University of São Paulo
"Luiz de Queiroz" College of Agriculture

Management zones and space-time prediction of soybean yield variability: machine learning techniques applied to soil physical quality parameters

Gislaine Silva Pereira

Thesis presented to obtain the degree of Doctor of Science. Area: Agricultural Systems Engineering

Piracicaba
2023

Gislaine Silva Pereira
Agricultural Engineer

Management zones and space-time prediction of soybean yield variability: machine learning techniques applied to soil physical quality parameters

Advisor:
Prof. Dr. **LEANDRO MARIA GIMENEZ**

Thesis presented to obtain the degree of Doctor of Science. Area: Agricultural Systems Engineering

Piracicaba
2023

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Pereira, Silva Gislaine

Management zones and space-time prediction of soybean yield variability: machine learning techniques applied to soil physical quality parameters / Gislaine Pereira. - - Piracicaba, 2023.

68 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. ACP 2. Agrupamentos 3. Random forest 4. Árvores de classificação 5. Aquacrop 6. CROPGRO 7. Soja I. Título

ACKNOWLEDGEMENTS

My sincere thankfulness,

For God for giving me the strength to complete this journey.

My parents, brother and family for their support and patience.

My husband for all the encouragement and admiration throughout this journey.

My supervisor for his support and patience throughout my PhD.

All my friends, who were with me throughout or at some point in this process.

Thank you All

CONTENTS

RESUMO

Zonas de manejo e predição espaço-temporal da variabilidade da produção de soja: Técnicas de machine learning aplicadas a parâmetros de qualidade física do solo

Os métodos e ferramentas da agricultura de precisão são chave para garantir o aumento da produção de soja. Para isso, conhecer as variabilidades intra-campo é a chave para auxiliar na tomada de decisão do produtor agrícola. Embora os métodos de modelagem para estimativa da produção sejam baseados em condições regionais e/ou modelos agroecossistêmicos que não representam escalas locais, esta tese tem como objetivo utilizar técnicas de aprendizado de máquina em busca de melhorar a qualidade de dados para previsão de produtividade a nível de zonas de manejo. Sendo assim, esta pesquisa foi dividida em três capítulos que utilizam técnicas e métodos com foco em agricultura de precisão para validar a necessidade de garantir um maior suporte ao produtor agrícola a nível local. O primeiro capítulo teve como objetivo utilizar machine learning para melhorar a qualidade de dados oriundos do mapeamento de produtividade e informações de sensores de alta resolução na geração de zonas de manejo (MZs) além de validar as diferenças entre e intra MZs sob os aspectos relacionados as variáveis de solo. A hipótese deste primeiro capítulo esteve centrada na necessidade de utilizar a técnica de análise multivariada de componentes principais (PCA) para melhorar a qualidade de predição das MZs a partir dos dados originais. O segundo capítulo teve como objetivo estimar a produtividade de soja em cada MZs para múltiplos anos, em função de mapas de água no solo e do desenvolvimento das culturas. Como hipóteses para o capítulo se avaliou a necessidade de comprovar a existência de variabilidade da produtividade intra-regiões. Uma segunda hipótese focou em testar a qualidade de superfícies de refletância no infravermelho próximo (NIR) para representar o desenvolvimento da cultura em comparação ao uso de índice de vegetação por diferença normalizada (NDVI). A terceira hipótese foi de que a técnica de machine learning Random Forest (RF) apresenta uma maior qualidade de predição da produtividade devido sua eficiência em trabalhar com dados desbalanceadas em comparação ao método convencional de análise de regressão múltipla (MLR). O objetivo do terceiro capítulo foi entender a sensibilidade de modelos de cultura (Aquacrop e CROPGRO) na estimativa da produtividade de soja a níveis de zona de manejo em função de fatores de solo. A hipótese deste capítulo verificou a capacidade de modelos de cultura em apresentar variabilidade reduzida para estimar a produtividade em função da variação desses fatores, principalmente água no solo. Os resultados do capítulo 1 evidenciaram que a técnica de PCA resulta em maior qualidade de agrupamento em relação ao método convencional de normalização, além de garantir uma maior estabilidade na definição do número de MZs. As variáveis de solo foram fundamentais para validação das especificidades em cada região, o que foi demonstrado com a técnica de árvore de classificação. Os resultados do capítulo 2 mostraram as diferenças entre as superfícies de água no solo em função das MZs, evidenciando a importância do manejo diferenciado nas regiões, mesmo em nível local. A refletância NIR melhorou a qualidade da previsão da produtividade de soja nas regiões em comparação ao uso do NDVI. O método de RF apresentou desempenho superior nas estimativas em comparação ao método de MLR. Os resultados do capítulo 3 evidenciaram que os modelos de cultura Aquacrop e CROPGRO apresentaram desempenho variável na estimativa da produtividade de soja nas zonas em decorrência de anos predominantemente secos ou úmidos. Mais estudos devem ser realizados com modelos de cultura para previsão da produtividade de soja a nível local. Por fim, como resultado do trabalho foi possível evidenciar a importância da avaliação em escala local e do uso de métodos de machine learning e mapeamento digital como suporte à agricultura de precisão. Verificou-se que o uso de MZs é adequado para conhecer a variabilidade de fatores de solo e planta que podem influenciar no planejamento para uso localizado de insumos e impactar nos resultados de produtividade em um mesmo talhão. Como estudos futuros, sugere-se aqueles envolvendo o uso de sensores locais para monitorar a variabilidade temporal do clima, solo e planta, como meios para elevar o desempenho de métodos de machine learning na agricultura.

Palavras-chave: ACP, Agrupamentos, Random forest, Árvores de classificação, Aquacrop, CROPGRO, Soja

ABSTRACT

Management zones and space-time prediction of soybean yield variability: machine learning techniques applied to soil physical quality parameters

Methods and tools for precision agriculture are the key to ensuring increased soybean production. In this respect, knowledge of intra-field variability is the key to helping the agricultural producer in the decision-making process. Although methods for modelling production are based on regional conditions and/or agroecosystem models that do not represent local scales. The aim of this thesis is to use machine learning techniques to improve data quality for predicting yield at the management-zone level. The research was divided into three chapters that use techniques and methods focused on precision agriculture to validate the need to guarantee greater support to the farmer at the local level. The first chapter sought to use machine learning to improve the quality of data and the information from high-resolution sensors in generating management zones (MZs). In addition to validating the differences between and within MZs related to soil factors. The hypothesis of this first chapter was centred on the need to use principal component analysis (PCA) to improve the quality of MZ prediction based on observed data. The second chapter aimed to estimate soybean yield in each MZ over several years based on maps of soil water and crop development. One hypothesis for the chapter was the need to confirm the existence of the variability of intra-regional yield. The second hypothesis focused on testing the quality of near infrared reflectance (NIR) surfaces to represent crop development compared to using vegetation index (NDVI). The third hypothesis was that the machine learning technique Random Forest (RF) affords better quality yield prediction due to its efficiency in working with unbalanced data compared to the conventional method of multiple linear regression analysis (MLR). The aim of the third chapter was to understand the sensitivity of crop models (Aquacrop and CROPGRO) in estimating soybean yield at the management-zone level, especially as a function of available soil water. The hypothesis of this chapter was in the ability of crop models to show less variability when estimating yield based on the variations in soil water. The results of Chapter 1 showed that the PCA techniques afforded higher-quality clustering compared to the conventional method of normalisation, besides ensuring greater stability in defining the number of MZs. Soil variables were fundamental for validating the specific characteristics of each region using the classification tree technique. The results of Chapter 2 showed the differences between digital soil water surfaces as a function of the MZs, demonstrating the importance of different management practices in each region, even at the local level. NIR reflectance improved quality predictions of soybean yield in each region compared to the use of NDVI. The RF method afforded higher-quality estimates compared to the MLR method. The results of Chapter 3 showed that the Aquacrop and CROPGRO models showed variable performance when estimating soybean yield in each zone in occurrence of wet and dry years. More studies should be carried out using crop models to predict soybean yield at local level. In this way, was possible to highlight the importance of evaluation on a local scale, with the use of machine learning methods and digital mapping to support precision agriculture. The use of MZs is the adequate to understanding the variability of soil and plant factors that will later influence planning for the localised use of inputs, impacting yield at same field. For future studies, the use of local sensors to continuously monitor variability of climate, soil and plant variability to improve precision of machine learning methods in agriculture.

Keywords: PCA, Clusters, Random forest, Classification trees, Aquacrop, CROPGRO, Soybean

1. GENERAL INTRODUCTION

The growing increase in soybean [*Glycine max* (L.) Merrill] production is one of the principal guarantees for meeting the global demand for food. In the 2023/24 season, Brazil will contribute with 38% of the world production of this oilseed (USDA, 2023). Soybean production in Brazil is estimated to reach 175 million tons, 16% higher than in 2022/23, with an increase of only 4% in planted area (USDA, 2023; CONAB, 2023). During the crop cycle, soybean development is mainly influenced by the soil and its intrinsic characteristics, which directly or indirectly impact the yield. The soil is the basis for successful soybean production under rainfed conditions, whose function is to support the system via a balance between physical-chemical and biological factors (HILLEL, 2007). A balanced soil is a guarantee of greater agricultural yield, especially the physical characteristics that are affected by anthropogenic management. Anthropogenic activity contributes to soil degradation, favoring on compaction process and reducing the water available to crops (FRANCHINI et al., 2017; MORAES et al., 2018a; MORAES et al., 2018b).

Understanding the dynamics of soil factors is one of the first steps in understanding the processes that are key to crop development (ROSSATO et al., 2017). Reichert et al. (2020) stated that the hydraulic properties of the soil are not continuous in an area, and vary depending on soil texture. Soil texture is one of the main natural factors that directly or indirectly affect other variables, such as soil water retention, bulk density, apparent electrical conductivity and organic matter. Soil texture is closely linked to water retention, which is lower in sandy soils than in clayey soils, a result of variations in the surface area of the particles, which can affect the water retention capacity (KIRKHAM, 2005). Under conditions where the soil physics is unsuitable, the chemical aspects (impediment to the absorption and availability of ideal amounts of nutrients for plants) and biological aspects (imbalance between the communities of microorganisms present in the soil) are also affected. At the field level, these factors can help boost an increase or decrease in soybean production, especially through interactions between the soil-plant system and the atmosphere. On the other hand, the main basis for decision-making by the farmer regarding soil fertility begins with the chemical properties of the soil, due to the time taken in collecting and measuring soil physics. Furthermore, even when using the chemical properties, agricultural areas can still be managed uniformly, not taking into account the spatial-temporal variability of the production system impacting production costs and the sustainable use of resources (PERRON et al., 2018; GAVIOLI et al., 2019).

On the scale of field, the integration between soil factors, plant development and climate seasonality is essential to understanding crop responses to water deficit, often related to the physical production environment of the plant (VIANNA, 2018). The use of climate, plant and soil predictors for estimating local crop yields should be investigated, as this can be one way of evaluating the spatial-temporal variability of crops. Although several studies may evaluate the spatial and temporal distribution of water in the soil (VIEIRA et al., 2008; ZHANG et al., 2013), few effectively monitor the influence of this variability during the crop growth cycle (HUANG et al., 2019), exploring its relationship to yield using high-resolution information (BOENECKE et al., 2018; YOST et al., 2019). Soil water can limit crop production, affecting the environment, atmospheric and hydrological processes. The spatial and temporal distribution of water in the soil can be obtained through high resolution maps, and interacts with such control factors as the type of crop, soil parameters and terrain (Huang et al., 2019). The tensions that act on the soil through potentials, adsorption forces and capillarity, control the movement of water and its availability for metabolising the crop, so that, of all the water stored in the soil, only a part is actually available to the plants (REICHARDT et al. al., 1979; JURY and HORTON, 2004). The water balance must therefore be considered and evaluated throughout the layer of soil covered by the plant roots (ROSSATO et al., 2004). At the field level, more researchs should be carried out in order to investigate the variation in soil water and its relationship with crop yield (VEREECKEN et al., 2016), especially in terms of management zones (MZs).

To ensure increased soybean production in an area, several factors should be considered, with precision agriculture (PA) being the key to this increase (BREUNIG et al., 2020). These techniques make it possible to recognise which factors influence the crop cycle, helping to understand the stability of agricultural yield over the years. A knowledge of the variations that govern the delineation of MZs in agricultural areas can help minimise the differences within a region and maximise the differences between regions, ensuring greater profit and production quality, and reducing environmental impact (PERRON et al., 2018). MZs are, by definition, differentiated management units within the same area (MORAL et al., 2010). To generate the MZs, clustering techniques based on mathematical methods are used. Among the information used to create these different management areas are apparent electrical conductivity (ECa) surfaces, altitude and yield maps (MOLIN et al., 2008; BUTTAFUOCO et al., 2010; BREUNIG et al., 2020). The ECa is widely used for large-scale measurements due to the ease of mapping in large areas over a short period (CORWIN and LESCH, 2005). According to the authors, ECa is influenced by the water content of the soil, bulk density, texture and organic matter. Bottega et al. (2022) used ECa surfaces as an alternative for determining MZs, obtaining a correlation with the clay content and reducing the need for soil sampling. As an aid in generating MZs, the first monitors of crop yield appeared during the 90s for mapping the variability of agricultural production in the field. Over the years, it has become possible to use the historic of harvest maps for decision-making in an attempt to understand patterns related to soil seasonality or climate conditions from harvest to harvest (LEROUX et al., 2018). For Blasch et al. (2020), the use of time series for crop yields based on machine data facilitates an understanding of the spatial-temporal variations in an agricultural area. On the other hand, there is a need to understand whether the use of machine learning (ML) can be an alternative in processing and mapping these patterns in large machine databases (LEUKEL et al., 2023).

With the advancement in digital agriculture resulting from the transformation of data into information, machine learning techniques must be understood for more-assertive decision-making. In this respect, the possibility of predicting the yield of agricultural crops using digital data becomes the next challenge to be overcome. This advance will only be possible with the use of artificial intelligence combined with supervised and unsupervised machine learning techniques (LEUKEL et al., 2023). ML consists of a set of techniques that improve the performance of systems through computational learning, developing learning algorithms that are built from a database to create predictive or observation models (ZHOU, 2021). The fuzzy c-means method (FCM) (BEZDEK, 1981) is widely used to understand clustering patterns in MZs using quantitative and/or qualitative agronomic variables (GAVIOLI et al., 2019; JENA et al., 2019). This technique of unsupervised learning allows the similarities between different soil and plant attributes to be understood. Li et al. (2008) determined MZs using FCM with ECa and yield maps, stating that the approach is suitable for delimiting the regions. Random forest (RF) is a supervised learning method combined with classification trees that searching for a correct combination will result in the final estimate. RF is highly accurate due to good outlier adjustment and is one of the most used methods in data mining (LIU et al., 2012). Another well-known method is the use of classification and regression trees (BREIMAN, 1984), which aim to predict the behaviour of a given factor as a function of predictor variables (LOH, 2011). Furthermore, according to the author, cutting conditions are defined starting from one root node, with the set repeatedly divided into internal nodes until the stopping point is reached, when the final classes of the predicted factor are determined. PCA (Principal Component Analysis) is a method that is widely used in agriculture, and consists in reducing the dimensionality of large data sets. Use of the PCA technique can help reduce noise and increase the quality of the information from these large data sets, in addition to performing better than univariate methods of analysis (HASAN et al., 2021).

Studies by Burdett and Wellen (2022) showed the importance of using RF and decision trees to estimate the yield of agricultural crops based on soil properties and terrain. The methods used in the study performed better than conventional methods such as multiple linear regression (MLR). Da Silva et al. (2020) used decision trees to predict soybean yield over two seasons as a function of different crop and ground vegetation indices, obtaining an estimate accuracy of 93%. According to the authors, indices that were related to the soil showed better predictive performance. Madarasz et al. (2021), worked with RF to predict the risk of soil erosion in agricultural areas. The authors showed that the method was suitable for predicting surface runoff and soil loss based on maize data. Metwally et al. (2019) classified MZs based on soil properties and PCA, with the first four components used to carry out the grouping and determine the regions. For the authors, the technique helped in reducing the dimensionality and variability of the properties under study, where the four PCs explaining 84% of the variance in the data. Kinoshita et al. (2021) also used PCA to estimate the variability of maize yield as a function of climate seasonality. The authors obtained a variance of between 60%-78% explained by the first 2 PCs, and recommended the technique. Jiang et al. (2020) used different machine learning methods to predict the spatial yield of rice in agricultural areas, stating that the use of regression trees and RF (RMSE ~ 2.0, R2 ~ 0.60) gave better performance than the conventional MLR method (RMSE ~ 2.3 and R2 ~0.5). According to the authors, this type of approach combined with decision-making on the rational use of agricultural inputs may be key to understanding field patterns and correlating these with soil and climate variability.

