., Pereira, C. G., Song, Q., Wu, D., Dong, Q., Cheng, X., Wang, F., Zheng, J., Liu, Y., & Lambers, H. (2020). Phosphorus-fertilisation has differential effects on leaf growth and photosynthetic capacity of Arachis hypogaea L. *Plant and Soil*, *447*(1), 99–116. https://doi.org/10.1007/s11104-019-04041-w

Silleos, N. G., Alexandridis, T. K., Gitas, I. Z., & Perakis, K. (2006). Vegetation Indices: Advances Made in Biomass Estimation and Vegetation Monitoring in the Last 30 Years. *Geocarto International*, *21*(4), 21–28. https://doi.org/10.1080/10106040608542399

Sishodia, R. P., Ray, R. L., & Singh, S. K. (2020). Applications of Remote Sensing in Precision Agriculture: A Review. *Remote Sensing*, *12*(19), Article 19. https://doi.org/10.3390/rs12193136

Tan, C. S., & Reynolds, W. D. (2003). Impacts of Recent Climate Trends on Agriculture in Southwestern Ontario. *Canadian Water Resources Journal*, *28*(1), 87–97. https://doi.org/10.4296/cwrj2801087

Tausch, R. J. (1989). Comparison of Regression Methods for Biomass Estimation of Sagebrush and Bunchgrass. *The Great Basin Naturalist*, *49*(3), 373–380.

Tian, J., Wang, S., Zhang, L., Wu, T., She, X., & Jiang, H. (2015). Evaluating different vegetation index for estimating lai of winter wheat using hyperspectral remote sensing data. *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 1–4. https://doi.org/10.1109/WHISPERS.2015.8075437

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

Wang, F., Yang, M., Ma, L., Zhang, T., Qin, W., Li, W., Zhang, Y., Sun, Z., Wang, Z., Li, F., & Yu, K. (2022). Estimation of Above-Ground Biomass of Winter Wheat Based on Consumer-Grade Multi-Spectral UAV. *Remote Sensing*, *14*(5), Article 5. https://doi.org/10.3390/rs14051251

Wei, L., Yang, H., Niu, Y., Zhang, Y., Xu, L., & Chai, X. (2023). Wheat biomass, yield, and straw-grain ratio estimation from multi-temporal UAV-based RGB and multispectral images. *Biosystems Engineering*, *234*, 187–205. https://doi.org/10.1016/j.biosystemseng.2023.08.002

Wu, C., Niu, Z., Tang, Q., & Huang, W. (2008). Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. *Agricultural and Forest Meteorology*, *148*(8), 1230–1241. https://doi.org/10.1016/j.agrformet.2008.03.005

Xie, Q., Huang, W., Liang, D., Chen, P., Wu, C., Yang, G., Zhang, J., Huang, L., & Zhang, D. (2014). Leaf Area Index Estimation Using Vegetation Indices Derived From Airborne Hyperspectral Images in Winter Wheat. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(8), 3586–3594. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. https://doi.org/10.1109/JSTARS.2014.2342291

Yu, J., Wang, J., Leblon, B., & Song, Y. (2022). Nitrogen Estimation for Wheat Using UAV-Based and Satellite Multispectral Imagery, Topographic Metrics, Leaf Area Index, Plant Height, Soil Moisture, and Machine Learning Methods. *Nitrogen*, *3*(1), Article 1. https://doi.org/10.3390/nitrogen3010001

Zhang, Y., Qin, Q., Ren, H., Sun, Y., Li, M., Zhang, T., & Ren, S. (2018). Optimal Hyperspectral Characteristics Determination for Winter Wheat Yield Prediction. *Remote Sensing*, *10*(12), Article 12. https://doi.org/10.3390/rs10122015

Zhao, S., Lü, J., Xu, X., Lin, X., Luiz, M. R., Qiu, S., Ciampitti, I., & He, P. (2021). Peanut yield, nutrient uptake and nutrient requirements in different regions of China. *Journal of*

Integrative Agriculture, *20*(9), 2502–2511. https://doi.org/10.1016/S2095-3119(20)63253-1

Zhu, Y., Liu, J., Tao, X., Su, X., Li, W., Zha, H., Wu, W., & Li, X. (2023). A Three-Dimensional Conceptual Model for Estimating the Above-Ground Biomass of Winter Wheat Using Digital and Multispectral Unmanned Aerial Vehicle Images at Various Growth Stages. *Remote Sensing*, *15*(13), Article 13. https://doi.org/10.3390/rs15133332

Chapter 3

3 Local Field-Scale Winter Wheat Yield Prediction Using VENμS Satellite Imagery and Machine Learning Techniques

3.1 Introduction

The growing global population has heightened the need for reliable food sources and food security, underscoring the importance of advancing efficient and sustainable agricultural practices. The agriculture industry today faces substantial challenges, including rising global food demand, crop diseases, pest outbreaks, limited arable land, and the impacts of climate change. Addressing these issues is vital for ensuring a resilient and productive agricultural sector. Research by Tan and Reynolds indicates that in southwestern Ontario, water supply and demand pose the greatest challenge to the agricultural sector (Hewer & Brunette, 2020). Interestingly, farmers in this region are less concerned about cli-mate change compared to those in areas more frequently affected by extreme weather events (Reid et al., 2007). The agriculture and agri-food sector contributed approximately 7% to Canada's gross domestic product (GDP) and accounted for one in every nine jobs in 2023 (Agriculture and Agri-Food Canada, 2024). While climate change may not present an immediate threat to the Canadian agricultural industry, it is wise to stay informed and proactively prepare for potential future climate variations.

Precision agriculture (PA) employs advanced technologies and data analysis techniques to optimize crop yields while minimizing resource use. This approach involves evaluating quantified spatial and in-situ plant data to inform agricultural practices such as the application of water, labor, and fuel, thereby reducing costs and preventing excessive waste, including pesticide and nutrient loss. PA integrates various spatial technologies, such as geographic information systems (GIS), handheld ground-based data collection devices, and remote sensing through ground-based or aerial vehicles, to develop and implement efficient agricultural strategies (Chlingaryan et al., 2018). Given the high demand for data collection, remote sensing techniques are employed in crop management

to precisely manage, produce, and predict crop data for analysis. Accurate crop yield prediction is crucial for helping farmers address production challenges and mitigate the effects of climate variability and change on crop yield (Hammer, 2000).

Among the various platforms of surface spectral data collection in PA, space-borne satellites are one of the most stable platforms (Liao et al., 2023; Shafi et al., 2019; Skakun et al., 2018; Yu et al., 2022; Zhang et al., 2020; Zhao et al., 2020). A key advantage of using optical satellite images for remote sensing is the ability to obtain spectral data over large land areas in a single snapshot with high resolution. Traditionally, researchers have faced challenges with optical satellite images due to their relatively lower spatial resolution compared to ground-collected data (Fu et al., 2020). This limitation has restricted research to regional scales rather than local, field-scale studies. For example, Landsat 8, launched in 2013 by the United States, features the Optical Land Imager (OLI) with a spatial resolution of up to 30 m (United States Geological Survey [USGS], 2019). Similarly, Sentinel-2, launched in 2015, features 13 multispectral bands with spatial resolutions of 10 m, 20 m, and 60 m, and a revisit time of 5 days with its constellation of twin satellites (European Space Agency [ESA], 2015). In contrast, VENμS's VSSC (Vegetation and Environment monitoring on a New Micro-Satellite Super-spectral Camera) captures optical images at a resolution as high as 5.3 m. Additionally, VENμS has a revisit time of 2 days, compared to Landsat 8's 16 days (USGS, 2019; Centre National d'Etudes Spatiales [CNES] & Israeli Space Agency [ISA], 2023). These advantages in both high spatial and temporal resolution make VENμS a superior choice for detailed crop monitoring and analyses, providing more frequent and precise data for agricultural applications.

The ease of access to satellite data offers a significant advantage over UAV-level remote sensing. Many satellite datasets, such as those from VENμS, Landsat series, Sentinel-2, MODIS, and SPOT series are publicly available. VENμS imagery can be downloaded free of charge in its predefined areas, thereby reducing both labor and monetary research costs compared to ground sampling and UAV flight operations. While crop monitoring has traditionally relied on satellite imagery, UAV-based systems often challenge their usability due to superior spatial and temporal resolution. Crop growth stages can vary

week to week, making some satellite images unsuitable for timely analysis. For instance, Sentinel-2 data have yielded unsatisfactory crop yield prediction results due to cloud coverage and lower temporal resolution (Bukowiecki et al., 2021). VENμS addresses this issue by providing higher spatial resolution data compared to most satellites, while maintaining frequent revisits of 2 days and offering a wide range of multispectral bands (CNES & ISA, 2023).

Furthermore, UAV operations are often constrained by weather conditions. Clear skies and low wind speeds are typically required to collect high-quality data. While UAVs offer flexible planning and scheduling, VENμS can achieve similar advantages by mitigating poor coverage with its short revisit period. Despite this, UAVs provide an edge over satellites by allowing researchers greater control over the location and timing of data collection. However, UAV flights with payload such as multispectral cameras are often restricted under aviation regulations, and additional procedures or certifications are often required if the flight is to be conducted in a regulated aerodrome in most countries. For instance, Transport Canada mandates the registration of any remotely piloted aircrafts (RPAs) weighing between 250 g and 25 kg, which encompasses most commercially available UAVs that can carry spectral sensors as pay-loads (Transport Canada, 2023). Additionally, operating these RPAs categorized by Transport Canada re-quires the pilot to have different classes of operation licenses based on the location of flight. In contrast, satellite data can often be obtained online free of charge and without any operational requirements, making the data widely accessible. Thus, VENμS effectively combines the advantages of both satellite and UAV systems, offering high spatial resolution, frequent temporal coverage, and ease of data access.

With the spectral data collected from remote sensing imagery, vegetation index (VI) calculations become feasible. VIs are mathematical transformations of spectral bands widely used in agricultural research to determine specific plant properties, such as leaf area index (LAI), chlorophyll content, and nutrient levels (Wu et al., 2008; Xie et al., 2014; Yu et al., 2022). Consequently, VIs are commonly employed for crop growth and health monitoring, including yield prediction (Silleos et al., 2006; Yu et al., 2022). For instance, vegetation indices that performed well in the study by Fu et al. were derived

using the red absorption portion of the spectrum (Fu et al., 2014). On multispectral cameras, this typically includes the red band and red-edge bands. Indices such as the normalized difference vegetation index (NDVI), normalized difference red edge (NDRE), and soil-adjusted vegetation index (SAVI) have been previously studied as effective indices in winter wheat yield monitoring (Fu et al., 2020; Panek et al., 2020). VENμS is specifically designed for vegetation monitoring, offering more bands in the red-edge and near-infrared range than most publicly available satellite data. This enhanced spectral capability improves its ability to detect vegetation properties. Therefore, it is important to further explore VENμS's yield prediction potential using a more diverse range of vegetation indices that may be unavailable from other satellites.

Recently, machine learning regression methods, such as Random Forest (RF) and Support Vector Regression (SVR), have been extensively investigated for biomass and yield estimation (Atkinson Amorim et al., 2022; Chlingaryan et al., 2018; van Klompenburg et al., 2020; Wang et al., 2022). These machine learning methods can capture complex patterns and relationships in the data that traditional methods might miss and was proven to be viable in yield prediction (Han et al., 2020; Nigam et al., 2019). RF, for example, can handle a large number of input variables and is less likely to over-fit due to its ensemble nature. Hunt et al. successfully mapped winter wheat yield using Sentinel-2 data and RF regression models, achieving a relatively low root mean square error at 0.66 t/ha (Hunt et al., 2019). This work suggests the potential of utilizing higher spatial resolution data to capture the within-field yield variability with a common machine learning algorithm.

SVR, conversely, focuses on optimizing a margin around a hyperplane, which can result in better generalization on unseen data. While traditional regression methods are straightforward and easier to interpret, machine learning regression methods like RF and SVR offer significant advantages in terms of handling complexity, scalability, and adaptability, making them suitable for a wide range of modern data-driven applications.

Compared to most publicly available satellites, such as Sentinel-2, VENμS offers additional bands in the red-edge and near-infrared ranges, which are particularly

advantageous for vegetation monitoring. It also provides relatively higher spatial and temporal resolution. Despite these benefits, VENμS has been rarely studied in yield estimation research. Therefore, to make a well-informed prediction of winter wheat yield at a local, field scale using VENμS data, it is essential to introduce an appropriate prediction model. The objective of this study is to (i) investigate the relationships between yield and VIs at difference growth stages; (ii) evaluate the effectiveness of RF and SVR models in predicting yield; (iii) determine the optimal combinations of dates (growth stages) for yield prediction in a winter wheat field located in southwestern Ontario; (iv) uncover insights in the ranked im-portance of VIs from different growth stages; and (v) produce a yield prediction map.

