Digital agriculture driven by big data analytics: A focus on spatio-temporal crop yield stability and land productivity

Nketia, K.A.[1], Ha., T.[1], Fernando, H.[1], van Steenbergen, S. [1], Shirtliffe, S.J.[1]
[1]Department of Plant Sciences, College of Agriculture and Bioresources, University of Saskatchewan, Canada

Abstract
Digital agriculture frontier has been propelled by the pressing imperativeness to address the growing food demands, effects of climate change, and sustainable land management. Canada has made a substantial commitment to reducing greenhouse gas emissions by ~45% by 2030 and to achieve net-zero by 2050. To support the attainment of these ambitious targets, our research delves into mapping spatio-temporal crop yield stability and land productivity. We utilized a multi-scale stacked machine learning approach to train a model using yield data from fields within ~5 m hectares spanning 5 years across Western Canada. Key model inputs included environmental-landscape explicit variables that heavily drives seasonal yields. Canola and wheat yields were predicted and subsequently used to classified arable lands into three productivity classes, i.e., stable-high, stable-low, and unstable. Validation assessment of predicted yields recorded temporal $RMSEs$ (R^2) ranging from 0.85 to 1.21 t ha^{-1} (0.56) for canola and 0.86 to 1.20 t ha^{-1} (0.66) for wheat. Biomass accumulation, precipitation, peak and end of season (day of years), and landscape derivatives such as valley depth and topographic position index were consistent key explicit variables driving yield prediction across space and time. Productivity classes mapped for the Canadian prairies is fundamental to the development of data-driven precision agriculture strategies. Such efforts are imperative in the formulation of crop input prescriptions that target areas that provides high economic returns while less productivity areas can be managed with strategies that contribute to the advancement of marginal land conservation and restoration efforts. These efforts have the potential to sequester carbon, reduce the need for agrochemical inputs, enhance biodiversity, and still provide economic benefits to producers.

Keywords
Crop yield prediction; digital agriculture yield stability; land productivity; stacked machine learning

Introduction: Improved agriculture technologies and big data analytics have been identified as a potential solution to addressing the global food demands, mitigating changing climate effects, and promoting sustainable land use. However, this backdrop puts enormous pressure on limited arable lands resources to increase production per unit area, and thus has stimulated questions about stability and productivity of arable lands. To adapt and sustain improved crop production over time while diversifying agricultural systems to maximize the efficiency of farm inputs as well as seeking the commitment to reducing greenhouse gas emissions, agricultural systems' stability is critical. For most producers, how agricultural stability drives economic predictability at low risk is crucial to optimizing farm inputs [1]. For this reason, finite arable land use needs continual intensification adaptations to optimize inputs to realize improved yields. Against this backdrop and gaps in the literature, we investigate (1) the robustness and validity of using big data analytics, artificial intelligence, and remote sensing data for productivity assessment and, (2) the synergetic use of long-term yield data and soil-landscape-climate in the assessment of crop stability [2]. It is therefore imperative to investigate at large scales to adequately influence production recommendations while devising agricultural systems that can effectively adapt to changing climate.

Study area: The study was conducted within the prairie landscapes of Western Canada covering an area of ~5 m hectares from 2017 to 2022. The selected pilot regions are limited to arable lands and are representative of the heterogeneities of the Canadian prairie. Grower's harvest data we voluntarily contributed to this study. Yield datasets were corrected for sensor and operational, header swath and switch, travel time, flow, and reading cycles errors. Production practices used in the study area follows standard production in the northern Great Plains of North America, with mechanized dry land agriculture including mineral fertilization, and limited to no irrigation. In this study, yield data from ~1000 farmer fields were compiled for model calibration and validation. In this study, the two dominant crops, canola and wheat were considered.

Methods: Modelling land productivity and yield stability were implemented under 5 main stages: (1) designing a yield-class ratings and spatially discretization yields into stratify sub-samples, which reflects within-field variability, (2) preparation of dynamic and static explicit variables, (3) designing a regression task, (4) fitting an stacked-machine learning model based on yield classes, (5) spatial validation of predicted yields, and (6) using a measure of fuzziness to threshold predicted yields into productivity classes [3]. Models were trained in R with $Ranger$ and ensembled using their squared errors as weights. All processes were executed on Compute Canada clusters and Google Earth Engine via R and Python programming languages.

Results and Discussion: After an iterative leave-field-out validation, precipitation, crop phenological stages including peak and end of season, biomass accumulation, as well as landscape derivatives such as valley depth and topographic position index were consistently ranked the optimal variables across years in both yield classes and crop types. These observations corroborate expert knowledge documented for other landscapes [4].
Predicted yields ranged from <1 t ha^{-1} to 9 t ha^{-1}. In correlating held-out fields with predicted yields, accuracies ($RMSEs$; R^2) for canola and wheat ranged from 0.85 to 1.21 t ha^{-1} (0.56), and 0.86 to 1.20 t ha^{-1} (0.66) respectively.
Fig. 1 shows yearly predicted maps and their respective productivity classes for a field. Consistently across crops, 2021 had the lowest yields because of drought. Yield difference between unstable and stable-low were similar. Yields were generally stable and high. A distinctive observation for unstable regions is the time dependencies of within-field variabilities. This could be attributed to an interaction of crop to variables that drive within field variability.
Generating a wider-scale model is to account for poor dispersion and limited to no discrete data points across an area. Conversely, it was observed that in sparse networks like those of this study, key regional-localized effects impact the global model's extrapolation power. This finding thus suggests the inclusion of approaches such as area of applicability and dissimilarities for the prediction models.
There were yield instability over years due to the negative skewed reference data across the study area. It is therefore important to expand our approach across multiple years, where retrospective predictions are considered.

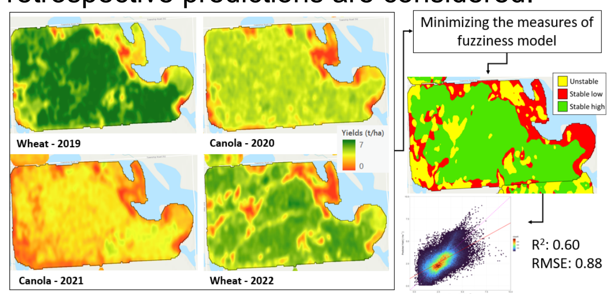

Fig. 1: Extracts of space-time predicted yields translated into three land productivity classes.

Conclusion: The comprehensive use of big data analytics and crop-specific data to understand yield stability and land productivity to draw demonstrative conclusions for Western Canada is fundamental to the development of sustainable precision agriculture strategies. Based on our findings, growers can identify alternative usage for less productive areas, such as the cultivation of more diverse perennial forages on marginal areas, which has the potential to sequester carbon, reduce the need for agrochemical inputs, enhance biodiversity, and still provide economic benefits to producers and the environment.

References:
[1] Basso, B., & Antle, J. (2020). Digital agriculture to design sustainable agricultural systems. Nature Sustainability, 3(4), 254–256
[2] Reckling, M., Ahrends, H., Chen, TW. et al. Methods of yield stability analysis in long-term field experiments. A review. Agronomy for Sustainable Development. 41, 27 (2021).
[3] Huang, L.-K., & Wang, M.-J. J. (1995). Image thresholding by minimizing the measures of fuzziness. Pattern Recognition, 28(1), 41–51.
[4] Thelemann, R., Johnson, G., Sheaffer, C., Banerjee, S., Cai, H., & Wyse, D. (2010). The effect of landscape position on biomass crop yield. Agronomy Journal, 102(2), 513-522.