Another common approach for estimating the yield of agricultural crops are the crop models widely used around the world for predicting soybean development and yield. However, this prediction is often based on general or point conditions, in which mean yield values are considered for a given region and condition without considering spatial variability at the management-zone level. Models such as Aquacrop-FAO are used to estimate crop development and yield as a function of the soil water balance, whereas models from the DSSAT platform, such as CROPGRO-Soybean, consider the variation in photoperiod to estimate yield. For current and future studies, understanding the impact of using these crop models to predict soybean yield at a spatial level is an opportunity to support agricultural studies and decision-making by the farmer (SINGH et al., 2023). Several studies have used crop models to estimate soybean yield under different conditions, considering the water deficit (BATTISTI et al., 2017; GIMENEZ et al., 2017; MORALES-SANTOS et al., 2023), climate seasonality (EJAZ et al., 2022), cultivar calibration (AKUMAGA et al., 2023) and soil and crop indices (SALMERÓN and PURCELL, 2016; MULAZZANI et al., 2022), obtaining variations in the RMSE of 30 to 2500 kg ha^{-1}.

Other alternatives consist in the use of remote sensing to predict crop yields. According to Da Silva et al. (2020), yield can be predicted from NDVI surfaces, ensuring greater understanding of the management of agricultural areas. However, the authors state that the use of indices that consider aspects and characteristics related to the soil have a greater correlation with yield than does the NDVI. Kross et al. (2020), also using NDVI surfaces to predict grain yield, state that this information is essential to understanding different crop behaviours and local field conditions. Breunig et al. (2020), determining MZs using remote sensing, showed the importance of vegetation indices to predict the yield of commercial crops. According to the authors, NIR reflectance and the NDVI proved to be important predictors of crop biomass, providing useful results that can help to reduce the need for soil and plant analysis.

The use of learning and modelling methods that seek to improve the predictive ability and understanding of these processes on a spatial and temporal scale will help reduce the response time and aid in agricultural management practices (DORIGO et al., 2011). Among the justifications for developing this research are (i) the need to increase the amount of research at the field level that evaluates the performance of machine learning to support the advancement of precision agriculture; (ii) seek alternatives for the farmer to make decisions based on the use of a large set of machine data transformed into reliable information; (iii) combine the use of artificial intelligence methods with management information from production systems to predict yield in regions with different soil characteristics; (iv) understand the performance of consolidated crop models in predicting soybean yield at the management-zone level. The present thesis comprises three chapters, and aims to answer various questions concerning the prediction of soybean yield using management zones. The first chapter, entitled "Machine learning to support management zones based on soil physics and soybean yield surfaces" aims to investigate the use of historical yield maps, electrical conductivity and altitude maps to define MZs using unsupervised (PCA and K-means) and understand soil relations trough supervised methods (classification tree). The second chapter, entitled "Multi-year simulation of soybean yield from the digital mapping of crops and soil water in management zones" aims to predict the spatial-temporal variability of soybean yield using a supervised technique (Random Forest) in three management zones based on digital soil water maps and vegetation indices from remote sensing. The third chapter, entitled "Performance of crop models to predict soybean yield on physical soil factors in management zones" aims to predict soybean yield based on variations in soil factors at the management-zone level using crop models (Aquacrop and CROPGRO), and its correlation with changes in the available soil water. At the end of the three chapters, final considerations were formulated in order to highlight the conclusions for continuing with the studies about themes addressed here.

References

AKUMAGA U, GAO F, ANDERSON M, DULANEY WP, HOUBORG R, RUSS A, et al. Integration of Remote Sensing and Field Observations in Evaluating DSSAT Model for Estimating Maize and Soybean Growth and Yield in Maryland, USA. **Agronomy** 2023;13:1540. https://doi.org/10.3390/agronomy13061540.

BATTISTI R, SENTELHAS PC, BOOTE KJ. Inter-comparison of performance of soybean crop simulation models and their ensemble in southern Brazil. **Field Crops Research** 2017;200:28–37. https://doi.org/10.1016/j.fcr.2016.10.004.

BEZDEK, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. **Springer US.** https://doi.org/10.1007/978-1-4757-0450-1

BLASCH, G.L.Z.; TAYLOR, J.A. Multi-temporal yield pattern analysis method to derive yield zones in agricultural production systems. **Precision Agric** 21, 1263–1290, 2020.

BOENECKE, E. et al. Determining the within-field yield variability from seasonally changing soil conditions. **Precision Agriculture**, v. 19, n. 4, p. 750-769, 2018.

BOTTEGA, E. L., SAFANELLI, J. L., ZERAATPISHEH, M., AMADO, T. J. C., QUEIROZ, D. M. DE, & OLIVEIRA, Z. B. DE. (2022). Site-Specific Management Zones Delineation Based on Apparent Soil Electrical Conductivity in Two Contrasting Fields of Southern Brazil. In **Agronomy** (Vol. 12, Issue 6, p. 1390). MDPI AG. https://doi.org/10.3390/agronomy12061390

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., & STONE, C. J. (2017**). Classification And Regression Trees. Routledge**. https://doi.org/10.1201/9781315139470

BREUNIG, F. M., GALVÃO, L. S., DALAGNOL, R., DAUVE, C. E., PARRAGA, A., SANTI, A. L., DELLA FLORA, D. P., & CHEN, S. (2020). Delineation of management zones in agricultural fields using cover–crop biomass estimates from PlanetScope data. **In International Journal of Applied Earth Observation and Geoinformation** (Vol. 85, p. 102004). Elsevier BV. https://doi.org/10.1016/j.jag.2019.102004

BURDETT, H., WELLEN, C. Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. **Precision Agric** 23, 1553–1574 (2022). https://doi.org/10.1007/s11119-022-09897-0

BUTTAFUOCO, G., CASTRIGNANÒ, A., CUCCI, G. ET AL. Geostatistical modelling of within-field soil and yield variability for management zones delineation: a case study in a durum wheat field. **Precision Agric** 18, 37–58 (2017). https://doi.org/10.1007/s11119-016-9462-9

CONAB. **National Supply Company. 2023.** Available at:<http://www.conab.gov.br/ /Boletim_de_Monitoramento_Vera_o_Fevereiro_2023_final.pdf>.

CORWIN, D. L., & LESCH, S. M. (2005). Apparent soil electrical conductivity measurements in agriculture. In **Computers and Electronics in Agriculture** (Vol. 46, Issues 1–3, pp. 11–43). Elsevier BV. https://doi.org/10.1016/j.compag.2004.10.005

DA SILVA, E. E., ROJO BAIO, F. H., RIBEIRO TEODORO, L. P., DA SILVA JUNIOR, C. A., BORGES, R. S., & TEODORO, P. E. (2020). UAV-multispectral and vegetation indices in soybean grain yield prediction based on in situ observation. In **Remote Sensing Applications: Society and Environment** (Vol. 18, p. 100318). Elsevier BV. https://doi.org/10.1016/j.rsase.2020.100318

DORIGO et al. The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements, **Hydrology and Earth System Sciences,** 15 (2011), pp. 1675-1698, 2011. Available in:< 10.5194/hess-15-1675-2011>.

EJAZ, M., ABBAS, G., FATIMA, Z. et al. Modelling Climate Uncertainty and Adaptations for Soybean-Based Cropping System. **Int. J. Plant Prod**. 16, 235–250 (2022). https://doi.org/10.1007/s42106-022-00190-8

FERREIRA, C. J. B., TORMENA, C. A., SEVERIANO, E. D. C., ZOTARELLI, L., & BETIOLI JÚNIOR, E. (2020). Soil compaction influences soil physical quality and soybean yield under long-term no-tillage. In **Archives of Agronomy and Soil Science** (Vol. 67, Issue 3, pp. 383–396). Informa UK Limited. https://doi.org/10.1080/03650340.2020.1733535

FRANCHINI, J. C. et al. Root growth of soybean cultivars under different water availability conditions. **Semina. Ciências Agrárias (online)**, v. 38, p. 715-724, 2017.

GAVIOLI, A., DE SOUZA, E. G., BAZZI, C. L., SCHENATTO, K., & BETZEK, N. M. (2019). Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods. In **Biosystems Engineering** (Vol. 181, pp. 86–102). Elsevier BV. https://doi.org/10.1016/j.biosystemseng.2019.02.019

GAVIOLI, A.; DE SOUZA, E.G.; BAZZI, C.L.; SCHENATTO, K.; BETZEK, N.M. Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods. **Biosystems Engineering**, vol. 181, p. 86–102, maio 2019.

GIMÉNEZ L, PAREDES P, PEREIRA LS. Water Use and Yield of Soybean under Various Irrigation Regimes and Severe Water Stress. Application of AquaCrop and SIMDualKc Models. **Water** 2017;9:393. https://doi.org/10.3390/w9060393.

HILLEL, D**. Environmental soil physics**. New York, Academic Press, 1998. 771p.

HUANG, J. et al. Unraveling location-specific and time-dependent interactions between soil water content and environmental factors in cropped sandy soils using Sentinel-1 and moisture probes. **Journal of Hydrology**, v. 575, p. 780-793, 2019.

IN B. LIU, M. MA, & J. CHANG. Information Computing and Applications. (2012). , **Lecture Notes in Computer Science.** Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34062-8

JENA, R. K., BANDYOPADHYAY, S., PRADHAN, U. K., MOHARANA, P. C., KUMAR, N., SHARMA, G. K., ROY, P. D., GHOSH, D., RAY, P., PADUA, S., RAMACHANDRAN, S., DAS, B., SINGH, S. K., RAY, S. K., ALSUHAIBANI, A. M., GABER, A., & HOSSAIN, A. (2022). Geospatial Modelling for Delineation of Crop Management Zones Using Local Terrain Attributes and Soil Properties. In **Remote Sensing** (Vol. 14, Issue 9, p. 2101). MDPI AG. https://doi.org/10.3390/rs14092101

JIANG, G., GRAFTON, M., PEARSON, D., BRETHERTON, M., & HOLMES, A. (2021). Predicting spatiotemporal yield variability to aid arable precision agriculture in New Zealand: a case study of maize-grain crop production in the Waikato region. In New Zealand **Journal of Crop and Horticultural Science** (Vol. 49, Issue 1, pp. 41–62). Informa UK Limited. https://doi.org/10.1080/01140671.2020.1865413

JURY, W.A.; HORTON, C.N. **Soil Physics**, (6th ed.), John Wiley, Hoboken, NJ, USA, 2004.

KINOSHITA, R., ROSSITER, D. & VAN ES, H. Spatio-temporal analysis of yield and weather data for defining site-specific crop management zones. **Precision Agric** 22, 1952–1972 (2021). https://doi.org/10.1007/s11119-021-09820-z

KIRKHAM, M.B. Field Capacity, Wilting Point, Available Water, and the Non-Limiting Water Range. **Principles of Soil and Plant Water Relations**. [S. l.]: Elsevier, 2005. p. 101–115.

KROSS, A., ZNOJ, E., CALLEGARI, D., KAUR, G., SUNOHARA, M., LAPEN, D. R., & MCNAIRN, H. (2020). Using Artificial Neural Networks and Remotely Sensed Data to Evaluate the Relative Importance of Variables for Prediction of Within-Field Corn and Soybean Yields. In **Remote Sensing** (Vol. 12, Issue 14, p. 2230). MDPI AG. https://doi.org/10.3390/rs12142230

LEROUX, C., JONES, H., TAYLOR, J., CLENET, A., & TISSEYRE, B. (2018). A zone-based approach for processing and interpreting variability in multi-temporal yield data sets. **In Computers and Electronics in Agriculture** (Vol. 148, pp. 299–308). Elsevier BV. https://doi.org/10.1016/j.compag.2018.03.029

LEUKEL, J., ZIMPEL, T., & STUMPE, C. (2023). Machine learning technology for early prediction of grain yield at the field scale: A systematic review. In **Computers and Electronics in Agriculture** (Vol. 207, p. 107721). Elsevier BV. https://doi.org/10.1016/j.compag.2023.107721

LI, X. et al. Spatial variability of soil water content and related factors across the Hexi Corridor of China. **Journal of Arid Land**, v. 11, n. 1, p. 123-134, 2018.

LI, Y., SHI, Z., WU, C., LI, H., & LI, F. (2008). Determination of potential management zones from soil electrical conductivity, yield and crop data. In **Journal of Zhejiang University** SCIENCE B (Vol. 9, Issue 1, pp. 68–76). Zhejiang University Press. https://doi.org/10.1631/jzus.b071379

LOH, W. (2011). Classification and regression trees. In **WIREs Data Mining and Knowledge Discovery** (Vol. 1, Issue 1, pp. 14–23). Wiley. https://doi.org/10.1002/widm.8

MADARÁSZ, B., JAKAB, G., SZALAI, Z., JUHOS, K., KOTROCZÓ, Z., TÓTH, A., & LADÁNYI, M. (2021). Long-term effects of conservation tillage on soil erosion in Central Europe: A random forest-based approach. In **Soil and Tillage Research** (Vol. 209, p. 104959). Elsevier BV. https://doi.org/10.1016/j.still.2021.104959

METWALLY, M. S., SHADDAD, S. M., LIU, M., YAO, R.-J., ABDO, A. I., LI, P., JIAO, J., & CHEN, X. (2019). Soil Properties Spatial Variability and Delineation of Site-Specific Management Zones Based on Soil Fertility Using Fuzzy Clustering in a Hilly Field in Jianyang, Sichuan, China. In **Sustainability** (Vol. 11, Issue 24, p. 7084). MDPI AG. https://doi.org/10.3390/su11247084

MOLIN J.P. et al. Establishing management zones using soil electrical conductivity and other soil properties by the fuzzy clustering technique. **Scientia Agricola** 65(6):567-573.

MORAES, M. T et al. Modelagem da dinâmica da água em sistemas de preparo de um Latossolo Vermelho. **Scientia Agraria** (UFPR. IMPRESSO), v. 19, p. 142, 2018b.

MORAES, M. T. et al. Corn crop performance in an Ultisol compacted by tractor traffic. **Pesquisa Agropecuária Brasileira**, v. 53, p. 464-477, 2018a.

MORAL, F. J., TERRÓN, J. M., & SILVA, J. R. M. DA. (2010). Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. In **Soil and Tillage Research** (Vol. 106, Issue 2, pp. 335–343). Elsevier BV. https://doi.org/10.1016/j.still.2009.12.002

MORALES-SANTOS A, GARCÍA-VILA M, NOLZ R. Assessment of the impact of irrigation management on soybean yield and water yield in a subhumid environment. **Agricultural Water Management** 2023;284:108356. https://doi.org/10.1016/j.agwat.2023.108356.

MULAZZANI RP, GUBIANI PI, ZANON AJ, DRESCHER MS, SCHENATO RB, GIRARDELLO VC. Impact of soil compaction on 30-year soybean yield simulated with CROPGRO-DSSAT. **Agricultural Systems** 2022;203:103523. https://doi.org/10.1016/j.agsy.2022.103523.

PERRON, I.; CAMBOURIS, A.N.; CHOKMANI, K.; VARGAS GUTIERREZ, M.F.; ZEBARTH, B.J.; MOREAU, G.; BISWAS, A.; ADAMCHUK, V. Delineating soil management zones using a proximal soil sensing system in two commercial potato fields in New Brunswick, Canada. **Canadian Journal of Soil Science**, v. 98, nº 4, p. 724–737, 1 dez. 2018.

REICHARDT, K. et al. Dinâmica da água em solo cultivado com milho. R. Bras. Ci. Solo, 3:1-5, 1979.

REICHERT, J.M.; ALBUQUERQUE, J.A.; SOLANO J.E.P.; DA COSTA, A. Estimating water retention and availability in cultivated soils of southern Brazil. **Geoderma Regional**, v.21, p.277, jun. 2020.

ROSSATO, L.; ALVALÁ, R.C.S.; MARENGO, J.A.; ZERI, M.; CUNHA, A.P.M.; PIRES, L.B.M.; BARBOSA, H. Impact of Soil Moisture on Crop Yields over Brazilian Semiarid. **Frontiers in Environmental Science,** v. 05, n.73, p.1-16, 2017.

ROSSATO, L.; ALVALÁ, R.C.S.; TOMASELLA, J. Spatiotemporal variation of soil moisture in Brazil: Analysis of the mean conditions in the period 1971-1990. **Journal of Meteorology,** v.19, n.2, p.113-122, 2004.