3.2 Materials and Methods

3.2.1 Study Area and Data Collection

The study site is in Strathroy-Caradoc, Ontario, Canada, near the village of Mount Brydges, which is about 23 km southwest of the urban center of London, Ontario (Figure 3-1). The studied period was in May to early July of 2020, during which the average temperature was recorded at 22 °C and the relative humidity averaging 73%. The climate in the area is classified as warm-summer humid continental climate (Dfb) according to the Köppen climate classification system. The area is predominantly agricultural croplands and its major field crops include winter wheat, corn, and soybeans (Ontario Ministry of Agriculture, Food and Rural Affairs, 2023). Winter wheat was selected as the focus of this study. A winter wheat field covering 53.7 hectares in this region was designated as the specific area for investigation

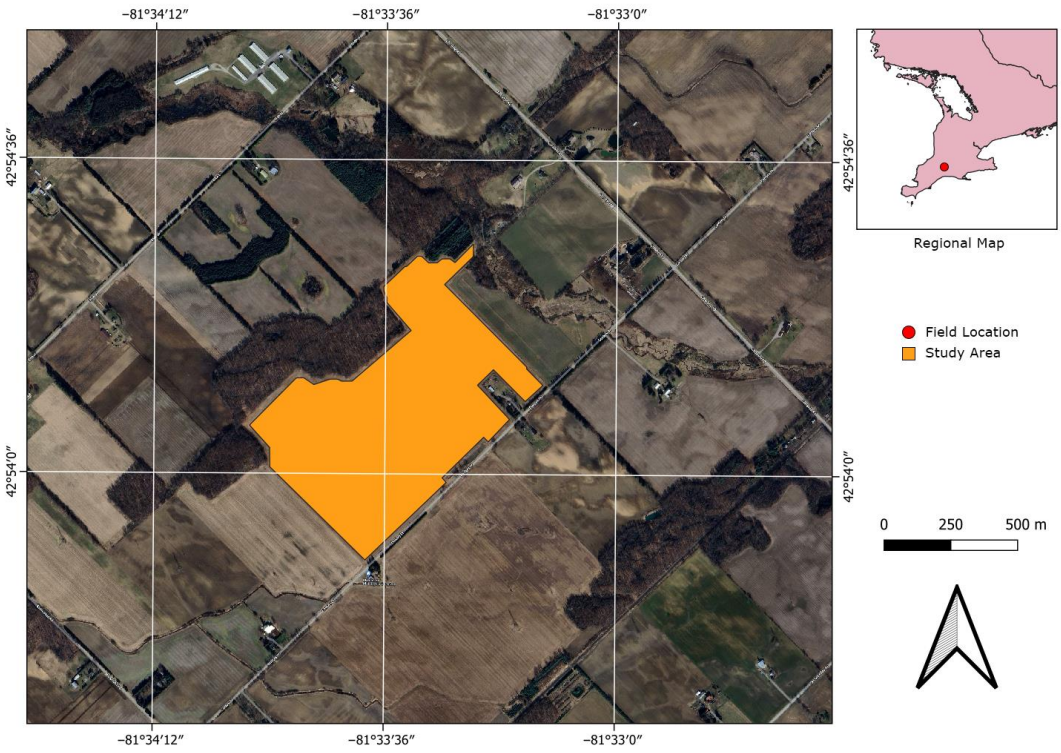


Figure 3-1. Location of the studied wheat field near Melbourne, ON, Canada over an ArcGIS Pro Basemap Image.

The cultivar in the studied field was soft red winter wheat, which was planted in October of 2019. In the region of Southwest Ontario, winter wheat typically lies dormant over the winter after planting, then commences shooting in late April of the following year and is harvested from early to mid-July. VENμS imagery acquisition was performed at each consequent growth stages starting at tillering, then stem elongation, booting, heading, flowering, early fruit development, and ripening. The growth stages were verified by gauging the plant's physical characteristics using the Biologische Bundesanstalt, Bundessortenamt and CHemical industry (BBCH) scale at the field, matching the satellite overpass dates (Table 3-1). Unfortunately, the satellite imagery taken on the date of visiting the field during late fruit development stage was later found to be covered by cloud shadows. In total, 8 cloud-free VENμS imagery was acquired.

Table 3-1. Growth stages at the study area with matching VENμS overpass dates.

Growth Stage	VENμS Overpass

Tillering - 1	20200503
Tillering - 2	20200513
Stem Elongation	20200521
Booting	20200525
Heading	20200606
Flowering	20200612
Early Fruit (Grain) Development	20200616
Late Fruit (Grain) Development	Cloud Cover
Ripening	20200706

3.2.2 VENμS Satellite Imagery and Preprocessing

The data used in this research was collected by VENμS (Vegetation and Environment monitoring on a New MicroSatellite), which was launched in August 2017. This satellite marks the first Earth observation collaboration between France and Israel, led by the Centre National d'Etudes Spatiales (CNES) and the Israeli Space Agency (ISA). The mission aims to monitor plant growth and health status, providing valuable in-sights into the impacts of environmental factors, human activities, and climate change on Earth's land surface (CNES & ISA, 2023). Since 2017, the VENμS VM1 mission has provided multispectral data from its 12 different bands, featuring a spatial resolution of 5.3 meters, a revisiting period of 2 days, and operating at an altitude of 720 kilometers above sea level (Table 3-2). As implied by its name, VENμS excels in monitoring Earth's surface vegetation, which is facilitated by its extensive red-edge and near-infrared bands.

Table 3-2. Spectral bands of the VENμS super-spectral camera.

Bands	Central Wavelength (nm)	Bandwidth (nm)
1	423.9	40
2	446.9	40
3	491.9	40
4	555	40
5	619.7	40
6	619.5	40
7	666.2	30
8	702	24
9	741.1	16
10	782.2	16
11	861.1	40
12	908.7	20

The imagery was categorized as level 2A (L2A) surface reflectance data, which each scene covering areas ranging from 27 × 27 km² to 27 × 54 km² at the spatial resolution of 5 × 5 m². The satellite imagery was processed and distributed by Theia MUSCATE (MUlti SATellite, multi-CApteurs, for multi-TEmporelles data), a component of the Theia Land Data Centre. This French inter-agency organization aims to provide satellite data and value-added products for scientific communities and public policy actors. MUSCATE facilitates the processing and distribution of large volumes of satellite imagery, particularly from VENμS, Sentinel-2, and Landsat satellites. This includes tasks such as atmospheric corrections and creating cloud-free surface reflectance syntheses. The processed data are used in various applications, including agriculture, forestry, urban planning, and environmental monitoring. Each of the VENμS L2A products contained two versions of surface reflectance data for the 12 bands, from B01 to B12. The first version of the surface reflectance rasters is denoted as SRE.DBL (Surface Reflectance), which is atmospherically corrected. The second version is denoted as FRE.DBL. (Flat REflectance), which are SRE.DBL files further corrected for slope effects. This correction suppresses apparent reflectance variations due to the orientation of slopes with regard to the sun, making the corrected image appear as if the land surface were flat. For this study, the FRE.DBL raster files were adopted.

The L2A surface reflectance rasters were encoded as 16-bit signed integers, necessitating preprocessing before any manipulation by dividing pixel values of each channel by 1000. This preprocessing was conducted in Python 3.9.19 using packages such as “rasterio”, “gdal”, and “numpy” to extract and obtain surface reflectance values from each band at the study site. Subsequently, the 12-band raster values were normalized to a range between 0 and 1 for use in later calculations.

3.2.3 Vegetation Indices

Vegetation indices (VIs) were used in this study as predictors of final harvested yield. The VIs were calculated as raster products using the 12 VENμS bands in Python 3.9.19, employing the same packages used in satellite image preprocessing. Additionally, several VIs that utilize spectral information in the red edge and near-infrared wavelengths, which are well-represented in VENμS data, have demonstrated strong correlations with crop

growth, health, and yield (Cao et al., 2016; Xie et al., 2014; Zhang et al., 2018). A total of 21 VIs were tested in this study, including 8 variations of existing VIs based on their original development formulas (Table 3-3). This was made possible by fitting the narrow bandwidth of VENμS bands into VI formulas initially developed with legacy sensors and satellites. For instance, NDVI was developed using the Landsat-1 Multi-spectral Scanner, where the NIR band 7 had a bandwidth range of 800 to 1100 nm. With the VSSC, both bands 11 and 12 fit within this NIR range, allowing for the inclusion of variations of existing VIs in the analysis.

Table 3-3. Vegetation indices to be tested in this study.

VI [1]	Formula [2]	Original Authors
ARVI	$\frac{\mathrm{NIR}_{11} - [\mathrm{Red}_7 - 1 \times (\mathrm{Red}_7 - \mathrm{Blue}_3)]}{\mathrm{NIR}_{11} + [\mathrm{Red}_7 1 \times (\mathrm{Red}_7 - \mathrm{Blue}_3)]}$	Kaufman and Tanre, 1992
DVI-1	$\mathrm{NIR}_{11} - \mathrm{Red}_7$	Richardson and Wiegand,1977
DVI-2	$\mathrm{NIR}_{12} - \mathrm{Red}_7$	
EVI	$\frac{2.5 \times (\mathrm{NIR}_{11} - \mathrm{Red}_7)}{\mathrm{NIR}_{11} + 6 \times \mathrm{Red}_7 - 7.5 \times \mathrm{Blue}_2 + 1}$	Huete et al., 2002
ISR-1	$\frac{\mathrm{Red}_7}{\mathrm{NIR}_{11}}$	Fernades et al., 2003
ISR-2	$\frac{\mathrm{Red}_7}{\mathrm{NIR}_{12}}$	
MCARI	$[(\mathrm{RE}_8 - \mathrm{Red}_7) - 0.2 \times (\mathrm{RE}_8 - \mathrm{Green}_4)] \times \mathrm{RE}_8 \div \mathrm{Red}_7$	Daughtry et al., 2000
MSAVI-1	$[2 \times \mathrm{NIR}_{10} + 1 - \sqrt{(2 \times \mathrm{NIR}_{10} + 1)^2 - 8 \times (\mathrm{NIR}_{10} - \mathrm{Red}_7)}\,] \div 2$	Qi et al., 1994
MSAVI-2	$[2 \times \mathrm{NIR}_{11} + 1 - \sqrt{(2 \times \mathrm{NIR}_{11} + 1)^2 - 8 \times (\mathrm{NIR}_{11} - \mathrm{Red}_7)}\,] \div 2$	
NDRE-1	$\frac{(\mathrm{NIR}_{10} - \mathrm{RE}_8)}{(\mathrm{NIR}_{10} + \mathrm{RE}_8)}$	Gitelson and Merzlyak, 1994
NDRE-2	$\frac{(\mathrm{NIR}_{10} - \mathrm{RE}_9)}{(\mathrm{NIR}_{10} + \mathrm{RE}_9)}$	
NDVI-1	$\frac{(\mathrm{NIR}_{11} - \mathrm{Red}_7)}{(\mathrm{NIR}_{11} + \mathrm{Red}_7)}$	Rouse et al., 1974
NDVI-2	$\frac{(\mathrm{NIR}_{12} - \mathrm{Red}_7)}{(\mathrm{NIR}_{12} + \mathrm{Red}_7)}$	
OSAVI	$[1.16 \times (\mathrm{NIR}_{11} - \mathrm{Red}_7)] \div (\mathrm{NIR}_{11} + \mathrm{Red}_7 + 0.16)]$	Rondeaux et al., 1996
RDVI	$(\mathrm{NIR}_{11} - \mathrm{Red}_7) \div (\sqrt{(\mathrm{NIR}_{11} + \mathrm{Red}_7)})$	Roujean and Breon, 1995
REP	$702 + 40\left(\frac{\left(\frac{\mathrm{Red}_7 + \mathrm{NIR}_{10}}{2}\right) - \mathrm{Red}_8}{\mathrm{Red}_9 - \mathrm{Red}_8}\right)$	Guyot and Baret, 1988

RVI-1	$\frac{\mathrm{NIR}_{11}}{\mathrm{Red}_{7}}$	Jordan, 1969
RVI-2	$\frac{\mathrm{NIR}_{12}}{\mathrm{Red}_{7}}$	
SAVI-1	$\frac{(\mathrm{NIR}_{10}-\mathrm{Red}_{7})}{(\mathrm{NIR}_{10}+\mathrm{Red}_{7}+0.5)}(1.5)$	Huete, 1988
SAVI-2	$\frac{(\mathrm{NIR}_{11}-\mathrm{Red}_{7})}{(\mathrm{NIR}_{11}+\mathrm{Red}_{7}+0.5)}(1.5)$	
SAVI-3	$\frac{(\mathrm{NIR}_{12}-\mathrm{Red}_{7})}{(\mathrm{NIR}_{12}+\mathrm{Red}_{7}+0.5)}(1.5)$	

[1] ARVI, atmospherically resistant vegetation index; DVI-1, 2, difference vegetation index; EVI, enhanced vegetation index; ISR-1, 2, infrared simple ratio; MCARI, modified chlorophyll absorption in reflectance index; MSAVI-1, 2, modified soil-adjusted vegetation index; NDRE-1, 2, normalized difference red edge; NDVI-1, 2, normalized difference vegetation index; OSAVI, optimized soil-adjusted vegetation index; RDVI, renormalized difference vegetation index; REP, red edge position; RVI-1, 2, ratio vegetation index; SAVI-1, 2, 3, soil-adjusted vegetation index.
[2] Blue, blue reflectance; green, green reflectance; red, red reflectance; RE, red edge reflectance; NIR, near-infrared reflectance. Subscripts are the equivalent VENμS bands.

3.2.4 Yield Dataset

The yield data was collected at harvest on July 25, 2020, with a combine harvester equipped with a 10-meter wide and 1.5-meter-long header. Yield data was generated as point shapefile, with yield data recorded approximately every second at the center of the harvester's track. To ensure accuracy, potential outliers located at the edges of the field were removed. For this study, the shapefile was interpolated into a 5 × 5 m² spatial resolution raster using QGIS 3.22 with inverse distance weighted (IDW) interpolation, matching the VENμS imagery and the derived vegetation indices (VIs). This approach was adopted to fully utilize the high-resolution advantage of VENμS data and to produce a detailed yield prediction map.