SALIH HASAN, B. M., & ABDULAZEEZ, A. M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. In J**ournal of Soft Computing and Data Mining** (Vol. 2, Issue 1). Penerbit UTHM. https://doi.org/10.30880/jscdm.2021.02.01.003

SALMERÓN M, PURCELL LC. Simplifying the prediction of phenology with the DSSAT-CROPGRO-soybean model based on relative maturity group and determinacy [Internet]. Vol. 148, **Agricultural Systems.** Elsevier BV; 2016. p. 178–87. Available from: http://dx.doi.org/10.1016/j.agsy.2016.07.016

SINGH RS, SINGH KK, GOHAIN GB. Simulating crop yield using the DSSAT v4.7-CROPGRO-soyabean model with gridded weather and soil data. **Model Earth Syst Environ** 2023. https://doi.org/10.1007/s40808-023-01807-1.

USDA, NRCS. 2023. **Grain: World Markets and Trade** Avalaible in< https://www.fas.usda.gov/data/grain-world-markets-and-trade>.

VEREECKEN, H. et al. Modeling Soil Processes: Review, Key Challenges, and New Perspectives. **Vadose Zone Journal**, v. 15, n. 5, p. vzj2015.09.0131, 2016.

VIANNA, M. dos S. **Functional, structural and agrohydrological sugarcane crop modelling: towards a simulation platform for Brazilian farming systems**. 2018. Tese (Doutorado) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, São Paulo.

VIEIRA, S. R.; GREGO, C. R.; TOPP, G. C. Analyzing spatial and temporal variability of soil water content. **Bragantia**, v. 67, n. 2, p. 463–469, 2008.

WANG, Y.-P., & SHEN, Y. (2014). Identifying and characterizing yield limiting soil factors with the aid of remote sensing and data mining techniques. In **Precision Agriculture** (Vol. 16, Issue 1, pp. 99–118). Springer Science and Business Media LLC. https://doi.org/10.1007/s11119-014-9365-6

YOST, J.; HUANG, J.; HARTEMINK, A. Spatial-temporal analysis of soil water storage and deep drainage under irrigated potatoes in the Central Sands of Wisconsin, USA. **Agricultural Water Management**, v. 217, p. 226-235, 2019.

ZHANG, M. et al. Temporal and spatial variability of soil moisture based on WSN. **Mathematical and Computer Modelling**, v. 58, n. 3-4, p. 826-833, 2013.

ZHOU, G. et al. Global Pattern for the Effect of Climate and Land Cover on Water Yield. **Nature Communications**, v. 6, n. 1, 9 jan. 2015. Disponível em: <http://dx.doi.org/10.1038/ncomms6918>.

2. MACHINE LEARNING TO SUPPORT MANAGEMENT ZONES BASED ON SOIL PHYSICS AND SOYBEAN YIELD SURFACES

Abstract

Machine learning techniques are the next step to transform data into information in smart agriculture. The aim of this research was to investigate the use of historical soybean yield, electrical apparent conductivity (ECa), and altitude surfaces, to define Management Zones (MZs) using machine learning techniques. The hypothesis was that machine learning would provide greater stability in predicting MZs (Z1, Z2, Z3 and Z4) than the standard method. The study used as input surfaces of soybean yield from six crop seasons, altitude and ECa of a 10-hectare field. MZs were generated using the Fuzzy C-means employing two methods: Normalizing high-resolution data (N-M) and applying principal components analysis to choose inputs layers (PC-M). To evaluate the quality of clusters to N-M and PC-M were obtained the Fuzziness Performance Index, Normalized Classification Entropy, Fuzzy Silhouette Index, Gap index, Davies-Bouldin test and Pseudo F. The soil particle size and soil water content (9 points per hectares) were used to validate the MZs by decision trees. The results showed that PC-M was suitable for presenting greater stability in predicting clusters, which converged into either three (FPI = 0.035, NCE = 0.017, and FSI = 0.775) or four MZs (GI=1.092, DB = 0.762, and Pseudo F = 7893). The classification tree method proved to be suitable for validating the generated MZs, mainly due to particle size, with accuracy greater than 50% (test) and 88% (training). The use of machine learning techniques allowed validation of MZs generated with high-resolution data.

Keywords: PCA, Fuzzy Clustering, Classification Tree, FC, PWP, Precision Agriculture

2.1. Introduction

Management operations in intensive grain production systems have been carried out without considering the spatial variability of soil and topographic factors, limiting the use of available technologies with economic and environmental consequences due to the pressure on natural resources (PERRON et al., 2018). On the other hand, the growing demand for food requires the rational use of water and soil resources. An alternative for rationalization and decision-making for the management of agricultural areas is to obtain data from sensors in agricultural machinery, which can be useful in the definition of management zones (MZs). In this context, high-resolution data such as crop yield information, electrical conductivity and altitude can be used to improve MZs.

The treatment of variable data obtained in high resolution scales can be done by means of geostatistical methods, which will subsidize the interpolation of data to define the MZs. The performance of kriging before the generation of MZs is extremely important to obtain more precise boundaries between regions with distinct characteristics. Other authors use the technique of interpolation of agricultural big data by kriging to generate surfaces with high resolution (SCUDIERO et al., 2018; ALI et al., 2022). Complementary approaches for the reduction and pre-selection of variables that contribute effectively to the definition of MZs tend to be increasingly used. These include data normalization and correlation techniques and more complex machine learning techniques (BLASCH et al., 2020). Methods for generating MZs with a focus on machine learning (ML) for use in agriculture have made considerable progress in recent decades (CHLINGARYAN et al., 2018; BLASCH et al. 2020). The precision agriculture makes it possible when combined with solutions that will allow better estimation and decision-making on factors related to production and the environment, which will also contribute to crop yield management. Blasch et al. (2020), created the method MYPA (Multi-temporal Yield Pattern Analysis) for evaluating historical yield data used in MZs.

The design of the MZs occurs from the identification of layers of input data, aiming to delimit regions that can be classified statistically homogeneous. The Fuzzy C-means (FCM) method is widely used to delineate MZs using very distinct input variables. Many authors recommend this approach for evaluating agricultural data (FRIDGEN et al., 2004; ALI et al., 2021), which is prominent in the identification of MZs. FCM is an effective multivariate approach and can be recommended with a focus on directing soil sampling and management of specific local regions. In this context, Ali et al. (2021) considers the method capable of generating an adequate number of MZs. The identification of the related variables in the definition of the patterns within and between the management units supports the rational use of available resources.

Soil is the main factor that contributes to the definition of MZs, especially some soil factors whose spatial variability is temporally stable. Von Hebel et al. (2018) mapped the apparent electrical conductivity of soil (ECa) to investigate possible patterns regarding soil texture and depth variability. Jiang et al. (2020), used yield map data from 4 season crops to delineate MZs and understand the variations that occurred as a function of soil particle size, subsidizing the prescription of variable rate application of seeds and fertilizers. Perron et al. (2018) stated that the use of MZs can increase the profitability and quality of production, as well as reduce the environmental impact on agricultural areas as excessive use and losses of inputs. The definition of MZs based on yield data is an option for identifying patterns that are consistent across the area, even with different crops and climatic conditions.

Among the hypotheses of this research, the use of PCA as a layer selection technique that can guarantee a better prediction quality of MZs compared to just normalizing and clustering the input layers. Due to the slowness of the evaluation and collection of soil physical factors, the use of big data from yield maps combined with soil sensing can help define MZs. Finally, the classification tree technique combined with soil physical factors can help to validate this hypothesis. Tittonel et al. (2008) argue that the classification tree statistical technique is useful for predictions related to heterogeneous crop and soil management, as well as being a more accurate approach to know the limitations related to agricultural production. Thus, the objective of this research was to investigate the use of high-resolution big data and ML techniques to generate MZs and validate their relationship with soil physical factors in an agricultural field in southern Brazil.

2.2. Materials and Methods

2.2.1. Study site

The study was carried out in an agricultural unit of soybean production located in the Northwest region of the state of Paraná (Fig. 1), southern Brazil (23° 24 'S, 52° 15'W;492 m.a.s.l). The production unit (10 hectares) presented a time series of soybean yield data from 6 seasons (2016-2021). The soil is classified as a eutrophic Red Oxisol (LVe) (USDA, 2014); however, some distinct patterns were found presenting possible transitions to Entisol (RL). The soil particle size ranges from clay-sandy to very clayey with variations between 30 - 85% of clay (Fig. 1).

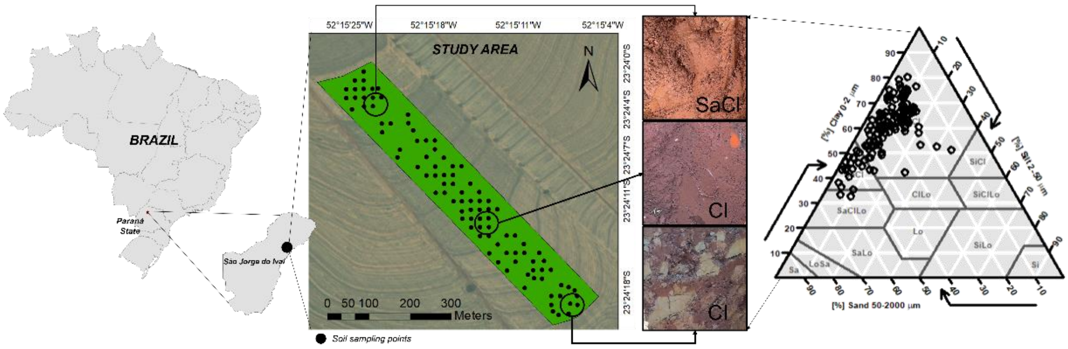


Figure 1. Spatial location of region, field, and soil characteristics (A); distribution of soil layer (0–0.60 m) particle size fractions in São Jorge do Ivaí, Paraná state, south of Brazil (Soil Survey Staff, USDA, 2014) (B).

The soybean crop cycle occurs between October and March, followed by the corn second season, which is seeded from March to July (Fig. 2). The region's climate is classified as humid subtropical with hot summers (ALVARES et al., 2013), concentrated rainfall in summer, and low occurrence of frosts (Fig. 2). The average annual temperature is 21°C, with a maximum of 27°C and minimum of 16°C. The average annual rainfall is around 1600 mm.

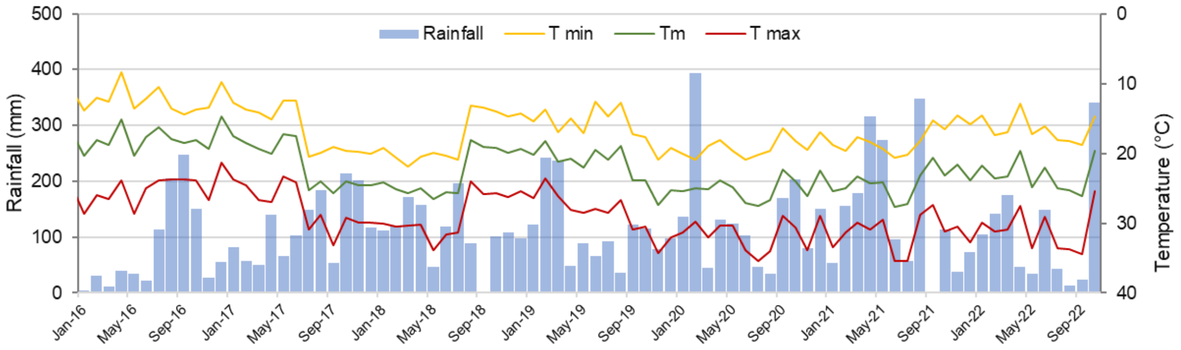


Figure 2. Variations of the daily rainfall (mm); daily mean, maximum and minimum air temperature (Tm, T max, T min - °C).

The soil chemical factors were determined for the characterization of the area (TEIXEIRA et al., 2017). A descriptive analysis of the data is presented in Table 1.

Table 1. Descriptive statistics of soil chemical factors (0-20 cm) in study site (N=20).

Descriptive Statistics	[1]pH	[2]OM	[3]P	[4]SSO$_2^{-4}$	[5]H + Al	[6]K	[7]Ca^{2+}	[8]Mg^{2+}	[9]EB	[10]CTC	[11]V
	(CaCl^2)	g dm^{-3}	mg dm^{-3}				mmol$_c$ dm^{-3}				%
Mean	5	25	34	6	37	5	47	11	62	99	61
Standard deviation	1	5	14	7	6	2	19	4	24	25	9
Min	4	14	21	0	26	0	22	5	27	63	43
Max	6	30	78	26	45	9	85	19	107	146	74
CV (%)	10	19	41	131	15	51	41	34	38	25	15

[1]Active Acidity in water; 2Organic Matter; 3Phosphorus; 4Oxide sulfur; 5Hidrogen + Aluminium; 6Potassium; 7Calcium; 8Magnesium; 9Echangeable bases; 10Cation Exchange Capability; 11Base saturation.

2.2.2. Management Zones (MZs)

Determination of MZs were based on historical maps of soybean yield, electrical conductivity (ECa) and altitude maps. To determine the MZs, yield maps from 6 harvests were used. Due to the low quality of yield data during the 2015/16 crop season, data from the 2016/17, 2017/18, 2018/19, 2019/20, and 2020/21 harvests were considered (Table 2). The yield monitor used was an AgLeader® PF 3000 model, coupled to a combine with a nominal power of 186 kW and a grain storage capacity of 7,050® L. The land altitude was model obtained from the combine's positioning data and used as an input layer for MZs estimation.

ECa mapping was performed on September 13, 2020, after the second crop yield. The Veris 3100 equipment, which measures ECa through six electrodes arranged to allow the simultaneous characterization of two layers (0 - 30 cm and 0 - 90 cm), was used. Data were collected continuously throughout the plot, with a spacing of 15 meters between strides and a frequency of 1 Hz. Sampling points were georeferenced using a Garmin GNSS (Global Navigation Satellite System) receiver with an accuracy of 3 to 5 m.

Table 2. Descriptive statistics of the historical of yield maps (t ha^{-1}), soil electrical conductivity (uS m^{-1}) and altitude (m) in study site.

Variable	N	Mean	Standard Deviation	Median	Min.	Max.	Amplitude	Skewness	Kurtosis
Soy$_{2016/17}$ (t ha^{-1})	5097	3,85	0,53	3,98	0,93	5,50	4,57	-1,50	5,90
Soy $_{2017/18}$ (t ha^{-1})	5097	2,90	0,35	2,95	1,00	4,44	3,44	-0.76	4,00
Soy $_{2018/19}$ (t ha^{-1})	5097	2,79	0,46	2,89	0,55	3,73	3,18	-1,30	4,70
Soy $_{2019/20}$ (t ha^{-1})	5097	2,18	0,34	2,24	0,73	3,00	2,27	-1,50	6,00
Soy $_{2020/21}$ (t ha^{-1})	5097	1,81	0,33	1,83	0,19	3,01	2,82	-0,58	4,30
EC$_{0-30}$ (uS m^{-1})	5097	5,02	1,72	5,10	1,98	8,60	6,62	-0,064	1,80
EC$_{0-90}$ (uS m^{-1})	5097	3,98	3,33	3,16	0,00	15,85	15,85	2,30	8,10
Altitude (m)	5097	474	13,12	475	446	495	49,00	-0,25	1,90

The yield and ECa data were filtered using the Mapfilter® software (SPEKKEN, ANSELMI and MOLIN, 2013) and normalized by the amplitude of variation method according to equation 1.

$$z = \frac{X - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where X corresponds to the sampling point, min (x) is the smallest value of the sample and max (x) is the largest value of the sample. The variation between the values will be from 0 to 1.

Vector layers were produced by interpolating each variable with the use of block kriging, using the Vesper 1.62® software, configured for use of local variogram, generating a surface with a resolution of 25 m^2 (5 m x 5 m). The next step of the process consisted of clustering data to generation of management zones (MZs), by the Fuzzy C-means (FCM) method. To compare the effects of different processing methods on clustering, two approaches were employed. The first method (N-M) utilized normalized data for clustering, while the second method (PC-M) Principal Component Analysis (PCA) + normalization on standardized input data. The suitability of variables for generating principal components (PC) was determined using a sampling adequacy measures (MSA) test (KAISER, 1974). The multivariate analysis was carried out using the MYPA method to transform the vectorized variables into a multiband raster containing the information of each variable in each pixel. Once the PCA was performed, spatialization of the PCs was done. Linear regression analysis was carried out between each PC and the normalized variables, and the coefficient of determination (R2) was obtained. The FCM algorithm uses the Mahalanobis distance to determine the similarity between the variables.

To perform the clustering of yield maps, ECa and altitude by PC-M and N-M, the Management Zone Analyst (MZA) software was used. To determine the optimal number of clusters, various combinations ranging from 2 to 6 were tested. The quality and the number of groups was defined and evaluated from 6 different tests: Fuzziness Performance Index, Normalized Classification Entropy, Fuzzy Silhouette Index, Gap index, Davies-Bouldin test and Pseudo F explained in the sequence by two approaches used (N-M and PC-M). The FPI assesses the degree of separation between fuzzy partitions and classified data during clustering. The NCE measures the optimal quality of clusters (FRIDGEN et al., 2004). The FSI, measures how well are grouped the data within the cluster. The GI is a statistical test that compares the similarity between the set of observations of a single cluster with the set of observations from the difference in each cluster (SENTELLE et al., 2007). The DBi is determined by the relationship between the sum of the level of dispersion in a cluster in relation to the dispersion between clusters (PHAKIRA et al., 2004). The PF (CALINSKI and HARABAS, 1974) aims to validate the clustering, being given by the relationship of the variance between clusters with the variance within the cluster. The higher PF value, the higher the cluster density and the greater the dissimilarity between clusters.

2.2.3. Sampling and analysis of soil factors

To characterize the physical properties of the soil, soil samples were collected at 9 points per hectare in a non-uniform grid, across three depths (0-20 cm, 20-40 cm and 40-60 cm) (Fig. 1). The sampling was carried out before soybean sowing in the 2021/22 crop season. The samples were collected using an auger, and two types of samples were collected: those without preserved structure, to measure soil particle size, and those with preserved structure, to measure bulk density (ϱd), following the method described by Grossman and Reinsch (2002).

The -10 kPa matrix potential was set to determine the upper limit of available soil water or field capacity (FC) using a voltage table similar to that described by Ball and Hunter (1988). Water contents at -1 and -6 kPa were also determined as indicator parameters of soil water retention easily available to plants. The soil water content in the permanent willing point (PWP) or the equivalent of the water content in the matrix potential of -1,500 kPa through the WP4-T psychrometer equipment was determined (MELO FILHO et al., 2015).

Vector layers were produced by interpolating each variable with the use of block kriging, using the Vesper 1.62® software with a resolution of 25 m² (5 m x 5 m). To evaluate the relationship between soil parameters and multitemporal soybean and ECa yield data, buffers with a radius of 15 m around the sampling points were determined and the averages of the data contained in the spatial regions were extracted through a geographic information system using the ArcMap software version 10.5 (ESRI, Redlands, CA, USA, 2011).

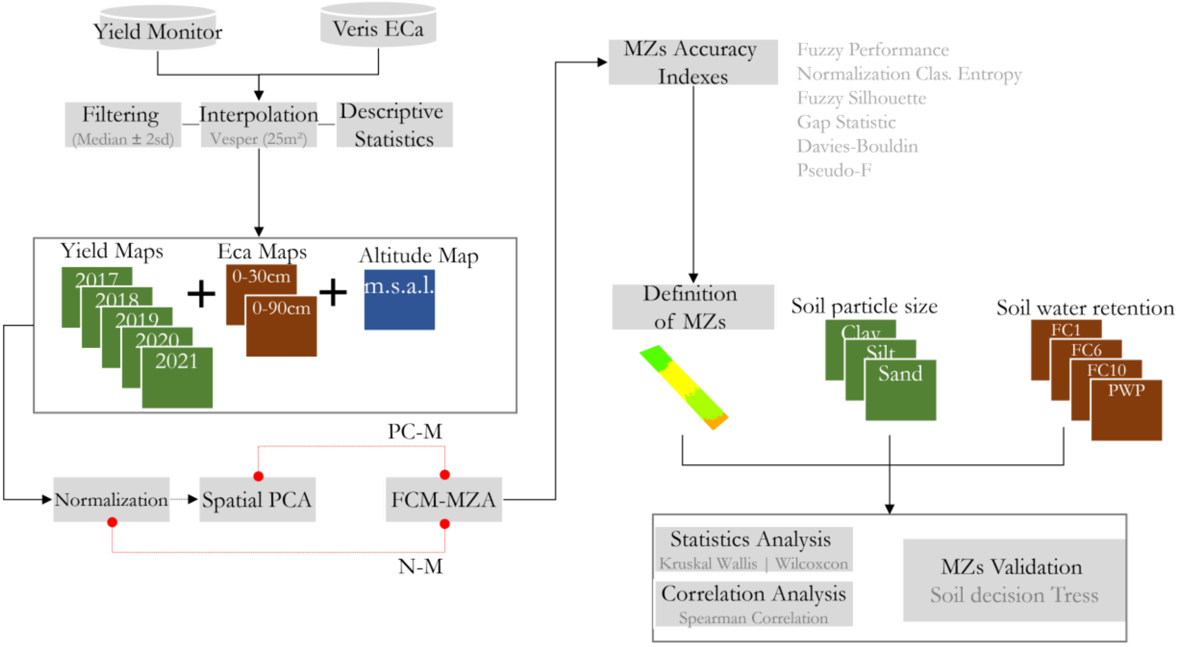


Figure 3. Flowchart of steps to generate management zones (MZs) by Normal and Principal Component Method (N-M; PC-M) and the correlation between soil factors.

Due to the ease of obtaining historical yield maps and soil sensing layers, soil texture and water retention data were used to evaluate the patterns of MZs trough classification tree analysis. The CART (Classification and Regression Tree) method aims to predict the response of a categorical variable as a function of predictor variables from machine learning supervised through rule-based recursive partitioning (BREIMAN et al., 1984). The test was performed using the Recursive Partitioning and Regression Trees (Rpart) package on the Rstudio (R Core Team, 2020). The data were partitioned into 80% for training and 20% for testing, determining the accuracy rate of the predictive model for training and testing. The predicted data were georeferenced and compared to the initial establishment of the MZs.

2.2.4. Statistical analysis

The Kruskal-Wallis test ($p<0.05$) was used to analyze the variances of the input layers (Table 4) in both approaches (N-M and PC-M). After verifying heterogeneity, the Mann-Whitney Wilcoxon test ($p<0.05$) was performed to determine the significant difference between the samples. The choice of the optimal approach (N-M and PC-M) and the appropriate number of groups was based on cluster quality indices and significant differences in input data layers. With the definition of MZs, each group was renamed based on the significant differences of the multitemporal data of yield, ECa and altitude.

The Spearman correlation test (ϱ) was performed to verify the association between the input layers of the MZs with the parameters of soil particle size, bulk density and soil water retention. To verify the interactions of soil parameters with layers and MZs, Kruskal-Wallis and Mann-Whitney Wilcoxon tests were performed. All statistical tests were performed using the R software.

2.3. Results

2.3.1. Definition, delimitation, and quality of MZs

The high-resolution ECa, altitude, and yield layers used to delineate the MZs are presented in Fig. 4. A pattern in the historical average of soybean yield (Table 2) was observed as a function of 53% reduction of yield between 2017/18 (3.85 ± 0.53 t ha^{-1}) and 2021/22 (1.81 ± 0.33 t ha^{-1}), which is related 38% reduction in the accumulated rainfall series (Fig. 2, R^2 = 82%) for same period. The reduction in accumulated rainfall occurred mainly in the 2020/21 and 2021/22. The reduction in rainfall was more intense in the first 3 months of 2020/21 and 2021/22 intensifying the relation to the historical average of rainfall for region (Sep = -86%, Oct = -64% and Nov = -60%). Certainly, the reduction in rainfall coincided with the reproductive phase of soybean and consequently provided a reduction in yield. Observing the spatial variability of yield maps (Fig. 4 D-H), it was possible verify patterns mainly at the south of field. The ECa$_{0\text{-}30cm}$ showed higher values in south of field (6-8 uS m^{-1}) and lower values at north (2-4 mS m^{-1}). A similar pattern was observed for ECa$_{0\text{-}90cm}$, except in a specific region with ECa < 4 uS m^{-1}. The region north of field has the highest altitude (484-495 m) while the southern portion with 446 and 460 m. The input variables presented in Fig. 4 were normalized and used in the design of MZs by the direct method (N-M, Fig. 3), preceding the FCM technique to multivariate method (PC-M).

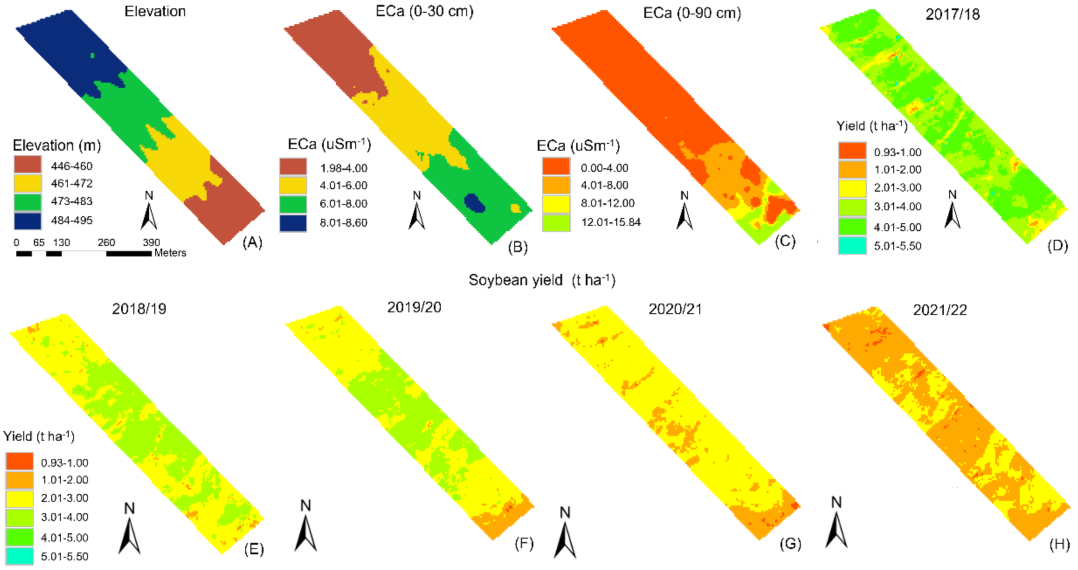


Figure 4. Spatial-temporal variability of altitude (A), electrical conductivity apparent: ECa$_{0\text{-}30cm}$ (B), ECa$_{0\text{-}90cm}$ (C) and multiple year soybean yield (D-H) to Fuzzy C-means analysis by normal method (N-M) for MZs=1,2,3 and 4.

The multivariate technique for selecting input layers, referred to as PC-M, was performed following the method described in Blasch et al. (2020) (Fig. 4). Initially, several data quality tests were performed. The Bartlett's sphericity test ($p<0.05$) was performed, which rejected the hypothesis of equal variances. Additionally, the MSA test was also performed, and all variables presented results higher than 60% of sample commonality. The result of the KMO test was 66% of mean sample commonality. The values of MSA and KMO must be greater than 50% to commonality to be suitable for the analysis (Kaiser, 1974). The percentage of variance explained by the PC1 + PC2 layers was 62% (Fig. 5), which were used as inputs in the FCM analysis. As the variance explained 62% of the data, the technique was used after verifying the quality of the clusters by the accuracy indices (Table 3).

The surfaces that most contributed to PC1 were those related to soil. Altitude contributed with 30%, $ECa_{0\text{-}30cm}$ = 30% and $ECa_{0\text{-}90cm}$ = 25%. For PC2, the surfaces that presented greatest contribution were those related to soybean yield, mainly in the agricultural years 2019 (30%), 2018 (24%) and 2020 (22%). It is possible to visualize the same pattern of yield variation in the southern and northern regions (Fig. 4) with PC2 (Fig. 5). The same behavior observed to PC1 (Fig. 5) with ECa and altitude (Fig. 4). The analysis was able to capture in two multivariate layers significant variations in the MZs.

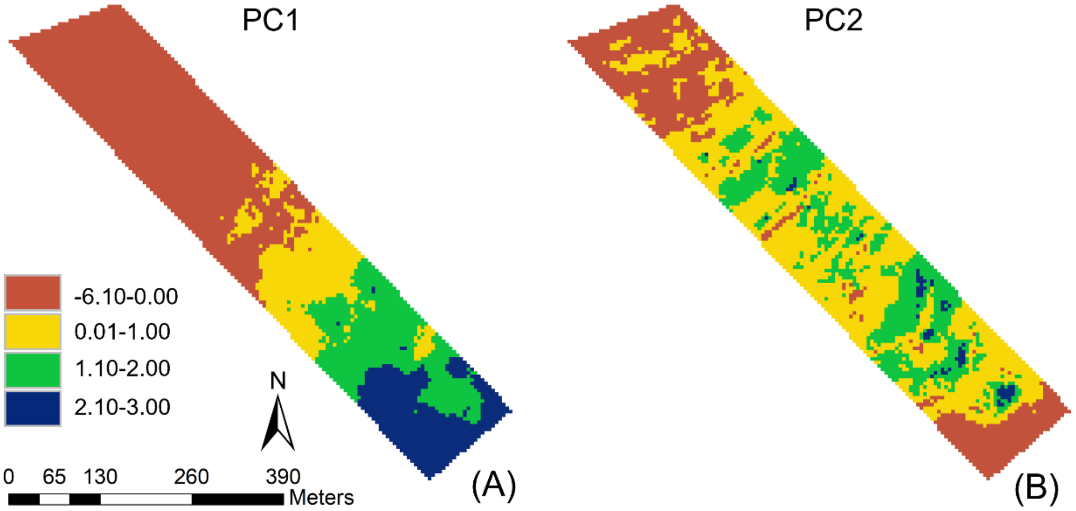


Figure 5. Principal Component 1 (PC1, A) and Principal Component 2 (PC2, B) to Fuzzy C-means analysis (FCM) by method PC-M.

In order to understand the quality of the clusters (k=2 to 6) as a function of the N-M and PC-M, accuracy indices were measured (Table 3). It was possible to observe a greater variation between the results of the quality indices and the number of clusters to N-M (k= 2 with NCE = 0.109, DBI = 0.723; k=3 ns; k=4 with FSI = 0.805; k=5 with GI = 1.501, PF = 10053 and k=6 with FPI = 0.165). On the other hand, PC-M presented a pattern in which the three quality indices were significant for the clustering of k=3 (FPI=0.035, NCE = 0.017 and FSI = 0.775) and three for k=4 (GI = 1.092, DBi = 0.762 and PF = 7893). Due to the greater stability to PC-M, it is possible to affirm that this multivariate method can be an alternative in relation to the standard method, obtaining clearer trends in relation to the number of ideal clusters. In order to obtain standards in relation to the number of clusters, k=4 was chosen as ideal since also in the N-M method it presented a significant quality index compared to k=3.

Table 3. Fuzzy C-Means quality indexes to defining the adequate number of yield zones for Normal (N-M) and PCA (PC-M) methods.

k	*Normal Method | N-M					
	FPI	NCE	FSI	GI	DBi	PF
2	0.306	**0.109**	0.799	1.3	**0.723**	9180
3	0.265	0.139	0.758	1.3	0.776	8887
4	0.203	0.125	**0.805**	1.24	0.823	8440
5	0.179	0.121	0.782	**1.501**	0.816	**10053**
6	**0.165**	0.117	0.76	1.475	0.794	9421

k	*Principal Component Method | PC-M					
	FPI	NCE	FSI	GI	DBi	PF
2	0.09	0.032	0.633	0.579	1.003	3511
3	**0.035**	**0.017**	**0.775**	0.617	1.031	3721
4	0.038	0.021	0.745	**1.092**	**0.762**	**7893**
5	0.05	0.031	0.676	1.002	0.883	7297
6	0.054	0.033	0.716	1.037	0.848	7776

The Fig. 6 represents the results of the FCM grouping considering the PC-M treatment. Each MZs was classified according to statistically significant average patterns ($p<0.05$) of layers between MZs and between seasons. Z1 is characterized by a region of significant low yield (~1.6 t ha^{-1}) and significant low ECa (~4 uS m^{-1}) ; Z2 is characterized by a significant high-yield region (~3.2 t ha^{-1}) but no significant ECa; Z3, a region of significant moderate ECa (~5 uS m^{-1}) but no significant differences in yield; and Z4, was characterized being a region of significant moderate yield (~2.4 t ha^{-1}), with significant high ECa (~10 uS m^{-1}). Z4 have incidence of soil source material not yet weathered (gravel) in the depth layer.