3.2.5 Machine Learning Regression Modelling and Cross-Validation

In machine learning, regression models are used to predict continuous outcomes based on input variables. Two notable techniques in this domain are Random Forest (RF) regression and Support Vector Regression (SVR), both of which offer robust solutions for complex regression problems. Advantages of machine learning regression also includes its ability to automatically learn from data without being explicitly programmed

for each specific task. Given that the regression models in this study were based on pixel-level analysis, machine learning regression methods were ideal for our needs as they excel in handing large data sizes. In our study, we used three key metrics to evaluate the performance of our regression models: Mean Absolute Error (MAE), R-squared (R^2), and Root Mean Squared Error (RMSE). These metrics were employed during both the cross-validation stage and the calibration and validation of the final model to ensure a comprehensive assessment of model accuracy and reliability.

RF is an ensemble learning method that constructs multiple decision trees during calibration and outputs the mean prediction of these trees. By using multiple trees, RF reduces overfitting, a common issue in single decision tree models. Each tree is built from a random sample of the data, with a random subset of features selected at each node to decide splits. This randomness helps make the model more resilient to noise and outliers. RF can handle large, high-dimensional datasets and identify important variables in the modeled relationships. Additionally, RF provides measures of feature importance, helping to understand the impact of each variable on the prediction.

SVR, on the other hand, extends the concepts of Support Vector Machines (SVMs) from classification to regression. Like RF, SVR is also generally robust to over fitting. It is a result of its margin maximization, the use of kernel functions, the epsilon-insensitive loss function, and the reliance on support vectors. Unlike traditional methods that minimize the error between predicted and observed values, SVR attempts to fit the error within a certain threshold. It involves the creation of a hyperplane in a multidimensional space where the distance between the data points and the hyperplane is minimized, ensuring that errors do not exceed a defined threshold. This makes SVR particularly useful in cases where a margin of tolerance is specified in the predictions. SVR is highly effective in handling non-linear relationships through the use of kernel functions, making it adaptable to various types of data (Chang & Lin, 2011).

In this study, the data collected over 8 dates were randomly divided into a 70% calibration set and a 30% validation set. A 10-fold K-fold cross-validation approach was employed in this study to ensure the robustness and generalizability of the machine

learning regression models. This method involved splitting the calibration data into multiple subsets (folds), using each subset in turn as the validation set while the remaining data was used for training. With 10 folds, each fold uses 90% of the data for training and 10% for validation. This approach ensured that each training set was large enough to effectively train the model, while each validation set was sufficient to provide a reliable evaluation without overfitting. During the cross-validation stage, MAE, R^2, and RMSE served as crucial indicators of model performance. Cross-validation involved partitioning the calibration dataset into multiple folds and iteratively training and validating the model on these folds. MAE provided the average magnitude of errors in the predictions, indicating the overall accuracy of the model without considering the direction of errors. RMSE, which penalized larger errors more significantly due to its squared component, offered insight into the model's ability to handle large deviations from observed values. R^2, representing the proportion of variance explained by the model, evaluated the goodness of fit, with values closer to 1 indicated a better fit. By averaging these metrics across all folds, we obtained a robust estimate of the model's performance and its variability, thus mitigating the risk of overfitting or underfitting to specific subsets of data. For the RF models, the RMSE dictated the optimal cross-validated RF model with an optimal number of splits at each tree node. The MAE value of that optimal RF model represents its average magnitude of the errors in the prediction. MAE served the same purpose in the SVR models, but RMSE determined the optimal SVR model with the optimal regularization parameter. The equations are as follows:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (1)$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and $\bar{y}_i$ is the mean of the observed values, and;

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

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \quad (3)$$

where $\hat{y}_i$ represents the predicted yield (t/ha), y_i denotes the observed yield (t/ha), n is the total number of observations, and i serves as the summation index, incrementing by one.

After cross-validation, the final model was trained on the entire cross-validated calibration dataset and then evaluated on both the calibration and validation datasets using the same metrics. In the calibration stage, RMSE assessed the model's fit to the data it was trained on, while R^2 measures how well the model captures the underlying data patterns. High R^2 values, coupled with low RMSE, suggest a good fit. However, it is crucial to compare these metrics with those from the validation stage. The validation stage involved assessing the model on unseen data, providing an indication of its generalization ability. Consistent performance across calibration and validation sets, characterized by similar R^2, and RMSE values, indicates a robust model.

Figure 3-2 displays the workflow of the methodology. The modeling was written in R programming language using RStudio by utilizing packages such as “randomForest” and “e1071” for RF and SVR, respectively. In both models, the independent variables were the VIs. Data collected over the 8 dates were ran individually, then divided into two groups of “pre-heading” and “post-heading”. For each dataset, a 10-fold cross-validation was performed using packages “caret” and “kernlab”. In the RF models, using the default setting of 500 decision trees yielded the best results. Additionally, the cross-validation process determined the optimal mtry to consider at each split for each model. The optimal mtry value varied as different combinations of the data were tested during cross-validation. In SVR, RBF was identified as the most suitable. The model's tuning parameters were automatically adjusted, optimizing for the cost value that produced the lowest RMSE, while the sigma (σ) parameter, which controls the width of the RBF kernel, was kept constant. The yield prediction raster was also created using the “raster”, “sp”, and “rasterVis” packages in RStudio.

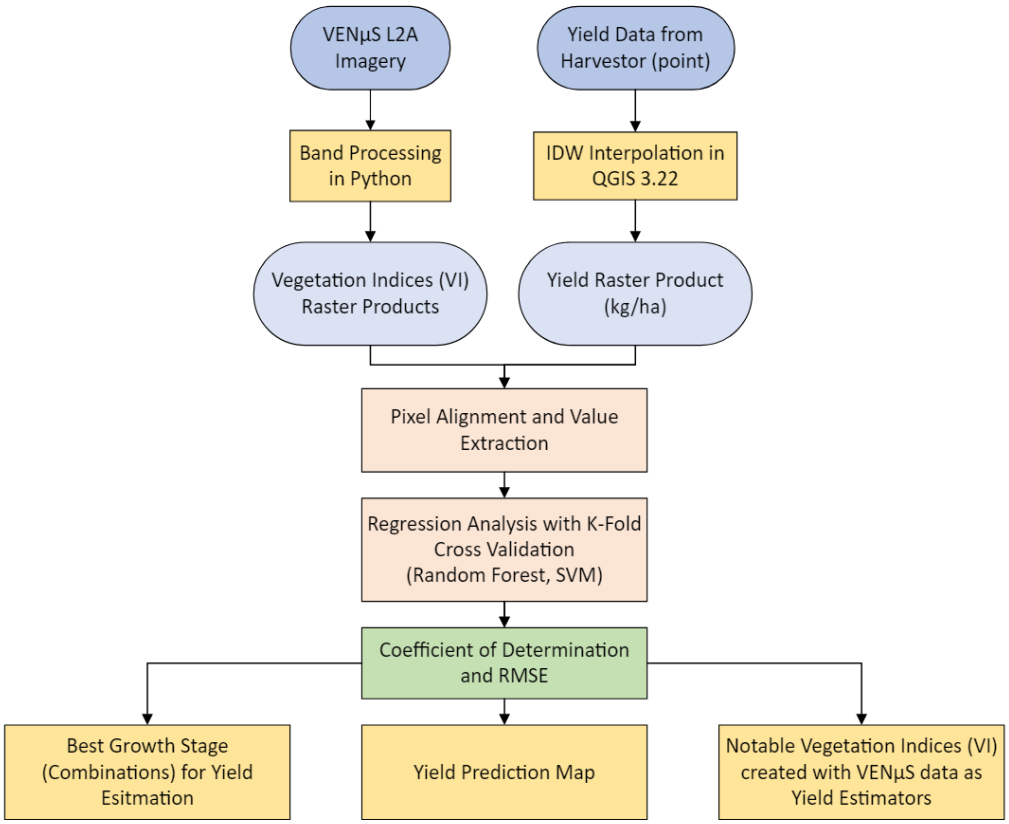


Figure 3-2. Methodology flowchart of this study.

3.3 Results

3.3.1 Cross-validation of Regression Models

In our study, we performed 10-fold cross-validation on a total of 13 datasets. These datasets were divided into two categories: 8 individual growth stages, and three groups of growth stages. The growth stage groups included 3 combinations of growth stages from pre-heading, 2 combinations of growth stages from post-heading, and all data. The rationale for forming grouped pre- and post-heading stage datasets is based on the fact that winter wheat undergoes a transition period marked by a slowdown in leaf growth due to shifts in developmental priorities and physiological changes. As the plant transitions to the reproductive phase, its focus shifts from vegetative to reproductive growth, including the formation and maturation of the inflorescence, causing the leaves to turn yellow. For the purpose of this study, ripening stage data was not included in the post-heading stage

and all data group because the model performance significantly dropped after early fruit development stage.

The mean of the evaluation metrics was used to test the models' generalizability on unseen data across all 10 folds. Among the individual growth stages, the early fruit development stage performed the best, while tillering-1 performed the worst, with both machine learning models performing similarly. As seen in figure 3-3, there was a trend in increasing $\overline{R^2}$, and decreasing $\overline{RMSE}$ and $\overline{MAE}$ as the growth stages progressed from tillering-1 to early fruit development. The evaluation metrics displayed a significant drop of model performance for both RF and SVR afterwards in the ripening stage. The RF model ($\overline{R^2}$ = 0.78, $\overline{RMSE}$ = 0.4832 t/ha, $\overline{MAE}$ = 0.3362 t/ha) explained the variance slightly better than the SVR model ($\overline{R^2}$ = 0.78, $\overline{RMSE}$ = 0.4834 t/ha, $\overline{MAE}$ = 0.3330 t/ha). However, the SVR model was slightly less sensitive to outliers compared to the RF model. Overall, the machine learning regression models within each growth stage were similar in terms of stability, as seen in the sizes of the whiskers on figure 3-3.

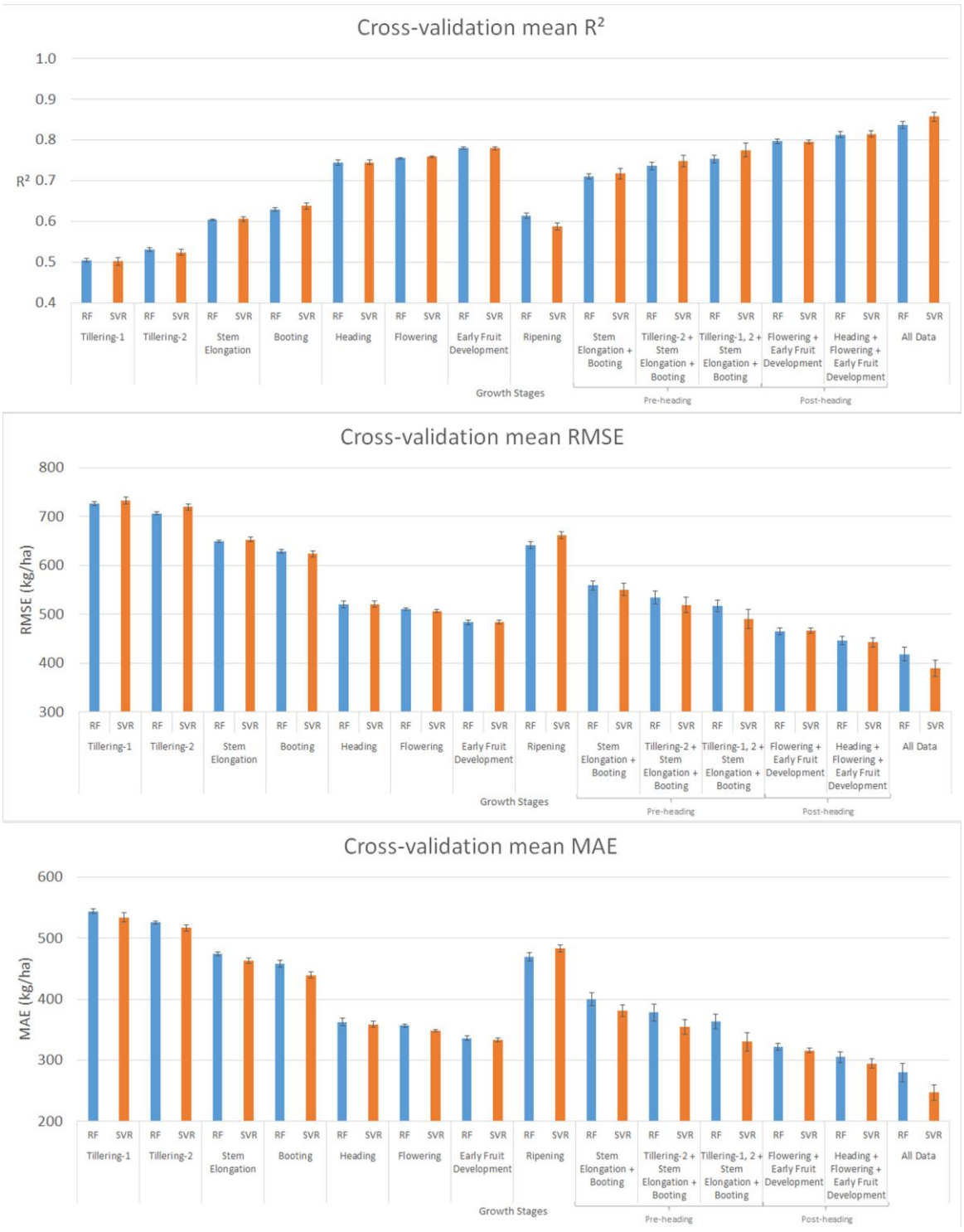


Figure 3-3. Mean cross-validation statistics histogram: analysis by growth stages datasets and modelling approach (RF and SVR) using 21 VI variables. The whiskers display the standard deviation of the metrics.