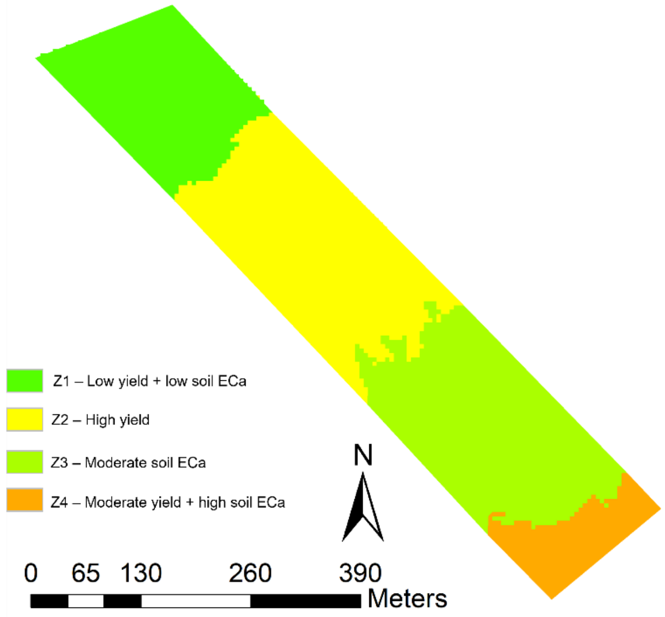

🟩 Z1 – Low yield + low soil ECa
🟨 Z2 – High yield
🟩 Z3 – Moderate soil ECa
🟧 Z4 – Moderate yield + high soil ECa
N
0 65 130 260 390
Meters

Figure 6. Management zones (MZs) derived from multiple year soybean maps (2017-2021), ECa (0-30 cm), ECa (0-90 cm) and altitude by principal component analysis method (PC-M).

2.3.2. Interactions between Management Zones and soil factors

Fig. 7 shows the correlations between the soil variables used in the PC-M method for the establishment of MZs as soil particle size, water and ϱs. Regarding the ECa surfaces, it was possible to observe a strong interaction with most soil factors. $ECa_{0\text{-}30cm}$ and $ECa_{0\text{-}90cm}$ were positively correlated with Clay (ϱ = 0.7 and 0.4, respectively) and negatively with Sand (ϱ = -0.8 and -0.4, respectively). ECa also correlated significantly with soil water retention factors. In both layers, there was a positive interaction between ECa and FC10kPa (ϱ = 0.6 for ECa0-30cm and ϱ = 0.4 for ECa0-90cm), with positive interactions also with PWP for the $ECa_{0\text{-}30cm}$. It was possible to observe correlations between the granulometry and the annual yields in the 2017 season (Sand: ϱ = - 0.4; Silt: ϱ = 0.5), 2018 season (Clay: ϱ = 0.4; Sand: ϱ = - 0.4) and 2019 season (Silt: ϱ = -0.5). The crop season of 2021/22 presented a different behavior from the other crops, presenting correlation with most of the variables (Clay: ϱ = 0.6; Sand: ϱ = -0.6) and with field capacity and soil water indicators easily available (FCs: ϱ = 0.4).

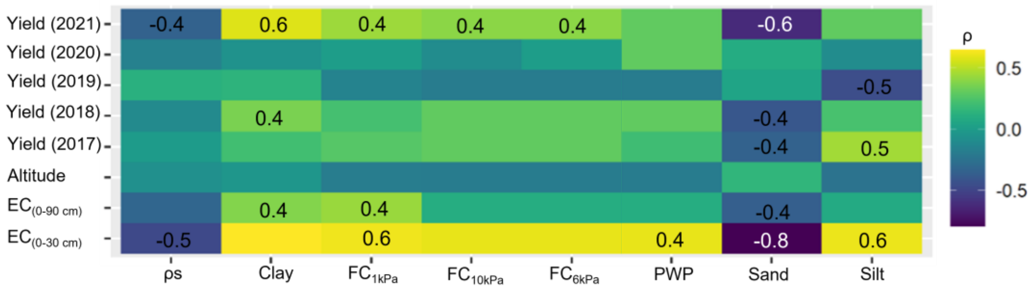


Figure 7. Heatmap of Spearman correlation (ϱ) between physical soil factors (Clay, Sand, Silt, FC, PWP, ϱs and SRP) and layers (ECa, historical soybean yields and altitude) used to delineate Management zones (MZs).

The results between differences in soil particle size (Clay, Silt and Sand) and soil water retention (FC and PWP) in soil depths of 0-20 cm, 20-40 cm and 40-60 cm and between MZs are presented in Table 4. The Mann-Whitney Wilcoxon test ($p<0.05$) was used to verify if the variable differed significantly within each of the layers of each MZs and, subsequently, if there were significant differences between MZs regardless of the soil layer. Regarding clay content, significant differences are observed between all layers within each MZ, significantly highlighting the 0-20 cm and 20-40 cm of soil depth. Increments in clay contents in the soil profile were also observed, independent of MZ. Z1 was a region that presented on average 44% less clay than Z3, which was characterized as the most clayey. The same pattern was observed for the silt contents, independent of the layer and MZ, which were mostly significant. The highest levels of silt were observed in the Z4, a region of transition to Entisol, being a region of shallow soil depth and high stoniness index. Regarding the sand contents, Z1 presented on average 53% more sand compared to Z3 and Z2 can be characterized as a transition region between Z1 and Z3. Z4, on the other hand, presents the presence of gravel (considerable level of stoniness) in some parts of the plot, with a variability distinct from the other regions.

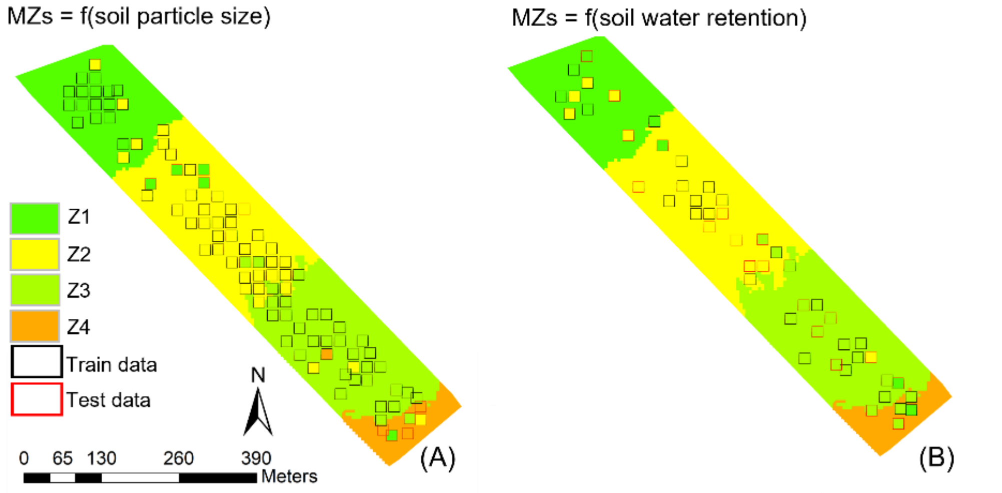
The differences between soil water retention within and between MZs was also distinct. It was possible to observe within each region, a reduction of FC with the reduction of the matrix potential, regardless of the depth. In relation to the depth layers, it was possible to observe an increase in FC from the superficial to the deeper layer, associated with the increase of clay in the soil. The Z1 presented FC on average 39% lower compared to Z4, which was the region with the highest FC levels. In Z2 it was also possible to observe a FC transition behavior between Z1 and Z3, with an average FC reduction of 20% compared to Z3 and a 10% increase in relation to Z1. Similar behavior was observed between the regions in relation to PWP. This research evidenced differences on MZs since distinctions between soil factors and variations of $ECa_{0\text{-}30cm}$.

Table 4. Soil variables interactions between and within Management zones (MZs) with soil particle size (Clay, Silt and Sand) and soil water retention (FC and PWP).

MZ	Soil depth	Clay (%)		Silt (%)		Sand (%)		FC_{1kPa}		FC_{6kPa}		FC_{10kPa}		PWP (%)	
Z1	00-20 cm	39.53	****	4.54	***	55.93	****	15.49	****	14.36	****	14.33	****	11.82	****
	20-40 cm	46.84	****	4.43	****	48.74	****	16.66	ns	15.50	ns	15.13	****	12.73	****
	40-60 cm	50.39	****	5.15	****	44.46	****	18.02	****	16.96	****	16.46	****	14.93	****
	Mean	45.59	****	4.71	****	49.71	****	16.72	****	15.60	****	15.31	****	13.16	****
Z2	00-20 cm	52.18	****	6.75	ns	41.07	****	17.91	****	16.85	****	16.56	****	13.26	****
	20-40 cm	57.33	ns	7.70	****	34.97	**	18.80	ns	17.87	**	17.51	***	14.39	****
	40-60 cm	62.80	****	5.92	****	31.28	****	19.32	****	18.27	****	17.81	****	14.61	****
	Mean	57.44	ns	6.79	****	35.77	****	18.68	****	17.66	****	17.29	****	14.09	****
Z3	00-20 cm	60.16	****	12.19	****	27.64	****	21.75	****	20.47	****	20.15	****	15.43	****
	20-40 cm	65.40	ns	12.53	****	22.07	****	22.56	ns	21.51	****	21.51	****	17.48	****
	40-60 cm	71.21	****	8.14	****	20.65	****	22.99	****	21.57	****	21.07	****	17.93	****
	Mean	65.59	****	10.96	****	23.46	****	22.43	****	21.19	****	20.91	****	16.95	****
Z4	00-20 cm	53.17	****	15.43	**	31.40	****	22.80	****	20.56	****	20.39	****	16.34	****
	20-40 cm	50.45	****	23.87	****	25.69	****	22.98	****	21.91	*	21.44	***	15.48	****
	40-60 cm	66.94	****	8.85	****	24.21	****	23.81	****	22.42	****	21.82	****	18.83	****
	Mean	56.85	ns	16.05	****	27.10	****	23.20	****	21.63	****	21.22	****	16.88	****

*Are significantly by Mann-Whitney Wilcoxon test ($p<0.05$).

The results of classification tree method for predicting MZs as a function of soil particle size and water retention are presented in Fig. 8. After cross-validation, the results obtained were georeferenced and spatialized. When evaluating the quality of training and test prediction, it was possible to observe high accuracy in the estimation of MZs by soil variables. The predictive validation of MZs as a function of soil particle size showed lower errors in Z1 (train = 97% and test = 88%) and Z3 (train = 90 and test = 94%). For the factors of soil water retention, the forecast in Z3 showed greater accuracy (train = 97% and test = 79%, as well as Z2 (train = 76% and test = 100%). For classification tree method, it was possible to observe the prediction of MZs generated from the PCA originated by ECa surfaces, altitude and historical yield as a function of the soil factors, presenting a direct relationship with the high-resolution information and with the soil factors measured.



Z1	Z2	Z3	Z4	Z1	Z2	Z3	Z4
Train: 97%	Train: 88%	Train: 90%	Train: 88%	Train: 76%	Train: 76%	Train: 97%	Train: -
Test: 88%	Test: 95%	Test: 94%	Test: 50%	Test: 67%	Test: 100%	Test: 79%	Test: -

Figure 8. Result of classification tree prediction and cross-validation from MZs (based on multiple years soybean maps and ECa) as a function of soil particle size factors (A) and soil water (B).

2.4. Discussion

2.4.1. The threshold between uniform management and spatial-temporal variability of agricultural fields

The results showed that the field management disregarding the spatial variability of physical soil factors and soybean crop yield is not an adequate practice, because the spatio-temporal variability of soybean yield maps and ECa confirm the need to perform a management considering these variations. Several authors have verified the importance of conducting management practices from the design of MZs, either by yield maps or by soil factors (SCUDIERO et al., 2018; KOUTSOS et al., 2021; ALI et al., 2022). The use of yield machinery data to know the variability in agricultural fields is a strategy that can help farmers in decision making do the low cost of obtaining, requiring only statistical and agronomic knowledge for the generation of MZs and their interpretation (KOUTSOS et al., 2021).

The use of not supervised machine learning algorithm by PC technique proved to be adequate for the detection and selection of variables and design of MZs. The PC-M technique was related to variations in soybean yield layers and apparent electrical conductivity, reaffirming the importance of having yield layers, soil factors and altitude to support the soil physical patterns found. PC1 was strongly influenced by the measured ECa and altitude, with was in accordance of Scudiero et al. (2018) in humid subtropical climate. This fact shows the importance of ECa as high-resolution information to find patterns in agricultural areas. ECa is influenced by soil texture and water, where the indirect use of this information can help detect soil variations in MZs. In this research, PC2 was strongly influenced by variation of the yield maps, which shows that the technique can quickly ensure the distinction between the patterns that will delineate MZs.

The PC-M approach ensured greater stability among MZs, evidenced by the accuracy indexes, with high agreement to 3 or 4 regions as the ideal number of MZs. This is due to the convergence of classification of each index in PC-M, where the division quality, similarity, and dispersion within and between MZs were evaluated. These results are similar to those of Ouazaa et al. (2022) also used PC-M to design MZs. Regarding the N-M, there was greater variability between the selection of the number of MZs suitable to quality indexes. Although N-M has not been considered the best method in this study, it is commonly used by other authors (SCHENATTO et al., 2017; BLASCH et al., 2020).

The four MZs identified in the area coincided from the farmer observations due distinct behaviors in certain regions. During soil sampling, it was discovered that the northern region had sandy soil, which was confirmed by low ECa and low yield patterns. These findings are in line with Ali et al. (2021), who reported that sandy regions showed high MZs instability in terms of yield. The influence of dry and wet years on Z1 was noted, with low yield patterns due to grouping, but there were wet years with moderate yields compared to other regions. Z2 was classified as a high-yield region, regardless of the variation in ECa. Despite being clayey, Z4 had different gravel formations, historically making agricultural operations difficult as seeding quality and emergence of plants. The region was classified as having a moderate yield, despite the high ECa in the top layer. As noted by Cordoba et al. (2016), regions with shallow soils may have high ECa values due to the clay horizon being closer to the topsoil layer, which may explain the unique characteristics found in Z4, as it is in a region classified as RL. These findings highlight the importance of understanding the physical characteristics of the soil and their influence on MZs to make informed decisions for successful crop management.

The delimitation of MZs is the basis to assist in decision making in areas of low yield, in the variation in seed population, reduction of downtime machine, reduction in the use of inputs, that is, intelligent rationalization of production costs, benefiting the greatest economic return to farmer (KOUTSOS, et al., 2021). Although the present research did not directly evaluate chemical and biological factors of the soil, the targeting of soil sampling from MZs helps in reducing large-scale collections and cost with analyses, making it possible to find the main characteristics inherent to the agricultural field and the evaluation of soil quality (OUAZAA et al., 2022).

2.4.2. Knowing the soil factors helps support decision making and directed management

In our study area, quantifying soil physical factors was a relevant step to increase the accuracy of the use of MZs. However, the uniform management of agricultural areas is a common practice, where the application of fertilizers and seeds is the same throughout the field, not considering the changes that are found in relation to the soil physical properties.

The increases in ECa levels (0-30cm and 0-90cm) were mainly influenced by the increase in clay contents and reduction in sand contents. Due soil particles sizes, sandy regions have low ECa, silty regions have medium ECa and clay regions, high ECa (KWEON, 2012). Our research also found a relation between MZs design, ECa variability and soil particle size as a function of soil layer. On superficial soil depth there was a higher intensity in ECa variation compared to deep layers, also seen in MZs outlined in the studies of Scudieiro et al. (2018) and Lajili et al. (2021).

Soil water retention was a factor that was positively correlated with ECa, especially in relation to FC and PWP, presenting variability between layers. From the design of the MZs it was possible to evidence such behavior, a result of soil water retention variability being higher at the surface, which makes the $ECa_{0\text{-}30cm}$ more sensitive to changes related to water and the deep ECa more stable. Lajili et al. (2021), also showed a similar trend for a humid climate region. The mapping of ECa as an input layer for delineation of MZs was the basis of the evaluation of soil variability at field and support in the correlation with temporal layers of yields. Allied to the fact that it is necessary to consider hydraulic variables, due to their heterogeneous influence on soils (OUAZAA et al., 2020).

The correlation between yield maps in wet and dry years with the soil factors showed the temporal variability of yield as a function of water deficit in field. The season 2020/21 and 2021/22 were considered drier years, directly influenced by changes in patterns with soil factors. In dry years, water stress is the main limiting factor for the crop, especially for sandy regions which is negatively correlated with yield, also evidenced in Scudiero et al. (2018). The behavior observed for wet years (2017, 2018 and 2019) was only due to soil particle size, also seen in Ali et al. (2021).

The season 2020/21 was considered a year of transition, where the reduction of water availability began, which impacted the 2021/22. Our research was conducted in a humid subtropical climate, which may further intensify the variability of crop yield. The production system is located in a dryland area, depending exclusively on rainfall, which causes soil factors to directly impact water retention and availability for crops. The 2020/21 crop season was influenced by most soil factors, possibly due to the reduction in rainfall volume in the last two seasons.

2.4.3. Machine learning tools as support for agricultural data analysis

In this study, machine learning tools have been shown to be useful for field data assessments. PC-M, an unsupervised machine learning technique, helped in the selection and adjustments of high-resolution spatial-temporal soil and yield layers to predict MZs. This technique is commonly used for soil factors (OUAZZA et al., 2020). However, the differential of this research was the use of historical yield maps of five crop seasons (BLASCH et al., 2020; KOUTSOS et al., 2021).