On the other hand, models that incorporated all data and combinations from the pre- and post-heading groups showed an increase in explanatory power as more data were added to the regression models. Models using post-heading stage datasets demonstrated greater robustness, with lower mean RMSE values, compared to those using pre-heading stage datasets. However, models using all data were the least stable, as indicated by their higher $\overline{\text{RMSE}}$ standard deviation. Overall, both models demonstrated the highest generalizability on unseen data when all data were combined. The SVR ($\overline{R^2}$ = 0.86, $\overline{\text{RMSE}}$ = 0.3899 t/ha, $\overline{\text{MAE}}$ = 0.2475 t/ha) explained data variance better and had lower prediction error than the RF ($\overline{R^2}$ = 0.84, $\overline{\text{RMSE}}$ = 0.4185 t/ha, $\overline{\text{MAE}}$ = 0.2800 t/ha) model in cross-validation.

3.3.2 Yield Prediction Using Regression Models

Table 3-4 displays the calibration and validation performance of the RF and SVR models with datasets from all 8 tested growth stages individually. Overall, the models best at explaining data variance could be found when the models were using early fruit development stage data. The calibration R^2 ranged between 0.54 and 0.96 and RMSE values ranged between 0.2039 and 0.7057 t/ha. The RF model has a significantly higher calibration R^2 = 0.96 compared to the SVR model R^2 = 0.79, indicating that the RF model fits the training data much better. This finding was consistent throughout the analysis, and is expected due to RF's ensemble nature, which excels in capturing complex patterns. In terms of validation model metrics, R^2 value ranged between 0.50 and 0.77 and RMSE value ranged between 0.5008 and 0.7421 t/ha. The validation results were also consistent with calibration that both RF and SVR had the highest R^2 when paired with data from the early fruit development stage. RF (R^2 = 0.77, RMSE = 0.5008 t/ha) slightly outperformed SVR (R^2 = 0.77, RMSE = 0.5039 t/ha), making it the best performing pre-diction model when using data from individual stages.

Table 3-4. Calibration and validation statistics: analysis by individual growth stage datasets and modelling approach (RF and SVR) using 21 VI variables[1].

Growth Stage	Model	Calibration		Validation	
		R^2	RMSE (t/ha)	R^2	RMSE (t/ha)
Tillering-1	RF	0.94	0.3017	0.50	0.7335

	SVR	0.54	0.7057	0.50	0.7421
Tillering-2	RF	0.94	0.2953	0.53	0.7116
	SVR	0.55	0.6971	0.53	0.7208
Stem Elongation	RF	0.94	0.2727	0.61	0.6510
	SVR	0.63	0.6358	0.61	0.6539
Booting	RF	0.95	0.2607	0.64	0.6264
	SVR	0.66	0.6032	0.65	0.6177
Heading	RF	0.96	0.2186	0.74	0.5283
	SVR	0.77	0.4989	0.74	0.5319
Flowering	RF	0.96	0.2181	0.75	0.5254
	SVR	0.77	0.4907	0.75	0.5183
Early Fruit Development	RF	0.96	0.2039	0.77	0.5008
	SVR	0.79	0.4696	0.77	0.5039
Ripening	RF	0.95	0.2653	0.61	0.6494
	SVR	0.61	0.6418	0.59	0.6709

[1] All models are significant at p-value < 0.001.

The analysis extended to using combinations of datasets from growth stages. Table 3-5 displays the calibration and validation performance of the RF and SVR models with dataset groups of pre-heading stage, post-heading stage, and all data. Collectively, models using dataset groups as variables outperformed models using individual datasets from tilling-1, tilling-2, stem elongation, and booting stage, which all of them were in the pre-heading stage group. In calibration, the R^2 values ranged between 0.75 and 0.98, while the RMSE value ranged between 0.1640 and 0.5189 t/ha. Both ranges were significantly narrower compared to their counterparts in individual growth stages as we saw improved performance of SVR. Overall, the calibration R^2 was consistently higher in RF models than in SVR models. However, the validation statistics shown that SVR models outperformed RF models in yield prediction as the validation, with higher R^2 values and the lower RMSE values in SVR models when paired with each of the three dataset groups. The R^2 values ranged between 0.72 and 0.86, while the RMSE value ranged between 0.3925 and 0.5465 t/ha in validation. The best yield prediction model was found to be SVR model using all data from tillering-1 to early fruit development stage (R^2 = 0.86, RMSE = 0.3925 t/ha). Although the RF model performed better in calibration, it predicted yield with slightly lower accuracy (R^2 = 0.83, RMSE = 0.4257 t/ha).

Table 3-5. Calibration and validation statistics: analysis by dataset groups and modelling approach (RF and SVR) using 21 VI variables[1].

Dataset Group	Growth Stage Combinations	Model	Calibration		Validation	
			R^2	RMSE (t/ha)	R^2	RMSE (t/ha)
Pre-heading Stage	Stem Elongation + Booting	RF	0.96	0.2241	0.72	0.5561
		SVR	0.75	0.5189	0.73	0.5465
	Tillering-2 + Stem Elongation + Booting	RF	0.97	0.2119	0.74	0.5353
		SVR	0.79	0.4788	0.75	0.5165
	Tillering-1, 2 + Stem Elongation + Booting	RF	0.97	0.2038	0.75	0.5210
		SVR	0.82	0.4431	0.78	0.4917
Post-heading Stage	Flowering + Early Fruit Development	RF	0.97	0.1902	0.79	0.4810
		SVR	0.81	0.4462	0.79	0.4814
	Heading + Flowering + Early Fruit Development	RF	0.97	0.1798	0.81	0.4570
		SVR	0.84	0.4116	0.81	0.4507
All Data (Ripening excluded)		RF	0.98	0.1640	0.83	0.4257
		SVR	0.89	0.3437	0.86	0.3925

[1] All models are significant at p-value < 0.001.

3.3.3 Ranked Importance of Vegetation Indices from Different Growth Stages

RF modeling, which utilizes numerous decision trees, was employed to generate a variable importance plot in RStudio using the "varImpPlot()" function. This plot displays increasing node purity (IncNodePurity) on the x-axis, representing the importance of each variable in predicting yield across different dates. A higher IncNodePurity value indicates that the variable is more significant as a predictor, helping to identify the key predictors in the models. The variable importance plot revealed that NDRE-1 and NDRE-2 from heading, flowering, and early fruit development stages were among the most important predictors of yield, with the top-ranked variable being NDRE-1 from flowering stage, as shown in figure 3-4. REP from flowering and early fruit development stage ranked 4th and 7th respectively, on the plot. NDRE-1, NDRE-2, REP, and ARVI from multiple growth stages constituted 17 of the top 20 ranked VI variables. Beyond the top 7 ranked VIs, the IncNodePurity values of the remaining VI variables were relatively similar and gradually decreased throughout the list of the 147 tested VI variables in total.

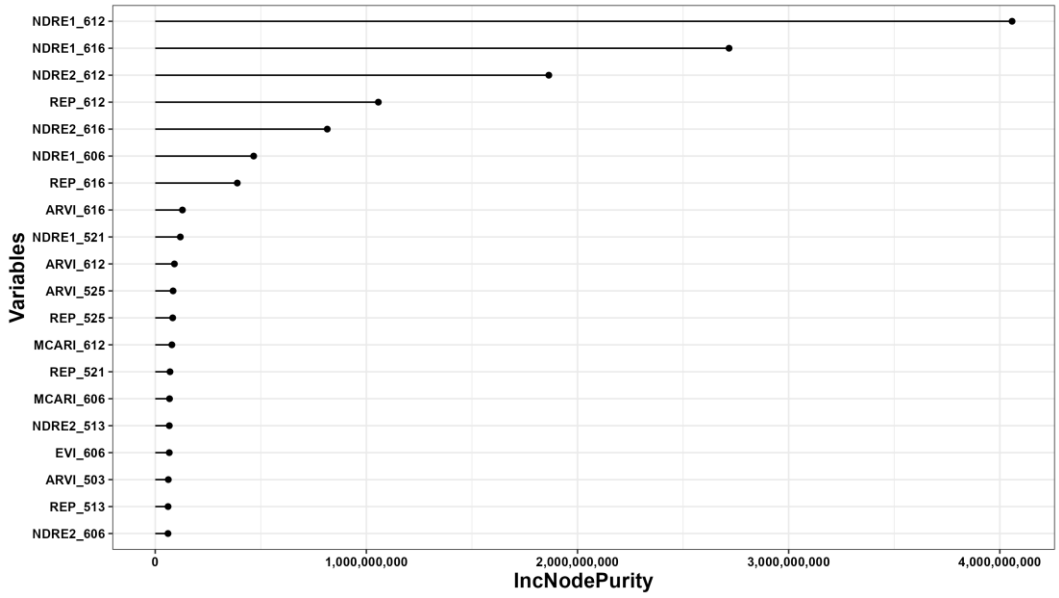


Figure 3-4. Variable importance plot produced with VIs with all data. Only the top 20 of the 147 VI variables were displayed. Refer to table 3-3 for the full names of the variables. The number denoted after the variables' abbreviation is the date of the VENμS imagery.

3.3.4 Visualization of Predicted Yield

A yield prediction map helps visualize the yield variations within a field, and VENμS' higher spatial resolution enabled readers to clearly identify areas of inaccuracies. Figure 3-5 demonstrates that the prediction generally captured the yield variations across the entire field, as reflected by both the map and the evaluation metrics. However, the prediction did not accurately capture the extreme values in the observed yield. For instance, the north side borders of the field showed extreme lows in the observed yield, but the predicted yield map did not reflect values as extreme. Similarly, in areas of extreme highs, the prediction failed to capture the highest values and some-times incorrectly predicted the yield as a markedly different value.

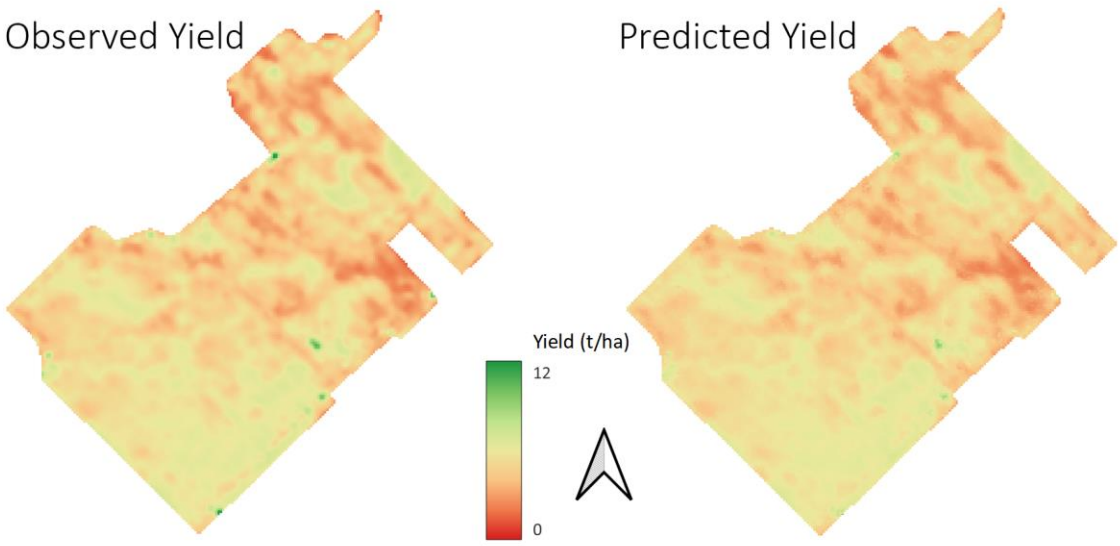


Figure 3-5. Visualized comparison between the observed and predicted yield.

3.4 Discussion

3.4.1 Implications of Model Performance on Yield Prediction with VENμS Imagery

PA uses advanced technologies and data analysis to optimize agricultural practices and assist in management decisions, with the goal of minimizing input while maximizing output and efficiency. We proposed using VENμS imagery for yield prediction as an alternative to other publicly available satellite data with its higher spatial and temporal resolution. The differences in performance of the machine learning regression models have been discussed in detail above. Though similar in prediction performance. SVR was the overall better machine learning regression model when more data was added to the regression, while RF was more accurate when predicting yield with data from individual stages. Our findings aligned with the study conducted by Han et al. and we were able to suggest that both RF and SVR were high performance techniques in yield prediction (Han et al., 2020; Nigam et al., 2019). However, potential overfitting was observed in the RF models, even after careful tuning. Although this overfitting decreased as more growth stage data was added, we recognize that RF models can be prone to overfitting, especially when dealing with complex data. This limitation prompted the inclusion of SVR in the

study, which exhibited less overfitting and proved to be the more reliable algorithm in this context.

The regression models were able to distinguish early fruit development stage was the best growth stage to predict yield from. This finding agreed with Hassan et al. in which the yield prediction accuracy increased as the growth stages progressed (Hassan et al., 2019). Overall, our study was able to achieve a higher accuracy by incorporating all data from tillering to early fruit development stage. Among all the test VIs, NDRE is the most important predictor of yield as tested, which is consistent with previous studies (Fu et al., 2020). As reported, NDRE-1, NDRE-2, REP, and ARVI from multiple growth stages made up 17 of the top 20 most important variables in predicting the final yield, with NDRE-1, NDRE-2, and REP being the major contributor to the prediction. The common characteristics of the three VIs is that they all used bands 8, 9, 10 of VENμS, which are at the central wavelengths of 702, 741.1, and 782.2 nm. These bands fall in between the red-edge and NIR regions of the spectrum and the VIs based on this wavelength range was previously proven to be effective in predicting grain yield (Zhang et al., 2018).