The classification tree technique proved to be a useful tool in understanding the interactions between soil particle size, water and their impact on the variability of different MZs, enabling informed decision-making in agricultural production (TITTONEL et al., 2008). However, few studies have utilized the present approach to predict MZs based on soil properties at the field scale. The use of predictive models based on soil properties can help to know more about the yield variation caused by these factors (CORDOBA et al., 2016). The soil particle size was the factor that presented the highest accuracy to differentiate the aspects related to the design of MZs, possibly because it is a stable factor over time. Although the region is predominantly clayey to medium clayey soils, sand content was the predictor of greater weight in the classification, being evidenced in other studies that did not use classification tree techniques (SCUDIERO et al., 2018).

Regarding water retention, the model showed good levels of accuracy. The correlation between soil water retention and sand contents contributes to the process of soil water drainage, seen in other studies using classification trees (AMORIM et al., 2022). FC is a factor dependent on soil structural organization and is related to the size and disposition of soil pores (REICHERT el al., 2020), which may explain FC importance as a predictor in this study, based on the instability of variation between regions. Through cross-validation it was possible to verify distinctions between and within regions in relation to machine learning predictions, evidencing the importance of high-resolution data to help in the evaluation of variation in agricultural fields.

The characterization of agricultural fields without prior data hinders decision-making and rationalization of resources (OUAZZA et al., 2022). Therefore, the present research is an ally in the use of high-resolution data as support for characterization of agricultural fields and design MZs, in the direction of determination of soil factors as well as in the rational use of inputs. Although the design techniques of MZs with temporal yield layers are extremely relevant, from this research we suggest that more studies be done to understand the climatic variability of dry and wet years on the existing characteristics in MZs, as they can directly influence the seasonality of agricultural production due to water deficit. Another alternative is the use of high-resolution space-time satellite images to replace gaps in the historical series or failures of yield maps.

2.5. Conclusions

The results of this study demonstrated the effectiveness of using high-resolution yield data and soil apparent electrical conductivity to identify management zones and analyze their relationships with soil water retention and particle size. By employing machine learning techniques, the specialized principal component analysis method showed superiority in predicting MZs based on principal components compared to conventional methods. The classification tree analysis revealed the important correlation between MZs, soil particle size, and soil water retention. Future studies that utilize historical spatial-temporal yield data can further enhance our understanding of the observed patterns. Our findings also highlight the crucial role of apparent electrical conductivity surfaces in designing MZs. To gain insights into the temporal variability of MZs due to water deficit, future research should investigate the factors responsible for such variability during wet and dry years. This study contributes to the development of more precise and effective agricultural management practices, ultimately leading to improved yields and sustainability.

Acknowledgment

We thank the farmer in southern Brazil who kindly authorized the use of the data and the area in favor of agronomic research and development. We would like to thank the Agronomy consultancy that kindly referred us to their client to carry out this research. Contributing to R&D is feeding the world.

References

ALI, ABID; MARTELLI, ROBERTA; SCUDIERO, ELIA; LUPIA, FLAVIO; FALSONE, GLORIA; RONDELLI, VALDA; BARBANTI, LORENZO. Soil and climate factors drive spatio-temporal variability of arable crop yields under uniform management in Northern Italy. **Archives of Agronomy and Soil Science.** [S. l.]: Informa UK Limited, 2 Aug. 2021. DOI 10.1080/03650340.2021.1958320. Available at: http://dx.doi.org/10.1080/03650340.2021.1958320.

ALI, ABID; RONDELLI, VALDA; MARTELLI, ROBERTA; FALSONE, GLORIA; LUPIA, FLAVIO; BARBANTI, LORENZO. Management Zones Delineation through Clustering Techniques Based on Soils Traits, NDVI Data, and Multiple Year Crop Yields. **Agriculture**. [S. l.]: MDPI AG, 5 Feb. 2022. DOI 10.3390/agriculture12020231. Available at: http://dx.doi.org/10.3390/agriculture12020231.

ALVARES, CLAYTON ALCARDE; STAPE, JOSÉ LUIZ; SENTELHAS, PAULO CESAR; DE MORAES GONÇALVES, JOSÉ LEONARDO; SPAROVEK, GERD. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, vol. 22, nº 6, p. 711–728, 1 dez. 2013.

AMORIM, RICARDO SANTOS SILVA; ALBUQUERQUE, JACKSON ADRIANO; COUTO, EDUARDO GUIMARÃES; KUNZ, MAURÍCIO; RODRIGUES, MIRIAM FERNANDA; SILVA, LUCAS DE CASTRO MOREIRA DA; REICHERT, JOSÉ MIGUEL. Water retention and availability in Brazilian Cerrado (neotropical savanna) soils under agricultural use: Pedotransfer functions and decision trees. **Soil and Tillage Research**. [S. l.]: Elsevier BV, Oct. 2022. DOI 10.1016/j.still.2022.105485. Available at: http://dx.doi.org/10.1016/j.still.2022.105485.

BALL, B.C., HUNTER, R., 1988. The determination of water release characteristics of soil cores at low suctions. **Geoderma** 43, 195–212.

BREIMAN, LEO; FRIEDMAN, JEROME H.; OLSHEN, RICHARD A.; STONE, CHARLES J**. Classification And Regression Trees.** [S. l.]: Routledge, 2017. DOI 10.1201/9781315139470. Available at: http://dx.doi.org/10.1201/9781315139470.

BLASCH, G., LI, Z. & TAYLOR, J.A. Multi-temporal yield pattern analysis method for deriving yield zones in crop production systems. **Precision Agric** 21, 1263–1290 (2020). https://doi.org/10.1007/s11119-020-09719-1

CAINSKI, T. AND HARABASZ, J. (1974), "A Dendrite Method for Cluster Analysis," ´ **Communications in Statistics—Theory and Methods**, 3, 1–27.

CHLINGARYAN, A., SUKKARIEH, S., & WHELAN, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. In **Computers and Electronics in Agriculture** (Vol. 151, pp. 61–69). Elsevier BV. https://doi.org/10.1016/j.compag.2018.05.012

CORDOBA, MARIANO A.; BRUNO, CECILIA I.; COSTA, JOSÉ L.; PERALTA, NAHUEL R.; BALZARINI, MÓNICA G. Protocol for multivariate homogeneous zone delineation in precision agriculture**. Biosystems Engineering.** [S. l.]: Elsevier BV, Mar. 2016. DOI 10.1016/j.biosystemseng.2015.12.008. Available at: http://dx.doi.org/10.1016/j.biosystemseng.2015.12.008.

ESRI 2011. **ArcGIS Desktop**: Release 10. Redlands, CA: Environmental Systems Research Institute.

GROSSMAN, R.B., REINSCH, T.G., 2002. Bulk Density and Linear Extensibility, in: Of, A.S., Agronomy, S.S.S. of A. (Eds.), **Methods of Soil Analysis**. Part. 4. Madison, Lincoln, Nebraska, pp. 201–228. https://doi.org/10.2136/sssabookser5.4.c9

JIANG, GUOPENG; GRAFTON, MILES; PEARSON, DIANE; BRETHERTON, MIKE; HOLMES, ALLISTER. Predicting spatiotemporal yield variability to aid arable precision agriculture in New Zealand: a case study of maize-grain crop production in the Waikato region. **New Zealand Journal of Crop and Horticultural Science**. [S. l.]: Informa UK Limited, 2 Jan. 2021. DOI 10.1080/01140671.2020.1865413. Available at: http://dx.doi.org/10.1080/01140671.2020.1865413.

KAISER, H. F. (1974). An index of factorial simplicity. **Psychometrika,** 39(1), 31-36.

KOUTSOS, THOMAS M.; MENEXES, GEORGIOS C.; MAMOLOS, ANDREAS P. The Use of Crop Yield Autocorrelation Data as a Sustainable Approach to Adjust Agronomic Inputs. **Sustainability.** [S. l.]: MDPI AG, 22 Feb. 2021. DOI 10.3390/su13042362. Available at: http://dx.doi.org/10.3390/su13042362.

KWEON, GIYOUNG. Delineation of site-specific productivity zones using soil properties and topographic attributes with a fuzzy logic system. **Biosystems Engineering.** [S. l.]: Elsevier BV, Aug. 2012. DOI 10.1016/j.biosystemseng.2012.04.009. Available at: http://dx.doi.org/10.1016/j.biosystemseng.2012.04.009.

LAJILI, ABDELKARIM; CAMBOURIS, ATHYNA N.; CHOKMANI, KAREM; DUCHEMIN, MARC; PERRON, ISABELLE; ZEBARTH, BERNIE J.; BISWAS, ASIM; ADAMCHUK, VIACHESLAV I. Analysis of Four Delineation Methods to Identify Potential Management Zones in a Commercial Potato Field in Eastern Canada. **Agronomy.** [S. l.]: MDPI AG, 26 Feb. 2021. DOI 10.3390/agronomy11030432. Available at: http://dx.doi.org/10.3390/agronomy11030432.

MELO FILHO, JOSÉ F. DE; SACRAMENTO, JOSÉ A. A. S. DO; CONCEIÇÃO, BRUNA P. SWater retention curve prepared by the psychrometer method for use in determining the "S" index of soil physical quality. **Agricultural Engineering**. [S. l.]: FapUNIFESP (SciELO), Oct. 2015. DOI 10.1590/1809-4430-eng.agric.v35n5p959-966/2015. Available at: http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v35n5p959-966/2015.

OUAZAA, SOFIANE; JARAMILLO-BARRIOS, CAMILO IGNACIO; CHAALI, NESRINE; AMAYA, YEISON MAURICIO QUEVEDO; CARVAJAL, JOHN EDINSON CALDERON; RAMOS, OMAR MONTENEGRO. Towards site specific management zones delineation in rotational cropping system: Application of multivariate spatial clustering model based on soil properties. **Geoderma Regional.** [S. l.]: Elsevier BV, Sep. 2022. DOI 10.1016/j.geodrs.2022.e00564. Available at: http://dx.doi.org/10.1016/j.geodrs.2022.e00564.

PERRON, I.; CAMBOURIS, A.N.; CHOKMANI, K.; VARGAS GUTIERREZ, M.F.; ZEBARTH, B.J.; MOREAU, G.; BISWAS, A.; ADAMCHUK, V. Delineating soil management zones using a proximal soil sensing system in two commercial potato fields in New Brunswick, Canada. (Newton Lupwayi, ed.). **Canadian Journal of Soil Science.** [S. l.]: Canadian Science Publishing, 1 Dec. 2018. DOI 10.1139/cjss-2018-0063. Available at: http://dx.doi.org/10.1139/cjss-2018-0063.

R CORE TEAM (2020). R: **A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

REICHERT, JOSÉ MIGUEL; ALBUQUERQUE, JACKSON ADRIANO; SOLANO PERAZA, JOSÉ EFRAIN; DA COSTA, ANDRÉ. Estimating water retention and availability in cultivated soils of southern Brazil. **Geoderma Regional.** [S. l.]: Elsevier BV, Jun. 2020. DOI 10.1016/j.geodrs.2020.e00277. Available at: http://dx.doi.org/10.1016/j.geodrs.2020.e00277.

SENTELLE, C., LUN HONG, S., GEORGIOPOULOS, M. A Fuzzy Gap Statistic for Fuzzy C-means. **11th Internacional Conference. Artificial Intelligence and Soft Computing.** ISB:978-0-88986-693-5

SCUDIERO, ELIA; TEATINI, PIETRO; MANOLI, GABRIELE; BRAGA, FEDERICA; SKAGGS, TODD; MORARI, FRANCESCO. Workflow to Establish Time-Specific Zones in Precision Agriculture by Spatiotemporal Integration of Plant and Soil Sensing Data. **Agronomy.** [S. l.]: MDPI AG, 7 Nov. 2018. DOI 10.3390/agronomy8110253. Available at: http://dx.doi.org/10.3390/agronomy8110253.

SOIL SURVEY STAFF, 2014 **Soil Survey Staff Keys to Soil Taxonomy** (12th), USDA (2014), pp. 1-410

SCHENATTO, KELYN; DE SOUZA, EDUARDO GODOY; BAZZI, CLAUDIO LEONES; GAVIOLI, ALAN; BETZEK, NELSON MIGUEL; BENEDUZZI, HUMBERTO MARTINS. Normalization of data for delineating management zones. **Computers and Electronics in Agriculture.** [S. l.]: Elsevier BV, Dec. 2017. DOI 10.1016/j.compag.2017.10.017. Available at: http://dx.doi.org/10.1016/j.compag.2017.10.017.

TEIXEIRA, P.C.; DONAGEMMA, G.K.; FONTANA, A.; TEIXEIRA, W.G. **Handbook of Methods Soil Analysis.** 3. ed. rev. e ampl. – Brasília, DF : Embrapa, 2017.

TITTONELL, P; SHEPHERD, K; VANLAUWE, B; GILLER, K. Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya—An application of classification and regression tree analysis. **Agriculture, Ecosystems &amp; Environment**. [S. l.]: Elsevier BV, Jan. 2008. DOI 10.1016/j.agee.2007.05.005. Available at: http://dx.doi.org/10.1016/j.agee.2007.05.005.

VON HEBEL, CHRISTIAN; MATVEEVA, MARIA; VERWEIJ, ELIZABETH; RADEMSKE, PATRICK; KAUFMANN, MANUELA SARAH; BROGI, COSIMO; VEREECKEN, HARRY; RASCHER, UWE; VAN DER KRUK, JAN. Understanding Soil and Plant Interaction by Combining Ground-Based Quantitative Electromagnetic Induction and Airborne Hyperspectral Data. Geophysical Research Letters. [S. l.]: **American Geophysical Union (AGU)**, 8 Aug. 2018. DOI 10.1029/2018gl078658. Available at: http://dx.doi.org/10.1029/2018GL078658.

3. MULTI-YEAR SIMULATION OF SOYBEAN YIELD FROM THE DIGITAL MAPPING OF CROPS AND SOIL WATER IN MANAGEMENT ZONES

Abstract

Predicting the yield of annual crops is a promising approach to increasing agricultural efficiency per unit area. The aim of this study was to estimate soybean yield in three management zones based on the spatial and temporal soil water balance and growth of the vegetation. Deficits and surpluses in the sequential water balance were estimated using surface texture, bulk density, organic matter fractions, and climate data, carrying out high-resolution mapping (25 m²) of the water available in the profile (RMSE ≈ 4 mm; R2 ≈ 68%). The vegetation indexes (NIR and NDVI) were obtained from satellite images with a spatial resolution of 3 m and temporal resolution of one month. Two modelling techniques (Multiple Linear Regression - MLR and Random Forest - RF) were evaluated for predicting soybean yield (2020/21) in three management zones (Z1, Z2 and Z3). Three temporal arrangements were considered, namely, the monthly predictors of water and vegetation for one, two or three earlier seasons. The quality of the models was assessed by accuracy index (RMSE, MAE, and R2), using maps and observed yield data. From the results, it was found that the RF method was more accurate, being used to predict yield in Z1, Z2 and Z3 for a rainy year (2018/19), a dry year (2019/20), and from observed data (2021/22). A 19% reduction in MAE and 17% reduction in RMSE, with a 64% increase in R2, were seen using RF compared to MLR. Including predictors from the three earlier seasons showed the greatest accuracy in Z1, Z2 and Z3 (MAE<0.21 kg.ha^{-1}, RMSE<0.29 kg.ha^{-1}, and R2>35%). The use of NIR (MAE≈0.19 kg.ha^{-1}, RMSE≈0.25 kg.ha^{-1} and R2≈45%) instead of NDVI (MAE≈0.20 kg.ha^{-1}, RMSE≈0.27 kg.ha^{-1} and R2≈36%) ensured greater reliability in the prediction model. The use of RF to predict soybean yield based on management zones proved to be applicable to precision agriculture, ensuring quick, early information for localized management in the field, replacing the conventional approach of homogeneous management.

Keywords: Available Soil Water, Remote sensing, Digital soil mapping, Temporal variability

3.1. Introduction

In the search for smarter farming, and with the transformation of data into information, understanding the use of remote sensing techniques, digital mapping and artificial intelligence is essential (SAGAN et al., 2021; SOUZA et al., 2022). In this respect, precision agriculture can help the producer decide on the rational management of agricultural practices for annual crops, especially in regions with similar characteristics. The use of management zones (MZs) consists in a grouping technique for precision agriculture that highlights heterogeneous characteristics between regions, and similarities within the same region (SCHWALBERT et al., 2018; SCUDIERO et al, 2018). Knowing the patterns in MZs can ensure rational decision-making by the farmer, aiming at saving resources by predicting the yield of commercial crops (BREUNIG et al., 2020; ALI et al., 2022). Modelling the production environments can help support digital decision agriculture by reducing the use of inputs, localized planning, and the conservation of environmental resources. Based on this need, good model development depends on intrinsic information of the soil, vegetation, and climate (MURUHANANTHAM et al., 2022).

Available Soil Water (SAW) is one of the factors that influence crop yield under local conditions. The SAW is obtained from the relationship between the water content at the potential corresponding to field capacity (FC) and the permanent wilting point of the crop (PWP) (VEREECKEN et al., 2010), and is calculated by laboratory analyses that are generally time-consuming. One solution is digital mapping of the SAW by modelling such primary soil data as textural fractions (AMIRIAM-CHAKAN et al., 2019; CAMPOS et al., 2021), organic matter and bulk density (CUEFF et al., 2021). Determining soil water storage together with climate data can help estimate the regional or local water balance (FUZZO et al., 2019; FERINA et al., 2021; AMIRI et al., 2022). The climate data can be estimated by remote sensing (TORSONI et al., 2023) using satellite images, such as those from the MERRA-2 satellite with a spatial resolution of 55 km (GMAO, 2015).

With crop growth and development, spatial and temporal monitoring is carried by means of indices, such as the NDVI (MERCANTE et al., 2010; ZENG et al., 2016; ALI et al., 2022), obtained through orbital sensing. Various studies have shown the importance of alternatives, such as reflectance at certain wavelengths (e.g. NIR), for the same purpose (LIU et al., 2014; ALABI et al., 2022). The use of vegetation indices by monitoring high-resolution spatial images from small satellite constellations (SKAKUN et al., 2021; RAO et al., 2021) is one way of improving the prediction quality of agroecosystem models, especially at the local level, such as management zones (MZs).

The use of bigdata makes it possible to generate new agroecosystem models, at a local level, which when supported by machine learning techniques, can estimate the growth, development and/or yield of annual agricultural crops as a function of soil, plant, and climate predictors (FILGUEIRAS et al., 2020; SONG et al., 2022). A few established agroecosystem models have been conventionally used in agricultural research (da SILVA et al., 2022); however, they need to be calibrated and adjusted for tropical and subtropical climate conditions (ADEBOYE et al., 2021). Alternatively, various studies have used machine learning techniques to predict, directly or indirectly, factors that influence agricultural yield (SZABÓ et al., 2019; FILGUEIRAS et al., 2020; AMORIM et al., 2022; TORSONI et al., 2023). Research should be carried out on a local scale, as one of the main challenges is increasing both the accuracy of these learning models and their applicability (MURUGANANTHAM et al., 2022).

One of the hypotheses of this study, therefore, is that the digital mapping of soil water based on a prediction model is more accurate when carried out for MZs compared to estimating an entire area. This hypothesis is based on the importance of understanding the specific characteristics of each production area, and of not considering them homogeneous. The second hypothesis is that the use of NIR reflectance surfaces to simulate spatial and temporal development in the soybean is more representative compared to surfaces based on NDVI indices. The third hypothesis is that the use of machine learning techniques to estimate the spatial and temporal variability of soybean yield by MZs is more accurate than using the multiple linear regression method due to its ability to deal with unbalanced data. The aim of this research, therefore, was to estimate multi-year soybean yield based on the digital mapping of soil water and crop development at the level of management zones.

3.2. Materials and Methods

3.2.1. Study site and Management Zones (MZ)

The study area is in the south of Brazil (23°24'S, 52°15'W, 492 m.a.s.l.), has a size of 10 hectares, and is intended for grain production (soybean and corn). The soil is classified as a Red Oxisol (LVe) in transition to an Entisol (USDA, 2014) (Fig. 1A). In outlining each MZ, layers of harvest maps from five consecutive soybean harvests (2017-2021) were used, together with the electrical conductivity at two depths, and an elevation map. The harvest maps were obtained using an AgLeader® model PF 3000 harvest monitor mounted on a combine harvester. The elevation was obtained from the positioning data of the harvester and was used as an additional layer. The apparent electrical conductivity (ECa) was determined during September 2020 using the Veris 3100 mobile sensor. This equipment measures ECa at a frequency of 1 Hz in two layers (0-30 cm and 0-90 cm) at a spacing 15 m between passes. The data were filtered using the Mapfilter® software (SPEKKEN, ANSELMI and MOLIN, 2013) and normalized using amplitude variation. The layers were estimated by interpolation using block kriging, employing the Vesper 1.62® software with the local variogram at a resolution of 25 m². The layers were then merged using principal component (PC) multivariate geospatialisation, and the significant components grouped using the Fuzzy C-means method (Fig. 1B). Four management zones were generated, classified by each significant layer as: Region Z1 of low yield and low ECa (yield: ~1.6 t ha^{-1} and ECa: ~4 uS m^{-1}); Z2 of high yield, but non-significant ECa (yield: ~3.2 t ha^{-1}); Z3 of moderate ECa and non-significant yield (ECa: ~5 uS m^{-1}), and Z4 of moderate yield and high ECa (yield: ~2.4 t ha^{-1} and ECa: ~10 uS m^{-1}). Z4 is a region that showed a high incidence of gravel and was not used for evaluation in the study (Fig. 1).

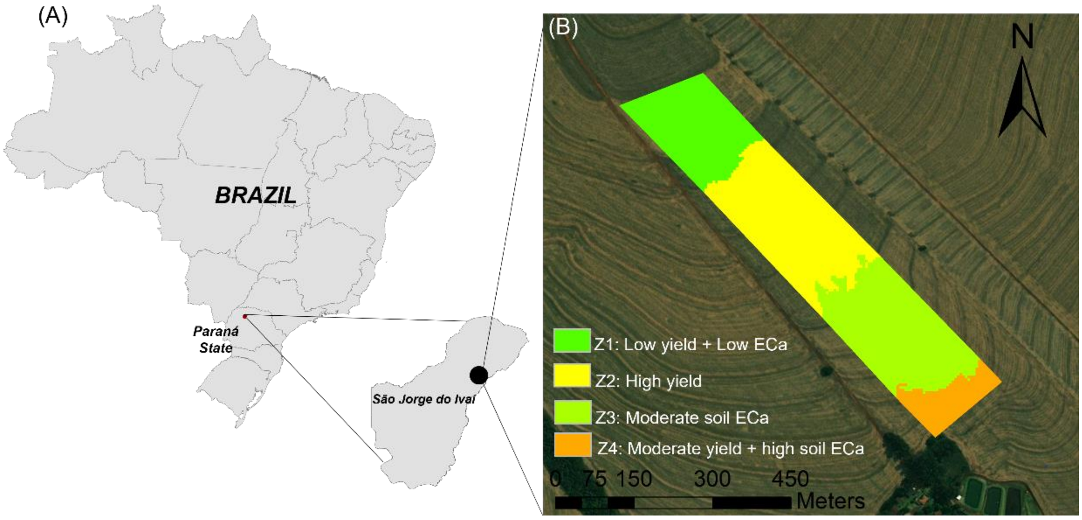


Figure 1. Geographical region of the study area in southern Brazil (A), and management zones evaluated in the field (Z1: Low yield + Low ECa, Z2: High Yield, Z3: Moderate soil ECa, and Z4: Moderate yield + high soil ECa) (B).

3.2.2. Weather and soil sampling

Soybean season occurs during the summer, usually from October to March. Climate seasonality occurred during the seasons considered in this study (2017-2021), showing variation in relation to accumulated rainfall and average temperature (Fig. 2). Due to rainfall above and average temperatures below the historical average (2000-2021), 2018 was considered a wet year. The other agricultural years (2017, 2019-2021) were considered dry years, with an accumulated rainfall below and temperatures above the historical average for most months.

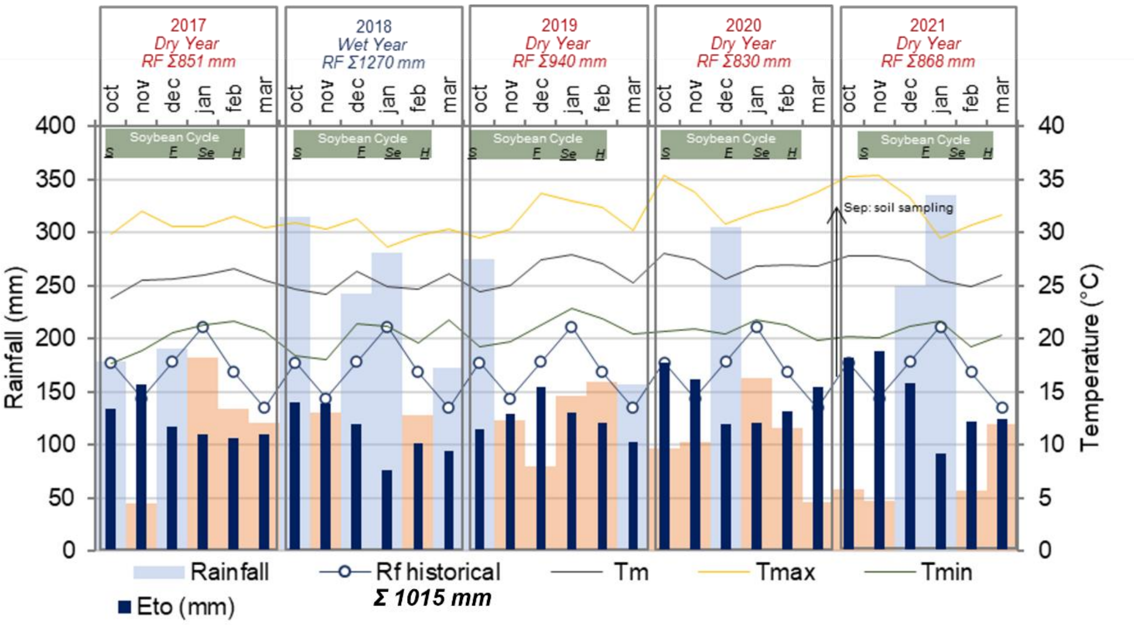


Figure 2. Seasonality of the local average, maximum and minimum air temperature (°C), accumulated rainfall (mm) and Eto for dry and wet years during the soybean crop seasons (2017, 2018, 2019, 2020 and 2021) with the historical averages for rainfall (Rf) and temperature (Tm) from 2000 to 2021.*V1 – Emergence, R2 – Flowering, R5.1 – Filling, and R8 – Maturation

Disturbed and undisturbed soil samples were collected before the start of the 2021/22 season. Undisturbed samples were collected from five points per hectare in three layers (0.00-0.20, 0.20-0.40 and 0.40-0.60 m). The bulk density (ϱs) was determined from the ratio between the dry weight and total known volume (TEIXEIRA et al., 2017). The water content at the PWP at a matrix potential of -1.5 Mpa was determined using the WP4-T Dewpoint potentiometer (Meter Group Inc.). The soil water content at a matrix potential of -10 kPa was also determined, and the upper limit of available water or field capacity (FC) was defined using a tension table similar to that described by Ball and Hunter (1988). The water content at -1 and -6 kPa was determined as indicator parameters of the retained soil water readily available to plants. Due to the shorter sampling time compared to the use of rings, disturbed samples were collected in the same three layers at a density of eight points per hectare to determine particle size (sand, silt and clay) employing sieving and chemical dispersion (ALMEIDA et al., 2012).

Using the undisturbed samples, a particle-size fractionation of the soil organic matter (SOM) was carried out and the levels of particulate organic matter (MOP) and mineral-associated organic matter (MOM) were determined as per the method for particle separation proposed by Cambardella and Elliott (1992). For this, approximately 20g of soil and 60mL of sodium hexametaphosphate solution (5g L^{-1}) were shaken for 15 hours in a horizontal shaker. The suspension was then passed through a 53µm sieve. The material retained on the sieve, considered as MOP, was oven-dried at 50°C. The material that passed through the sieve was considered the MOM fraction. To determine the C content of the MOP and MOM fractions, the individual fractions were ground in a porcelain mortar and later analysed by weight using the dry combustion method in a CN-2000® elemental analyser (Leco, St. Joseph, MI, USA). The total C content of the soil was obtained by summing the MOP and MOM fractions.

3.2.3. Procedure for predicting soybean yield by MZs based on water balance and vegetation indexes

3.2.3.1. Spatial and temporal soil-water database

A geostatistical analysis of the primary soil variables (sand, silt, clay, SOM, MOP, MOM, ϱs, FC and PWP) was carried out using the kriging technique in the ArcMap software (ESRI, 2011), with the aim of obtaining 25m² high-resolution surfaces at each depth (0.00-0.20, 0.20-0.40 and 0.40-0.60 m). Principal component analysis was carried out for each depth to understand the interactions between the soil attributes in Z1, Z2 and Z3. This analysis used the factoextra package (KASAMBARA and MUNDI, 2020) of the RStudio platform (R CORE TEAM, 2022).

The estimated soil surfaces (sand, silt, clay, SOM, MOP, MOM and ϱs) at the three depths were used to estimate SAW simulated by multiple linear regression (MLR) for Z1, Z2, Z3 and throughout the area (Fig. 3). Stepwise regression elimination was used to explain the weight of each of the predictor surfaces in the MLR (JAMES et al., 2014; BRUCE et al., 2017). The SAW measured surfaces were used to validate SAW simulated. The layer was calculated from SAW measured = [(FC-PWP) x ϱs/ 10], where z corresponds to the soil layer. SAW measured was determined and validated for the different matrix potentials (-1, -6 and -10 kPa) for each MZs and for the entire area. Cross-validation was then carried out using the K-Fold method (LI, 1987), considering 10 subdivisions.

Following validation and selection of the SAW simulated digital surface, temporal surfaces for water deficit (DEF) and water excess (EXC) were determined, resulting from calculating the sequential water balance (WB) as per the methodology proposed by Thornthwaite and Mather (1955). Climate data on rainfall and temperature were obtained from the Nasa Power platform, which provides daily data with a spatial resolution of 50 km and a temporal resolution of one day (STACKHOUSE et al. 2017). Data on the potential evapotranspiration (ETo) was calculated as per Thornthwaite and Mather (1955). The automatic method for localized determination of the water balance by Rolim et al. (1998) was adapted for use on a spatial scale. The spatial variation in the DEF and EXC surfaces occurred as a function of the spatial variation in water storage (ARM) in each of the MZs, i.e. when P-ETp > 0, then ARM = DSM of SAW. Thirty DEF and EXC surfaces were obtained for the months between November and March 2016 to 2022 that correspond to the soybean cycle.

3.2.3.2. Vegetation indexes (VIs) and boundary conditions

To account for spatial and temporal variability on crop development, the vegetation index surfaces were calculated by normalized difference (NDVI) and near-infrared reflectance (RefNIR). The scenes were extracted from sensors of the PlanetScope satellite constellation (Planet, 2020) at a spatial resolution of 3 m and a temporal resolution of a month, with the data acquired in four spectral bands (blue: 0.485 μm, green: 0.545 μm, red: 0.630 μm, and near-infrared: 0.820 μm). The criteria for scene selection were the absence of clouds and the NDVI calculation (NIR-RED)/(NIR+RED) using the ArcMap GIS. Nineteen NDVI surfaces were obtained between November and March (2016 to 2022). To determine RefNIR, the raster's were extracted and transformed, and the reflectance calculated for a total of 20 surfaces (November to March, 2016-2022) from the relationship between the digital numbers (DN) given by PlanetScope (Planet, 2020), where RefNIR = DN/10000.

In order to obtain the best representation of the prediction model, three temporal arrangements were considered. Determination accuracy was evaluated using Model 1: $\text{Yield}_{2020/21}$ = f(ΔWB Monthly + IV $\text{Monthly})_{2020/21}$, i.e. soybean yield predicted as a function of the spatial and temporal variability of the predictors for the current season. Model 2: $\text{Yield}_{2020/21}$ = f(ΔWB Monthly + IV $\text{Monthly})_{2020/21}$ + (ΔWB Monthly + IV $\text{Monthly})_{2019/20}$, i.e. yield predicted as a function of the spatial and temporal variability of the predictors for the current and previous season. Model 3: $\text{Yield}_{2020/21}$ = f(ΔWB Monthly + IV $\text{Monthly})_{2020/21}$ + (ΔWB Monthly + IV $\text{Monthly})_{2019/20}$ + (ΔWB Monthly + IV $\text{Monthly})_{2018/19}$, i.e. soybean yield predicted as a function of the spatial and temporal variability of the predictors for the current and the two previous seasons. Where ΔWB represents the variation in soil water balance from the water deficits (DEF) and surpluses (EXC), and VI the indices that represent the quality of the vegetation, using either NDVI or RefNIR.