Our results, when compared to studies based on different satellite data, displayed a similar or even higher prediction accuracy, often accompanied by lower errors (Skakun et al., 2018; Zhao et al., 2020). Given that grouped stages of data performed better in the prediction models, we conducted an additional test with Sentinel-2 data, applying the same methodology and using overpass dates as close as possible to those used with VENμS data. Table 3-6 presents the regression statistics of the yield prediction models using Sentinel-2 data with grouped growth stages. Of the 21 VIs tested, 20 were recreated using Sentinel-2 data. Unfortunately, data from the Tillering-2 stage could not be included in the analysis due to cloud cover. The best prediction results obtained was at R^2 = 0.79 and RMSE = 0.5147 t/ha with SVR using all data.

Table 3-6. Calibration and validation statistics: analysis by dataset groups and modelling approach (RF and SVR) using 20 VI variables created using Sentinel-2 bands, matched to equivalent VENμS bands[1].

Dataset Group	Growth Stage Combinations	Model	Calibration		Validation	
			R^2	RMSE (t/ha)	R^2	RMSE (t/ha)
Pre-heading Stage	Stem Elongation + Booting	RF	0.96	0.2557	0.70	0.6280
		SVR	0.76	0.5452	0.72	0.5997
	Tillering-1 + Stem Elongation + Booting	RF	0.96	0.2476	0.71	0.6106
		SVR	0.78	0.5237	0.74	0.5835
Post-heading Stage	Flowering + Early Fruit Development	RF	0.97	0.2167	0.78	0.5379
		SVR	0.82	0.4728	0.79	0.5238
	Heading + Flowering + Early Fruit Development	RF	0.97	0.2121	0.78	0.5310
		SVR	0.83	0.4615	0.79	0.5190
All Data (Ripening excluded)		RF	0.97	0.2091	0.78	0.5287
		SVR	0.84	0.4434	0.79	0.5147

[1] All models are significant at p-value < 0.001.

The accuracy of the prediction model plateaued when using post-heading stage data only, with minor decrease in RMSE as more data was added. Additionally, compared to the best prediction model using VENμS imagery, the model prediction error with Sentinel-2 data was still significantly higher. This could be contributed to Senti-nel-2's lower spatial resolution, as VENμS has four times more data than Sentinel-2. Although we were able to optimize and successfully create a robust and accurate winter wheat yield prediction model with VENμS data at a local, field-scale, it is not without its drawbacks. Contrast to most publicly available satellites which provide frequent coverage of the Earth's land surface, VENμS does not cover the entire Earth's land surface. Instead, it focuses on specific sites of interest and revisits these selected sites frequently. This means that the site of researcher's interest may not be in coverage even though the data is publicly available. Researchers are required to apply for VENμS coverage at their location of interest.

3.5 Conclusions

This study evaluated the effectiveness of VENμS multispectral imagery in predicting winter wheat yield in southwestern Ontario using machine learning methods. A total of 21 VIs, including 8 variations of existing VIs based on their original development formulas, were tested. The best prediction result demonstrated a high correlation between VENμS data and observed yield, with an $R^2 = 0.86$ and an RMSE = 0.3925 t/ha using an SVR model. According to our results, a reliable prediction of yield can be achieved two months prior to harvest using the combined pre-heading stage data, and the best result can be obtained 39 days prior when using all data from pre- and post- heading stage. The findings suggest that VENμS data can offer superior yield prediction accuracy compared to other publicly available satellites and could potentially serve as a viable alternative to UAV data for local, field-scale studies.

Though machine learning algorithms are effective in capturing complex patterns among variables, it is important to recognize the empirical nature of these models. They rely on existing datasets for validation and can only approximate the observed yield, which is only verifiable at harvest. This intrinsic limitation highlights the potential discrepancies between predicted and actual outcomes. Such limitations underscore the necessity for ongoing calibration and testing of these models under varied agricultural conditions and across different crop cycles to ensure their reliability and accuracy.

Additionally, while k-fold cross-validation and a 70/30 train-test split were employed in this study due to its broad adoption and effective use of the data, future work could explore spatial splitting as a viable alternative. Spatial splitting, which divides the dataset based on geographic location rather than random subsets, may provide a more realistic evaluation of the model’s robustness across different parts of the field by better addressing spatial autocorrelation. Investigating this approach could enhance the model's performance in capturing spatial variability within the field.

VENμS, as mentioned above, does not provide worldwide coverage, which is a significant drawback limiting the use of its superior high-resolution multispectral data. Although this study showed that Sentinel-2 is a less effective alternative, it remains the

next best option for predicting yield with publicly available satellite data using our method. Its worldwide frequent coverage can produce comparable results to VENμS at the field scale and potentially similar results at a regional scale. This research highlights the potential of high-resolution satellite data with multispectral cameras for yield prediction. Future studies may also consider using commercial satellites such as PlanetScope and WorldView-3 as an alternative for high-resolution multispectral data. Combining these satellite data with yield estimation models could lead to advancements in low labor cost, non-destructive, yet highly accurate yield predictions, providing a more detailed understanding of crop yield potential and distribution.

3.6 References

Agriculture and Agri-Food Canada. (2024, June 27). *Overview of Canada's agriculture and agri-food sector*. https://agriculture.canada.ca/en/sector/overview

Atkinson Amorim, J. G., Schreiber, L. V., de Souza, M. R. Q., Negreiros, M., Susin, A., Bredemeier, C., Trentin, C., Vian, A. L., de Oliveira Andrades-Filho, C., Doering, D., & Parraga, A. (2022). Biomass estimation of spring wheat with machine learning methods using UAV-based multispectral imaging. *International Journal of Remote Sensing*, *43*(13), 4758–4773. https://doi.org/10.1080/01431161.2022.2107882

Bukowiecki, J., Rose, T., & Kage, H. (2021). Sentinel-2 Data for Precision Agriculture?—A UAV-Based Assessment. *Sensors*, *21*(8), Article 8. https://doi.org/10.3390/s21082861

Cao, Q., Miao, Y., Shen, J., Yu, W., Yuan, F., Cheng, S., Huang, S., Wang, H., Yang, W., & Liu, F. (2016). Improving in-season estimation of rice yield potential and responsiveness to topdressing nitrogen application with Crop Circle active crop canopy sensor. *Precision Agriculture*, *17*(2), 136–154. https://doi.org/10.1007/s11119-015-9412-y

Centre National d'Etudes Spatiales (CNES), & Israeli Space Agency (ISA). (2023). *The VENμS mission and products* (pp. 1–26). CNES and ISA. https://venus.bgu.ac.il/Links/VENuS_mission_summary_VM05_v02.pdf

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27. https://doi.org/10.1145/1961189.1961199

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61–69. https://doi.org/10.1016/j.compag.2018.05.012

Daughtry, C. S. T., Walthall, C. L., Kim, M. S., de Colstoun, E. B., & McMurtrey, J. E. (2000). Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. *Remote Sensing of Environment*, *74*(2), 229–239. https://doi.org/10.1016/S0034-4257(00)00113-9

European Space Agency (ESA). (2015). *Sentinel-2 User Handbook* (pp. 1–64). ESA. https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook

Fernandes, R., Butson, C., Leblanc, S., & Latifovic, R. (2003). Landsat-5 TM and Landsat-7 ETM+ based accuracy assessment of leaf area index products for Canada derived from SPOT-4 VEGETATION data. *Canadian Journal of Remote Sensing*, *29*(2), 241–258. https://doi.org/10.5589/m02-092

Fu, Y., Yang, G., Wang, J., Song, X., & Feng, H. (2014). Winter wheat biomass estimation based on spectral indices, band depth analysis and partial least squares regression using

hyperspectral measurements. *Computers and Electronics in Agriculture*, *100*, 51–59. https://doi.org/10.1016/j.compag.2013.10.010

Fu, Z., Jiang, J., Gao, Y., Krienke, B., Wang, M., Zhong, K., Cao, Q., Tian, Y., Zhu, Y., Cao, W., & Liu, X. (2020). Wheat Growth Monitoring and Yield Estimation based on Multi-Rotor Unmanned Aerial Vehicle. *Remote Sensing*, *12*(3), Article 3. https://doi.org/10.3390/rs12030508

Gitelson, A., & Merzlyak, M. N. (1994). Quantitative estimation of chlorophyll-*a* using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology B: Biology*, *22*(3), 247–252. https://doi.org/10.1016/1011-1344(93)06963-4

Guyot, G., & Baret, F. (1988). Utilisation de la Haute Resolution Spectrale pour Suivre L'etat des Couverts Vegetaux. *Spectral Signatures of Objects in Remote Sensing*, *287*, 279–286.

Hammer, G. (2000). Applying Seasonal Climate Forecasts in Agricultural and Natural Ecosystems—A Synthesis. In G. L. Hammer, N. Nicholls, & C. Mitchell (Eds.), *Applications of Seasonal Climate Forecasting in Agricultural and Natural Ecosystems* (pp. 453–462). Springer Netherlands. https://doi.org/10.1007/978-94-015-9351-9_27

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., & Zhang, J. (2020). Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China. *Remote Sensing*, *12*(2), Article 2. https://doi.org/10.3390/rs12020236

Hassan, M. A., Yang, M., Rasheed, A., Yang, G., Reynolds, M., Xia, X., Xiao, Y., & He, Z. (2019). A rapid monitoring of NDVI across the wheat growth cycle for grain yield prediction using a multi-spectral UAV platform. *Plant Science*, *282*, 95–103. https://doi.org/10.1016/j.plantsci.2018.10.022

Hewer, M. J., & Brunette, M. (2020). Climate change impact assessment on grape and wine for Ontario, Canada's appellations of origin. *Regional Environmental Change*, *20*(3), 86. https://doi.org/10.1007/s10113-020-01673-y

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1), 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, *25*(3), 295–309. https://doi.org/10.1016/0034-4257(88)90106-X

Hunt, M. L., Blackburn, G. A., Carrasco, L., Redhead, J. W., & Rowland, C. S. (2019). High resolution wheat yield mapping using Sentinel-2. *Remote Sensing of Environment*, *233*(Complete). https://doi.org/10.1016/j.rse.2019.111410

Jordan, C. F. (1969). Derivation of Leaf-Area Index from Quality of Light on the Forest Floor. *Ecology*, *50*(4), 663–666. https://doi.org/10.2307/1936256

Kaufman, Y. J., & Tanre, D. (1992). Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. *IEEE Transactions on Geoscience and Remote Sensing*, *30*(2), 261–270. IEEE Transactions on Geoscience and Remote Sensing. https://doi.org/10.1109/36.134076

Liao, C., Wang, J., Shan, B., Shang, J., Dong, T., & He, Y. (2023). Near real-time detection and forecasting of within-field phenology of winter wheat and corn using Sentinel-2 time-series data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *196*, 105–119. https://doi.org/10.1016/j.isprsjprs.2022.12.025

Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019). Crop Yield Prediction Using Machine Learning Algorithms. *2019 Fifth International Conference on Image Information Processing (ICIIP)*, 125–130. https://doi.org/10.1109/ICIIP47207.2019.8985951

Ontario Ministry of Agriculture, Food and Rural Affairs. (2023, November 8). *Census farm data collection*. Ontario Data Catalogue. https://data.ontario.ca/dataset/census-farm-data-collection

Panek, E., Gozdowski, D., Stępień, M., Samborski, S., Ruciński, D., & Buszke, B. (2020). Within-Field Relationships between Satellite-Derived Vegetation Indices, Grain Yield and Spike Number of Winter Wheat and Triticale. *Agronomy*, *10*(11), Article 11. https://doi.org/10.3390/agronomy10111842

Qi, J., Chehbouni, A., Huete, A. R., Kerr, Y. H., & Sorooshian, S. (1994). A modified soil adjusted vegetation index. *Remote Sensing of Environment*, *48*(2), 119–126. https://doi.org/10.1016/0034-4257(94)90134-1

Reid, S., Smit, B., Caldwell, W., & Belliveau, S. (2007). Vulnerability and adaptation to climate risks in Ontario agriculture. *Mitigation and Adaptation Strategies for Global Change*, *12*(4), 609–637. https://doi.org/10.1007/s11027-006-9051-8

Richardson, A. J., & Wiegand, C. L. (1977). Distinguishing vegetation from soil background information. *Photogrammetric Engineering and Remote Sensing*, *43*, 1541–1552.

Rondeaux, G., Steven, M., & Baret, F. (1996). Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment*, *55*(2), 95–107. https://doi.org/10.1016/0034-4257(95)00186-7

Roujean, J.-L., & Breon, F.-M. (1995). Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. *Remote Sensing of Environment*, *51*(3), 375–384. https://doi.org/10.1016/0034-4257(94)00114-3

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. *NASA Special Publications*, *351*(1), 309.

Shafi, U., Mumtaz, R., García-Nieto, J., Hassan, S. A., Zaidi, S. A. R., & Iqbal, N. (2019). Precision Agriculture Techniques and Practices: From Considerations to Applications. *Sensors*, *19*(17), Article 17. https://doi.org/10.3390/s19173796

Silleos, N. G., Alexandridis, T. K., Gitas, I. Z., & Perakis, K. (2006). Vegetation Indices: Advances Made in Biomass Estimation and Vegetation Monitoring in the Last 30 Years. *Geocarto International*, *21*(4), 21–28. https://doi.org/10.1080/10106040608542399

Skakun, S., Franch, B., Vermote, E., Roger, J.-C., Justice, C., Masek, J., & Murphy, E. (2018). Winter Wheat Yield Assessment Using Landsat 8 and Sentinel-2 Data. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 5964–5967. https://doi.org/10.1109/IGARSS.2018.8519134

Transport Canada. (2023, March 16). *Flying your drone safely and legally*. AARV. https://tc.canada.ca/en/aviation/drone-safety/learn-rules-you-fly-your-drone/flying-your-drone-safely-legally

United States Geological Survey (USGS). (2019). *Landsat 8 (L8) Data Users Handbook* (pp. 1–93). USGS. https://d9-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/LSDS-1574_L8_Data_Users_Handbook-v5.0.pdf

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

Wang, F., Yang, M., Ma, L., Zhang, T., Qin, W., Li, W., Zhang, Y., Sun, Z., Wang, Z., Li, F., & Yu, K. (2022). Estimation of Above-Ground Biomass of Winter Wheat Based on Consumer-Grade Multi-Spectral UAV. *Remote Sensing*, *14*(5), Article 5. https://doi.org/10.3390/rs14051251

Wu, C., Niu, Z., Tang, Q., & Huang, W. (2008). Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. *Agricultural and Forest Meteorology*, *148*(8), 1230–1241. https://doi.org/10.1016/j.agrformet.2008.03.005

Xie, Q., Huang, W., Liang, D., Chen, P., Wu, C., Yang, G., Zhang, J., Huang, L., & Zhang, D. (2014). Leaf Area Index Estimation Using Vegetation Indices Derived From Airborne Hyperspectral Images in Winter Wheat. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(8), 3586–3594. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. https://doi.org/10.1109/JSTARS.2014.2342291

Yu, J., Wang, J., & Leblon, B. (2021). Evaluation of Soil Properties, Topographic Metrics, Plant Height, and Unmanned Aerial Vehicle Multispectral Imagery Using Machine Learning Methods to Estimate Canopy Nitrogen Weight in Corn. *Remote Sensing*, *13*(16), 3105–3105. https://doi.org/10.3390/rs13163105

Yu, J., Wang, J., Leblon, B., & Song, Y. (2022). Nitrogen Estimation for Wheat Using UAV-Based and Satellite Multispectral Imagery, Topographic Metrics, Leaf Area Index, Plant Height, Soil Moisture, and Machine Learning Methods. *Nitrogen*, *3*(1), Article 1. https://doi.org/10.3390/nitrogen3010001

Zhang, C., Marzougui, A., & Sankaran, S. (2020). High-resolution satellite imagery applications in crop phenotyping: An overview. *Computers and Electronics in Agriculture*, *175*, 105584. https://doi.org/10.1016/j.compag.2020.105584

Zhang, Y., Qin, Q., Ren, H., Sun, Y., Li, M., Zhang, T., & Ren, S. (2018). Optimal Hyperspectral Characteristics Determination for Winter Wheat Yield Prediction. *Remote Sensing*, *10*(12), Article 12. https://doi.org/10.3390/rs10122015

Zhao, Y., Potgieter, A. B., Zhang, M., Wu, B., & Hammer, G. L. (2020). Predicting Wheat Yield at the Field Scale by Combining High-Resolution Sentinel-2 Satellite Imagery and Crop Modelling. *Remote Sensing*, *12*(6), Article 6. https://doi.org/10.3390/rs12061024

Chapter 4

4 Conclusion

This chapter summarizes the thesis, discussing the research outcomes and evaluating whether the objectives were met. It also outlines the limitations of this research and explores possibilities for future studies.

4.1 Summary

Remote sensing applications are pivotal in precision agriculture as they facilitate efficient agricultural practices through the power of data, science, and technology. The continuous development and research of remote sensing applications for detecting plant characteristics and predicting crop productivity aim to help farmers anticipate and respond to future conditions. With the increase in frequency of extreme climate events causing environmental damage and threatening food security, the agricultural industry must strategically manage and apply its resources to maximize harvest output and avoid crop failure.

In Chapter 2, RF and SVM regression methods were used to predict the AGB of winter wheat, utilizing UAV multispectral MicaSense bands, associated VIs, plant biophysical parameters (plant height and LAI), and plant biochemical parameters (nutrient content levels and ratios). Both single-date and multi-date data were tested, and the best model's variable importance plot was used to identify key variables related to winter wheat AGB by variable selection based on the importance ranking. The result allowed evaluation of whether biochemical parameters were effective additions to the prediction models, as they were not commonly used in a biomass estimation study.

In Chapter 3, RF and SVM regression methods were used to predict the yield of winter wheat, utilizing VIs derived from VENµS satellite imagery, which offers higher spatial and temporal resolution compared to other popular satellites. The data were analyzed based on each growth stage of winter wheat by categorizing them individually and in combinations of growth stages. The variable importance plot of the best performing

model highlighted the key VIs at different growth stages that were strongly related to winter wheat yield. This analysis identified the most effective VENμS satellite bands and growth stages for yield prediction. A yield prediction map was then created and compared to the observed yield to visualize the prediction model's effectiveness in capturing field variations.

4.2 Conclusions

The research objectives for this thesis were completed in two separate journal articles detailed in Chapters 2 and 3, here are the responses to each as follows:

i. In Chapter 2, machine learning RF and SVR models were adopted to estimate winter wheat AGB using UAV multispectral data, plant biophysical, and plant biochemical parameters. The testing was first conducted across single- and multi-date datasets using all of the available variables. The accuracy of the machine learning models generally increased as more dates of datasets were used, and he best performing model was the RF model, which achieved an $R^2 = 0.80$ and RMSE = 152.71 g/m^2 using all four dates of data. The SVR models consistently showed slightly lower performance across different datasets, with the best performing SVR model achieving an $R^2 = 0.77$ and RMSE = 156.61 g/m^2 using the last three of the four dates of data. Afterwards, the variable importance plots for each of the best performing machine learning regression models were then utilized for variable selection based on their ranking to reduce redundancy and complexity in the regression models.

ii. In Chapter 2, variable selection was performed after determining the best date (growth stage) combinations for the RF and SVR models. The overall best-performing model was RF, which employed a combination of the top 20 variables. This model achieved an $R^2 = 0.81$ and an RMSE of 149.95 g/m^2, generally outperforming the SVR models. This was accomplished using data from all four dates (growth stages), ranging from the heading stage to the ripening stage. Among the 20 variables, there was a close to even split between spectral variables and nutrient content variables.

The variable importance plots for both the best-performing RF and SVR models showed that NDVI, ISR, ARVI, and RVI were consistently among the top four most important variables in relation to AGB. In the RF model, macronutrients such as N and K were ranked 6^{th} and 9^{th}, respectively, while in the SVR model, K and P ranked 5^{th} and 6^{th}, respectively.

iii. In Chapter 3, machine learning RF and SVR models were adopted to predict winter wheat yield using VIs derived from VENμS satellite imagery. The testing was conducted across the growth stages of the plants, from when they first become large enough to be detected to maturity. The best performing model was the SVR model, which achieved an $R^2 = 0.86$ and RMSE = 0.3925 t/ha, outperforming the best performing RF model, which achieved an $R^2 = 0.83$ and RMSE = 0.4257 t/ha. The variable importance plot for the best performing model was further studied to understand the effectiveness of VIs at different growth stages contributing to the prediction.

iv. In Chapter 3, the best performing prediction model was RF utilizing the "all data" dataset, which included satellite imagery captured from the winter wheat's tillering stage to early fruit development stage. This indicates that the early fruit development stage is the optimal time to use VENμS satellite imagery for yield prediction. Generally, fairly accurate predictions can be produced using combinations of datasets from the post-heading stage. The results also show that VENμS is a slightly superior alternative to the publicly available satellite data from Sentinel-2, as detailed in the conclusion section of Chapter 3.

The variable importance plot of the best performing model revealed that NDRE-1, NDRE-2, REP, and ARVI from multiple growth stages constituted 17 of the top 20 ranked VI variables. The common characteristics of NDRE-1, NDRE-2, and REP, the major contributors to the prediction, are that they all utilized bands 8, 9, and 10 of VENμS, which are all in the red-edge and NIR regions of the spectrum. Using the prediction model, a yield prediction map was created and compared with the observed yield map to evaluate the spatial variations of yield and

visualize the ability of the prediction model to capture extreme values in the observed yield.

4.3 Significance of the Research

In Chapter 2, the novelty of the research lay in the ability to use plant nutrient content levels and ratios as predictors of winter wheat AGB, demonstrating them to be among the most important predictors in the study. In the current research field, these parameters are not widely used, potentially due to the lack of access to plant laboratory analyses. Previous studies have also shown that senescence can cause a decline in model performance when predicting biomass (Sharma et al., 2022). Our study, utilizing a lower-cost UAV multispectral camera setup combined with biophysical and biochemical parameters, particularly at the later growth stages of winter wheat post-flowering, demonstrates a cost-effective method to predict AGB.

In Chapter 3, the effectiveness of VENμS satellite imagery in predicting winter wheat yield at a local, field scale was tested. Although the use of satellite imagery and machine learning for yield predictions has been widely studied, accurate yield prediction models have often been developed only on a regional scale due to challenges such as limited clear weather coverage and low spatial resolution (Hunt et al., 2019; Ma et al., 2022). The success of producing an accurate yield prediction model with VENμS imagery at a local, field scale highlights the potential of using high-resolution satellite imagery as an alternative to UAV-based imagery. This approach benefits farmers or researchers who lack the equipment to collect high-resolution multispectral data.

4.4 Limitations and Future Work

The limitations of the study varied between the approaches as Chapters 2 and 3 focused on different aspects of predicting crop productivity. However, both encountered the same fundamental limitation: the reliance of these machine learning regression models on existing datasets for validation. For biomass, it was in-situ destructive sampling, and for yield, it was at harvest. This inherent constraint reveals the possible gaps between predicted and observed results. These constraints emphasize the importance of

continuous calibration and testing of these models in diverse agricultural settings and across various crop cycles to maintain their dependability and precision.

Data access was also a major limitation. Studies similar to the one conducted in Chapter 2 often required in-situ measurements in the field, which are typically associated with intensive labor, costs, and conditions. UAV operations can also be disrupted by weather conditions, making it challenging to collect data in a timely manner when crop growth conditions can change within days. Access to high-quality satellite data can be difficult due to factors like cloud coverage. In Chapter 3, we noted that although VENμS data access is public and free, it does not provide worldwide coverage. As technology advances, future studies may consider using other high-resolution earth observing satellites as an alternative to some in-situ data collection, UAV imagery, and VENμS satellite imagery. Commercial satellites such as PlanetScope and WorldView-3 offer worldwide coverage of high-resolution satellite data, making them viable non-destructive alternatives for obtaining plant properties relevant to biomass and yield as discussed in Chapters 2 and 3.

Random Forest and Support Vector Regression were the two machine learning techniques used in this thesis due to their overall ease of use and commonality in precision agriculture studies. However, adopting machine learning techniques generally bears the limitation of requiring very large training datasets for effective model building, not to mention the associated computational cost of processing the analysis and storing the large amounts of data. With the recent boom in the development of large language models and generative artificial intelligence (AI), future studies may be able to implement these advancements in machine learning, alleviating the computational cost of running complex analyses.

Overall, promising results were discovered for both the aspects of crop productivity of winter wheat by integrating biochemical data in biomass estimation, as well as using VENμS derived VIs for yield prediction. The developed models need to be tested on additional datasets to determine their effectiveness and to better understand their

applicability and feasibility in precision agriculture. This further testing will help validate the models and potentially refine them for broader use in agricultural practices.

4.5 References

Hunt, M. L., Blackburn, G. A., Carrasco, L., Redhead, J. W., & Rowland, C. S. (2019). High resolution wheat yield mapping using Sentinel-2. *Remote Sensing of Environment*, *233*(Complete). https://doi.org/10.1016/j.rse.2019.111410

Ma, C., Liu, M., Ding, F., Li, C., Cui, Y., Chen, W., & Wang, Y. (2022). Wheat growth monitoring and yield estimation based on remote sensing data assimilation into the SAFY crop growth model. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-09535-9

Sharma, P., Leigh, L., Chang, J., Maimaitijiang, M., & Caffé, M. (2022). Above-Ground Biomass Estimation in Oats Using UAV Remote Sensing and Machine Learning. *Sensors*, *22*(2), Article 2. https://doi.org/10.3390/s22020601

5 Appendices

5.1 Appendix A – Fieldwork Photos

Figure A-1. UAV flight mission conducted during the 2020 fieldwork at the wheat field.

Figure A-2. Equipment testing at the field with the LI-COR LAI-2200C

Figure A-3. AGB samples sorted in paper bags at each sample point. Contained in a backpack for ease of transport when walking in the field.

5.2 Appendix B – Data Samples

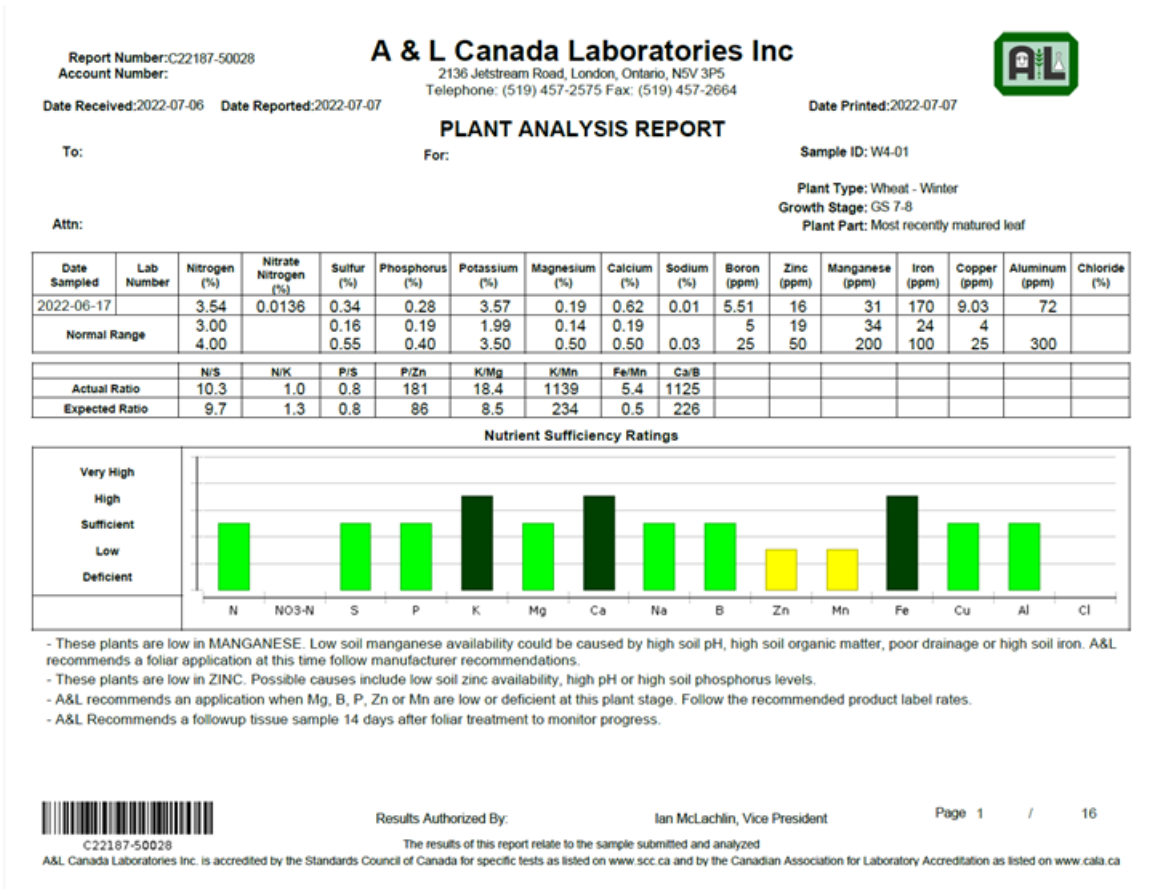
Biomass Lab Datasheet (Wheat)

Recorded by: MARCO Date: JUNE 17 Weight of bag/box (g): 11 g

Sample site	Plant density (plants/m²)	Number of plants harvested	Net fresh weight harvested (g)	Net dry weight harvested (g)	Notes
W4-01	20 cm		94	37	
02			93	25	
03			107	39	
04			104	38	
05			97	35	
06			91	32	
07			102	38	
08			101	36	
09			90	33	
10			81	28	
11			88	32	
12			94	32	
13			102	37	
14			112	39	
15			73	26	
16			86	29	
Notes					

Page:

Figure B-1. Example of biomass lab datasheet for sample points W4-01 to W4-16. Recorded on June 17th, 2022



Report Number:C22187-50028
Account Number:

A & L Canada Laboratories Inc
2136 Jetstream Road, London, Ontario, N5V 3P5
Telephone: (519) 457-2575 Fax: (519) 457-2664

Date Received:2022-07-06 Date Reported:2022-07-07 Date Printed:2022-07-07

PLANT ANALYSIS REPORT

To: For: Sample ID: W4-01

Plant Type: Wheat - Winter
Growth Stage: GS 7-8
Attn: Plant Part: Most recently matured leaf

Date Sampled	Lab Number	Nitrogen (%)	Nitrate Nitrogen (%)	Sulfur (%)	Phosphorus (%)	Potassium (%)	Magnesium (%)	Calcium (%)	Sodium (%)	Boron (ppm)	Zinc (ppm)	Manganese (ppm)	Iron (ppm)	Copper (ppm)	Aluminum (ppm)	Chloride (%)
2022-06-17		3.54	0.0136	0.34	0.28	3.57	0.19	0.62	0.01	5.51	16	31	170	9.03	72	
Normal Range		3.00 4.00		0.16 0.55	0.19 0.40	1.99 3.50	0.14 0.50	0.19 0.50	0.03	5 25	19 50	34 200	24 100	4 25	300	

	N/S	N/K	P/S	P/Zn	K/Mg	K/Mn	Fe/Mn	Ca/B
Actual Ratio	10.3	1.0	0.8	181	18.4	1139	5.4	1125
Expected Ratio	9.7	1.3	0.8	86	8.5	234	0.5	226

Nutrient Sufficiency Ratings

- These plants are low in MANGANESE. Low soil manganese availability could be caused by high soil pH, high soil organic matter, poor drainage or high soil iron. A&L recommends a foliar application at this time follow manufacturer recommendations.
- These plants are low in ZINC. Possible causes include low soil zinc availability, high pH or high soil phosphorus levels.
- A&L recommends an application when Mg, B, P, Zn or Mn are low or deficient at this plant stage. Follow the recommended product label rates.
- A&L Recommends a followup tissue sample 14 days after foliar treatment to monitor progress.

C22187-50028

Results Authorized By: Ian McLachlin, Vice President

Page 1 / 16

The results of this report relate to the sample submitted and analyzed

A&L Canada Laboratories Inc. is accredited by the Standards Council of Canada for specific tests as listed on www.scc.ca and by the Canadian Association for Laboratory Accreditation as listed on www.cala.ca

Figure B-2. Example of plant analysis report for AGB biomass. Samples collected on June 17th, 2022.

```
LAI_FILE        617W42
VERSION 2.0.3
DATE    20220617 10:57:40
PROMPT1 WHERE
RESP1   1
PROMPT2 WHAT
RESP2
TRANSCOMP       APS
MODEL   HORIZONTAL
GPSLAT  42.786433
GPSLONG -81.593161
GPSALT  222.5
GPSHDOP 0.99
GPSNUM  12
LAI     3.608
SEL     0.06062
ACF     0.9934
DIFN    0.07179
MTA     55.40
SEM     3.288
SMP     8
MASK    1       1       1       1       1
ANGLES  7.000   23.00   38.00   53.00   68.00
AVGTRANS        0.1896  0.1394  0.07953 0.03756 0.01705
ACFS    0.9423  0.9894  0.9973  0.9966  0.9958
CNTCT#  1.750   1.832   1.999   1.981   1.531
STDDEV  0.4708  0.1832  0.09170 0.08876 0.06964
DISTS   1.008   1.087   1.270   1.662   2.670
GAPS    0.1713  0.1365  0.07899 0.03714 0.01676

### Contributing Sensors
SENSOR  W1      PCH4836 4152.   1299.   1013.   1000.   1297.


### Observations
A       1       20220617 10:57:47       W1      193.4   193.4   187.1   201.8   252.2
G       2       20220617 10:57:47       G0      42.786437       -81.593155      223.1   8       0.95    20220617 15:20:01
A       3       20220617 10:57:49       W1      201.9   201.8   194.8   208.0   263.1
G       4       20220617 10:57:49       G0      42.786438       -81.593152      222.9   7       1.11    20220617 15:20:03
A       5       20220617 10:57:50       W1      208.2   203.7   193.6   208.1   260.0
G       6       20220617 10:57:50       G0      42.786438       -81.593151      223.0   8       1.11    20220617 15:20:04
A       7       20220617 10:57:52       W1      199.0   199.3   193.0   205.5   270.6
G       8       20220617 10:57:52       G0      42.786438       -81.593154      222.9   8       0.95    20220617 15:20:06
B       9       20220617 10:58:08       W1      25.87   27.38   13.06   7.197   3.344
G       10      20220617 10:58:08       G0      42.786432       -81.593167      222.5   8       0.95    20220617 15:20:22
B       11      20220617 10:58:09       W1      33.01   22.49   14.00   6.096   5.454
G       12      20220617 10:58:09       G0      42.786433       -81.593169      222.7   7       1.07    20220617 15:20:23
B       13      20220617 10:58:11       W1      12.63   22.03   13.47   7.241   4.753
G       14      20220617 10:58:11       G0      42.786433       -81.593173      223.2   8       0.95    20220617 15:20:25
B       15      20220617 10:58:12       W1      34.88   29.79   16.23   7.075   4.050
G       16      20220617 10:58:12       G0      42.786431       -81.593175      223.6   8       0.95    20220617 15:20:26
B       17      20220617 10:58:20       W1      36.33   22.74   13.99   7.009   3.707
G       18      20220617 10:58:20       G0      42.786430       -81.593162      222.0   8       0.95    20220617 15:20:34
B       19      20220617 10:58:22       W1      57.78   29.79   17.32   10.06   6.099
G       20      20220617 10:58:22       G0      42.786429       -81.593159      221.5   8       0.95    20220617 15:20:36
B       21      20220617 10:58:24       W1      34.47   26.24   17.57   8.903   4.596
G       22      20220617 10:58:24       G0      42.786429       -81.593159      221.3   8       0.95    20220617 15:20:38
B       23      20220617 10:58:26       W1      66.93   41.91   17.18   8.167   4.903
G       24      20220617 10:58:26       G0      42.786430       -81.593161      221.7   8       0.95    20220617 15:20:40
```

Figure B-3. Example of LAI data on June 17, 2022.

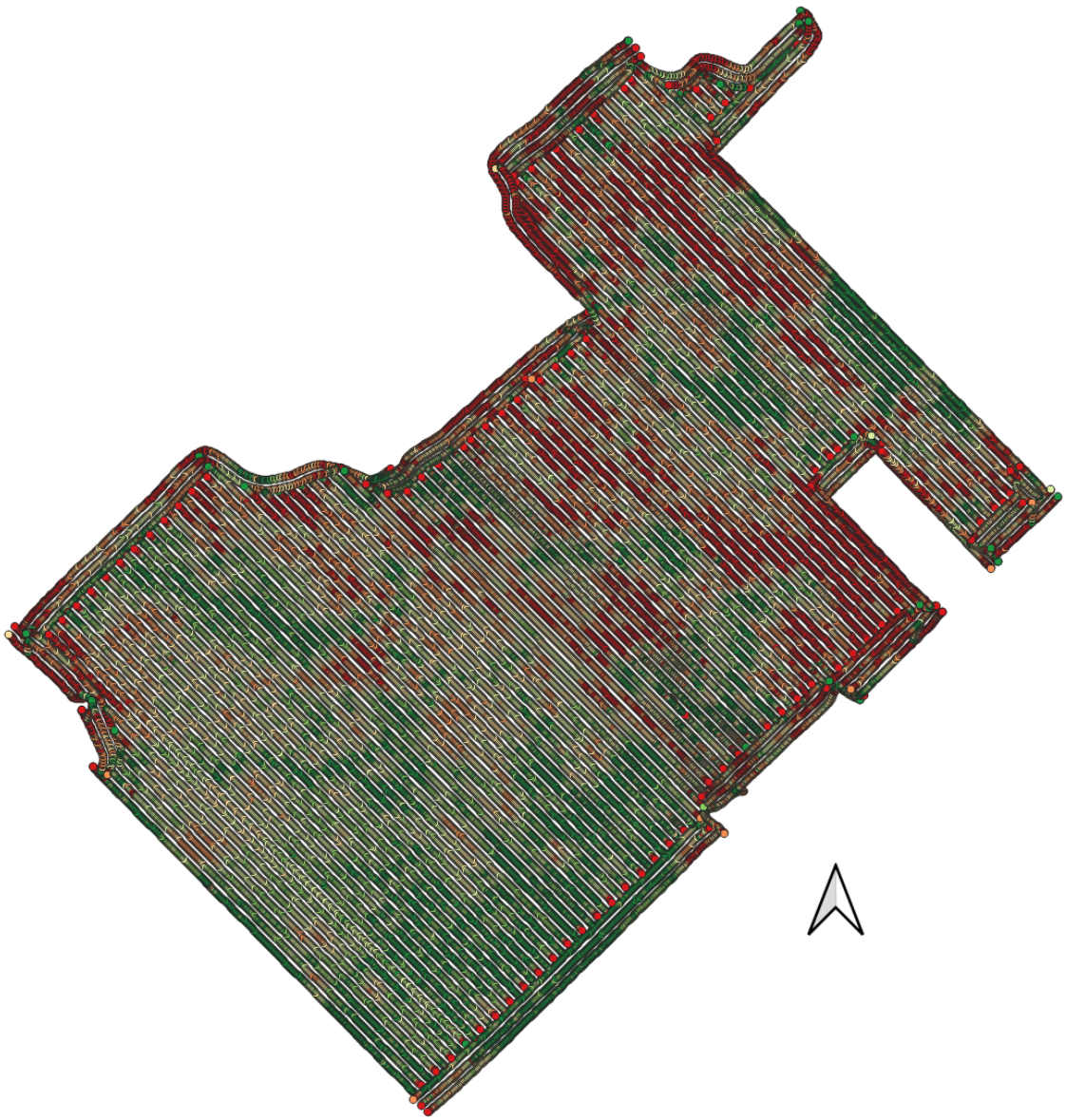


Figure B-4. Example of raw yield data of the studied field in 2020. Generated from the yield sensor mounted on the harvester. Green represents high and red represents low.

5.3 Appendix C – Remote Sensing Imagery

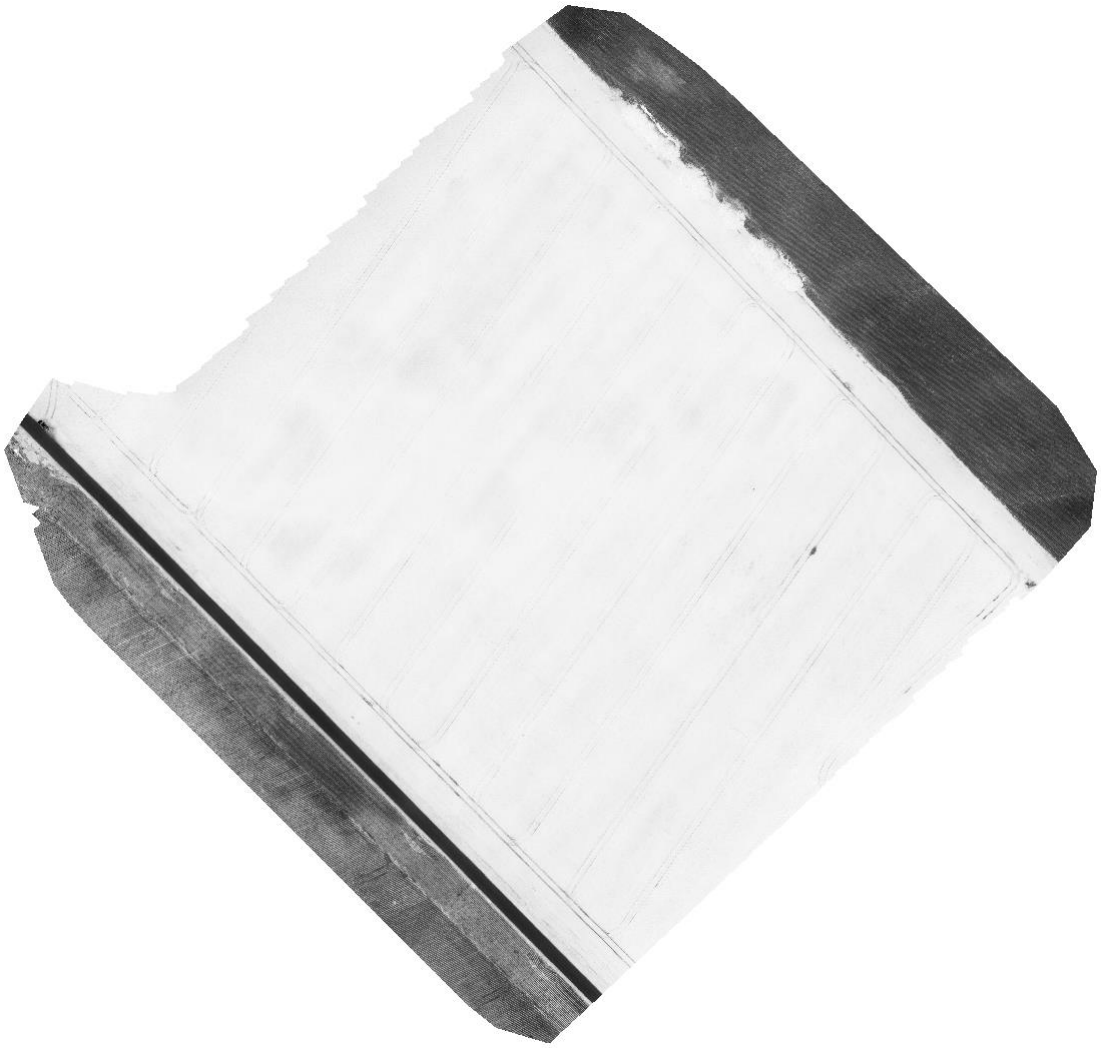


Figure C-1. Example of NDVI orthomosaic generated from imagery captured using a MicaSense multispectral camera mounted on a UAV on June 19th, 2022. Brighter equals higher NDVI value.

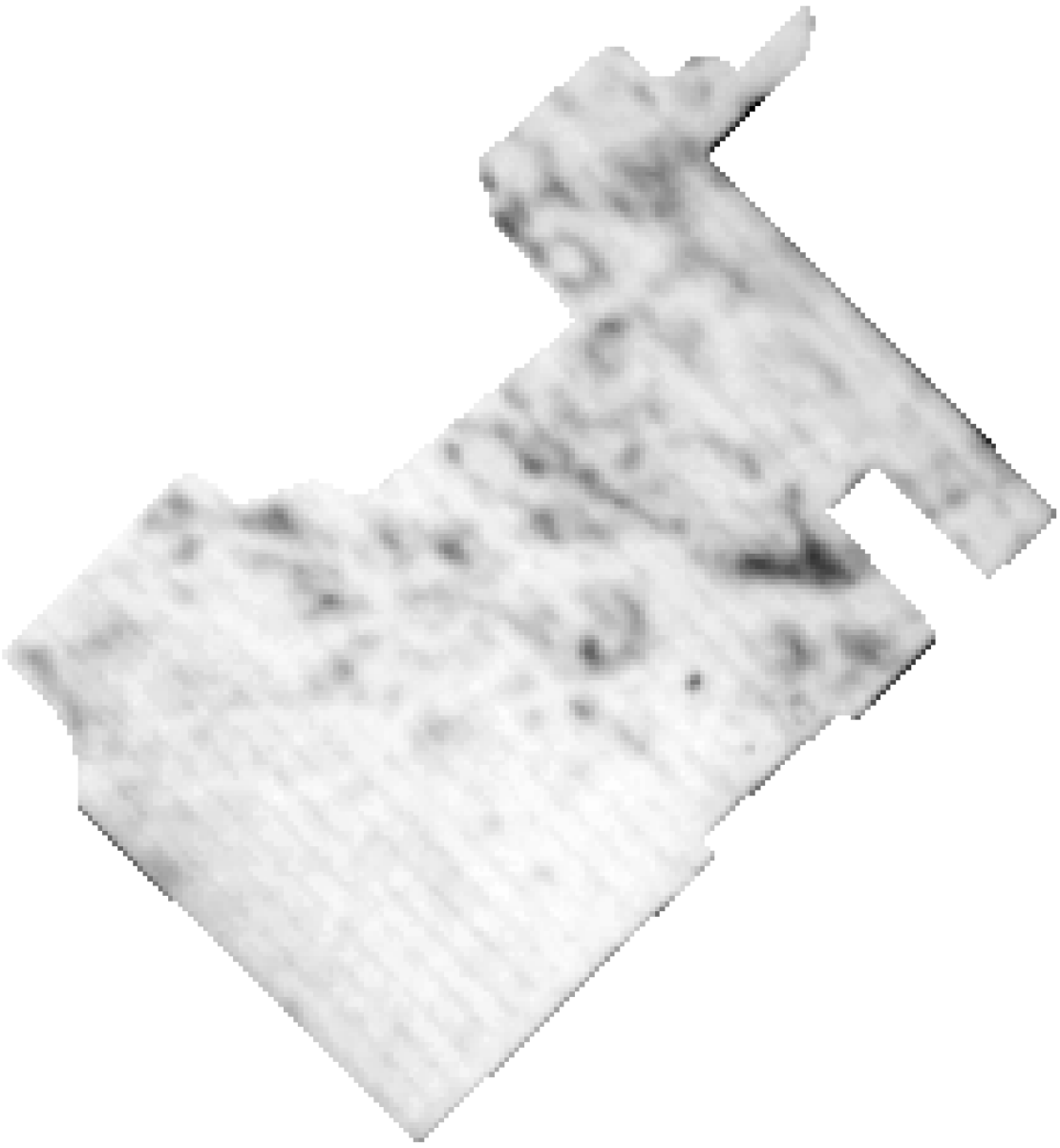


Figure C-2. Example of NDVI-1 orthomosaic generated from imagery captured by VENµS on June 16th, 2020. Brighter equals higher NDVI value.

5.4 Appendix D – Code

```
library(randomForest)
library(e1071)

# Load data
library(readxl)

wheat <- read_excel("E:/data.xlsx")
View(wheat)

# Data split
set.seed(001)
train <- sample(nrow(wheat), 0.7*nrow(wheat), replace = F)
trainset <- wheat[train,]
validset <- wheat[-train,]

# RF of training set
RFmodelT <- randomForest(Biomass ~., data = trainset)

# Variable importance plot
varImpPlot(RFmodelT)
plot(RFmodelT)

# RF
predictT <- predict(RFmodelT, newdata = trainset)
predictV <- predict(RFmodelT, newdata = validset)

# RF metrics
rmsemodT <- sqrt(mean((predictT - trainset$Biomass)^2))
rmsemodV <- sqrt(mean((predictV - validset$Biomass)^2))

RsqRFmodT <- lm(trainset$Biomass ~ predictT, data = trainset)
RsqRFmodV <- lm(validset$Biomass ~ predictV, data = validset)

# SVM
SVRmodelT <- svm(Biomass ~., data = trainset)

predictTsvr <- predict(SVRmodelT, newdata = trainset)
predictVsvr <- predict(SVRmodelT, newdata = validset)

# SVM metrics
rmsemodTsvr <- sqrt(mean((predictTsvr - trainset$Biomass)^2))
rmsemodVsvr <- sqrt(mean((predictVsvr - validset$Biomass)^2))

RsqSVRmodT <- lm(trainset$Biomass ~ predictTsvr, data = trainset)
RsqSVRmodV<- lm(validset$Biomass ~ predictVsvr, data = validset)
```

Figure D-1. R code of Random Forest and Support Vector regression models used in AGB estimation.

```
library(randomForest)
library(e1071)
library(caret)
library(parallel)
library(doParallel)
library(kernlab)

# Load data
csv_path <- "E:/data.csv"
wheat <- read.csv(csv_path, fileEncoding="UTF-8-BOM")

# Data split
set.seed(001)
trainIndex <- createDataPartition(wheat$Paul, p = 0.7, list = FALSE)
trainset <- wheat[trainIndex, ]
validset <- wheat[-trainIndex, ]

# Set up parallel processing
numCores <- detectCores() - 1
cl <- makeCluster(numCores)
registerDoParallel(cl)

# Define cross-validation
control <- trainControl(method = "cv", number = 10)

# Train Random Forest model with cross-validation
RFmodel <- train(Paul ~ ., data = trainset, method = "rf", trControl = control)
RFmodel

# Variable importance plot
varImpPlot(RFmodel$finalModel)
plot(RFmodel$finalModel)

# RF predictions and evaluation
predictTrain <- predict(RFmodel, newdata = trainset)
predictValid <- predict(RFmodel, newdata = validset)

# RMSE for RF
rmseTrain <- sqrt(mean((predictTrain - trainset$Paul)^2))
cat("Training RMSE for RF:", rmseTrain, "\n")
rmseValid <- sqrt(mean((predictValid - validset$Paul)^2))
cat("Validation RMSE for RF:", rmseValid, "\n")

# R2 for RF
R2Train <- cor(trainset$Paul, predictTrain)^2
cat("Training R-squared for RF:", R2Train, "\n")
R2Valid <- cor(validset$Paul, predictValid)^2
cat("Validation R-squared for RF:", R2Valid, "\n")

# Train SVR model with cross-validation
SVRmodel <- train(Paul ~ ., data = trainset, method = "svmRadial", trControl = control)
SVRmodel

# Predictions and evaluation for SVR
predictTrainSVR <- predict(SVRmodel, newdata = trainset)
predictValidSVR <- predict(SVRmodel, newdata = validset)
```

```
# RMSE for SVR
rmseTrainSVR <- sqrt(mean((predictTrainSVR - trainset$Paul)^2))
cat("Training RMSE for SVR:", rmseTrainSVR, "\n")
rmseValidSVR <- sqrt(mean((predictValidSVR - validset$Paul)^2))
cat("Validation RMSE for SVR:", rmseValidSVR, "\n")

# R2 for SVR
R2TrainSVR <- cor(trainset$Paul, predictTrainSVR)^2
cat("Training R-squared for SVR:", R2TrainSVR, "\n")
R2ValidSVR <- cor(validset$Paul, predictValidSVR)^2
cat("Validation R-squared for SVR:", R2ValidSVR, "\n")
```

Figure D-2. R code of Random Forest and Support Vector regression models used in yield prediction.

5.5 Appendix E – Copyrighted Material & Permissions

Figure E-1. Certificate of publication for chapter 2

FigureE-2. Certificate of publication for chapter 3

Curriculum Vitae

Name: Marco Spencer Chiu

Post-secondary Education and Degrees:
University of Western Ontario
London, Ontario, Canada
2017-2021 B.Sc. (Hons) Geography and Environment

University of Western Ontario
London, Ontario, Canada
2021-2024 M.Sc. Geography and Environment

Honours and Awards:
Western University Dean's Honour List, Geography and Environment
2018-2021

Western Gold Medal, B.Sc. (Hons) Geography and Environment
2021

Related Work Experience
Lab Technician (co-op)
A&L Canada Laboratories Inc.
2020

Teaching Assistant
University of Western Ontario
2021-2023

Research Assistant
GITA Lab, University of Western Ontario
2021-2024

Field Technician
A&L Canada Laboratories Inc.
2023-2024

Publications:

Chiu, M. S., & Wang, J. (2024). Evaluation of Machine Learning Regression Techniques for Estimating Winter Wheat Biomass Using Biophysical, Biochemical, and UAV Multispectral Data. *Drones*, *8*(7), Article 7. https://doi.org/10.3390/drones8070287

Chiu, M. S., & Wang, J. (2024). Local Field-Scale Winter Wheat Yield Prediction Using VENμS Satellite Imagery and Machine Learning Techniques. *Remote Sensing*, *16*(17), Article 17. https://doi.org/10.3390/rs16173132