3.2.4. Soybean yield prediction: methods and performance

The behavior of the DEF, EXC, RefNIR and NDVI prediction surfaces in each of the management zones (Z1, Z2 and Z3) were evaluated using principal component analysis. Two methods were then used to predict the yield of the soybean crop (2020/21 season) by MZ. The first method employed the Random Forest supervised learning technique (RF). The premise of the RF is to increase the accuracy of the estimate and is able to deal with the reduced dimensionality of the data (FILGUEIRAS et al., 2020). The comparative method used was MLR, based on stepwise regressive elimination of the input variables. To evaluate the performance of the model, RF and MLR were cross-validated using the K-Fold method. In choosing the best prediction model, the RMSE (Root Mean Square Error), MAE (Mean Absolute Error), adj R2 (Adjusted Coefficient of Determination), R2 (Coefficient of Determination), MSE (Mean Standard Error), AIC (Akaike Criterion), BIC (Bayesian Information Criterion) and PRE (Percent Relative Error) were also determined. (Equation 1 to 8).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{n-k}}$$
1)

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{\sum_{i=1}^{n}(yi-\bar{y})^2}$$
2)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|yi-\hat{y}i|$$
3)

$$\text{Adj R}^2 = 1\text{-}\frac{(n-1)}{(n-k)}(1-R^2)$$
4)

$$\text{RSE} = \frac{SE(y)}{\bar{y}} \times 100$$
5)

$$\text{AIC} = \text{-2log(L)+2k}$$
6)

$$\text{BIC} = \text{-2log(L)+klog(n)}$$
7)

$$\text{PER} = \frac{(yi-\hat{y}i)}{yi} \times 100$$
8)

Where n represents the number of observations in the simulation; k represents the number of estimated parameters; yi represents the observed value (harvest maps); y ̂i represents the estimated value (prediction models).

After choosing the best prediction method for each MZs (Z1, Z2 and Z3), the last method was applied to predict soybean yield for both a wet year (2018/19 season) and a dry year (2019/20 season). The model was also used to estimate soybean yield during the 2021/22 season and validated using observed yield data. To do this, soybean plants were collected manually in plots of 2 m², using an irregular mesh with a density of 2.7 points per hectare in Z1, Z2 and Z3. Fig. 3 shows each development stage of the research.

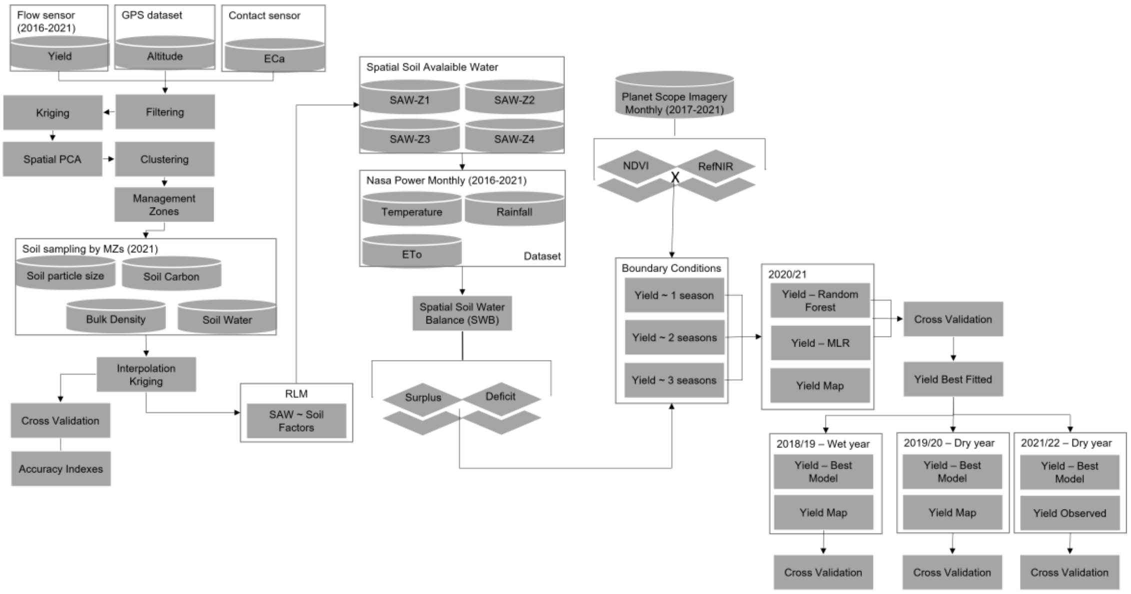


Figure 3. Analysis procedure to generate the spatial variability of soil water (Water Step) by management zone (Z1, Z2 and Z3) and selection of the temporal model of estimated yield (Yield Step) as a function of the water balance, and monthly NDVI and NIR indices.

3.3. Results

3.3.1. Multivariate physical and chemical analysis of the soil in the MZs

A biplot of the contribution of the soil variables in each of the layers to the PCs can be seen in Fig. 4. The variance explained by PC1 + PC2 was 86% for the 0 to 20 cm layer (Fig. 4A), 87% for 20 to 40 cm (Fig. 4B), 93% for 40 to 60 cm (Fig. 4C), and 83% for the entire profile (Fig. 4D). The soil variables with the greatest contribution to PC1 were related to FC and MOP. For PC2, the greatest contribution was from Sand (0-20 cm), while at the deeper layers the greatest contribution was from total organic carbon (MOS). In the soil profile (Fig. 4D), the contribution to PC1 showed the same pattern (FC + MOP) between layers, while for PC2, there were contributions from Sand, MOS and Silt. Of the regions, Z1 was considered to show the characteristics for low potential, Z2 for high potential, and Z3 for medium potential based on historical factors of yield and ECa. A group characterized by high values for MOS, MOP and ϱs (Fig. 4D) was seen in Z1, while Z2 presented a group in the central region of the biplot that can be characterized as a region of transition from Z1 to Z3. Z3 showed grouping between quadrants 1 and 4, classified with high values for FC, PWP, MOM, Clay and Silt.

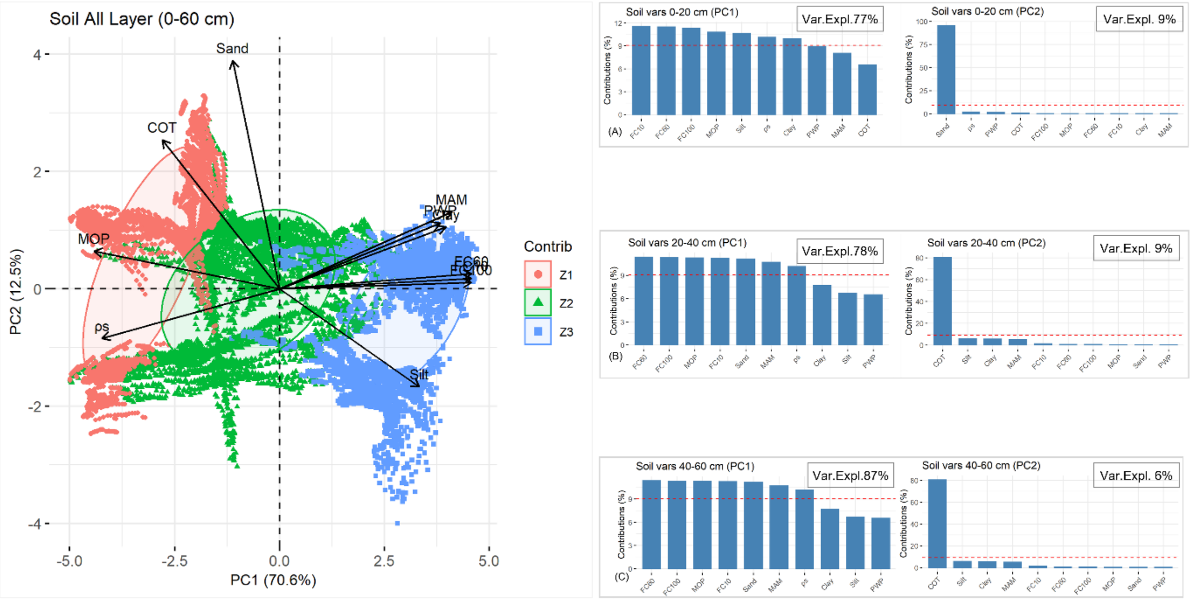


Figure 4. Principal component analysis of soil particle size (silt, sand and clay), carbon fractions (MOS, MOM and MOP) and soil density (ps) by management zone (Z1, Z2 and Z3) at different depths (0-20, 20-40 and 40-60 cm).

3.3.2. Prediction and spatial variability of the SAW between MZs

The performance of the SAW prediction for Z1, Z2, Z3, and for the entire area is shown in Table 1. In an effort to determine the variation in SAW for differing conditions of the available soil water, different potentials (-1, -6 and -10 kPa) were evaluated (OTTONI FILHO et al., 2014). According to the performance indices, the SAW in region Z1 showed greater accuracy under a matric potential of -10kPa (RMSE: 2.8 mm, R2: 0.75, MAE: 2.19 mm, adj R2: 0.75, MSE: 2.7 mm, AIC: 5981 and BIC: 6073). While in Z2 and Z3, the performance of SAW was superior at a matric potential of -6kPa (Z2 = MSE: 5.53 mm and AIC: 11956; Z3 = MSE:4.09 mm, AIC: 9898 and BIC: 10018). The percent relative error (PRE) of the SAW in region Z2 was 43% higher than in Z1 and Z3, which shows that the region can be characterized as transitioning, demonstrating the importance of using MZs in studying the water dynamics of agricultural areas.

Table 1. Accuracy indices for predicting the available soil water (SAW) by multiple linear regression (MLR) using the backward stepwise method and cross-validation for different soil water potentials (-1, -6 and -10kPa).

MZs	Predicted	Cross Validation			Performance					
		RMSE	R2	MAE	adj R2	RSE	AIC	BIC	PER	p.value
Z1	SAW1	2.94	0.75	2.33	0.75	2.92	6109	6201	**0.07**	<0.001
	SAW6	2.88	0.71	2.27	0.70	2.86	6058	6145	0.10	<0.001
	*SAW10	2.80	0.75	2.19	**0.75***	**2.77**	**5981**	**6073**	0.11	<0.001
Z2	SAW1	5.62	0.63	4.09	0.63	5.62	12017	12117	**0.11**	<0.001
	*SAW6	5.54	0.63	4.02	0.63	**5.53**	**11956**	12084	0.14	<0.001
	SAW10	5.53	0.61	4.01	0.61	5.54	11958	**12052**	0.15	<0.001
Z3	SAW1	4.25	0.71	3.30	**0.71**	4.23	10015	10124	**0.07**	<0.001
	*SAW6	4.10	0.65	3.15	0.65	**4.09**	**9898**	**10018**	0.09	<0.001
	SAW10	4.23	0.70	3.29	0.70	4.21	10001	10105	0.10	<0.001
All Field	SAW1	5.69	0.71	4.39	**0.71**	5.69	33461	33605	**0.11**	<0.001
	*SAW6	5.48	0.70	4.20	0.70	**5.49**	**33069**	**33214**	0.14	<0.001
	SAW10	5.67	0.68	4.37	0.69	5.66	33396	33541	0.15	<0.001

* SAW 1, 6 and 10 represents soil water content w/ FC in -1, -6 and -10 kPa, respectively. RMSE is Root Mean Squared Error, R2 is determination coefficient, MAE is Mean Average Error, RSE is Residual Standard Error, AIC is Akaike criteria, BIC is Bayesian criteria, PER is Percentage Relative Error. Numbers in bold represent the best values according to the accuracy indexes.

Region Z1 (SAW ~ 10-30 mm) presented lower SAW values than Z2 (SAW ~ 20-40 mm) or Z3 (SAW ~ 20-50 mm) (Fig. 5A). The SAW simulated surface (Fig 5A) was seen to be smoother and more homogeneous compared to SAW measured (Fig. 5B). Regions Z1 and Z3 showed greater stability and a smaller range of variation compared to Z2, again showing Z2 as a region in transition between the first and third regions (Fig. 5C). It can also be inferred that the model was not able to estimate the observed SAW values more accurately above 50 mm, which may have influenced the spatial smoothing of the estimated surface.

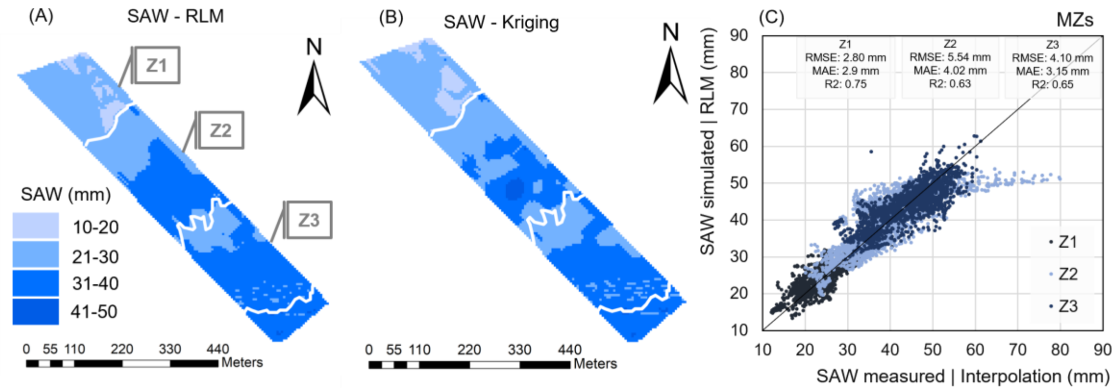
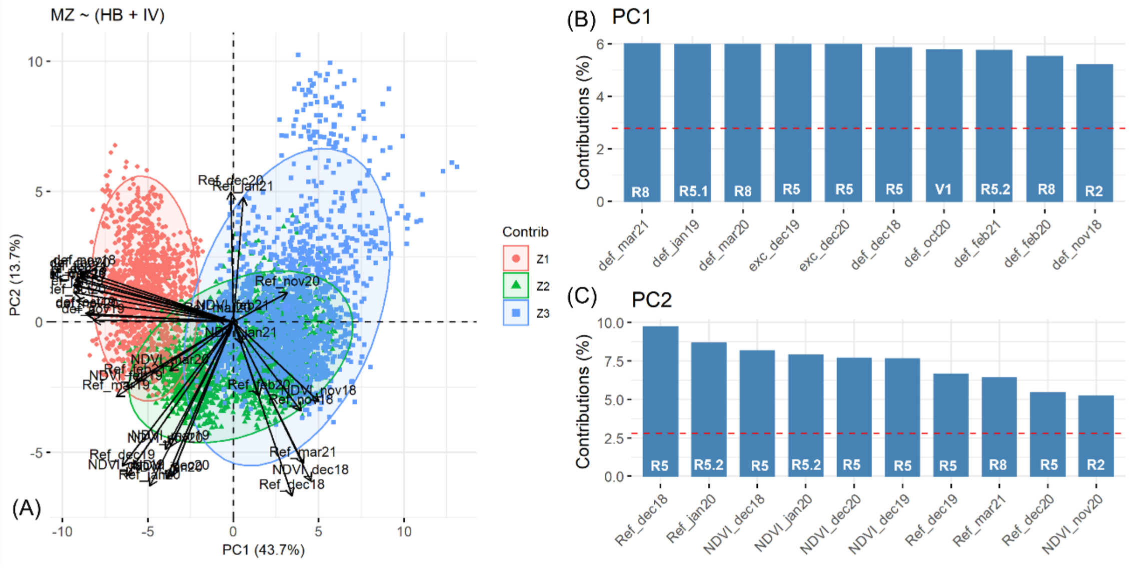


Figure 5. Spatial simulation of Soil Available Water predicted by RLM (A), Observed SAW by interpolation of measured points (B), and 1:1 curve by management zone (Z1, Z2 and Z3).

3.3.3. Multivariate analysis and predicting soybean yield as a function of WB, NIR and NDVI

The result of the PCA of the predictors used to estimate soybean yield and the contributions of each variable can be seen in Fig. 6. The explained variance of the predictors as a function of the MZs for the first two PCs was 57%. Z1 showed a grouping between quadrants two and three, characterized by high values for DEF and EXC (Fig. 6A). Z3 showed grouping in quadrants one and four, characterized by variations in the vegetation indices, particularly in RefNIR. Z2 was characterized as a region in transition to Z3, with a larger grouping of VIs. Fig. 6B shows that PC1 was strongly influenced by contributions related to the soil water balance, and PC2 by contributions from the vegetation indices. Considering the temporal aspect of the contributing predictors, the contributions were greater between the flowering stage (R5) and the grain filling stage (R5.2).



0

Figure 6. Principal component analysis of the predictors (monthly vegetation indices - VI and soil water balance - HB) (A) and contributions of PC1 (B) and PC2 (C) used to estimate soybean yield. *Ref represents the temporal NIR.

Regardless of the methods (RF or MLR), MZs (Z1, Z2 and Z3) or VIs (NIR and NDVI), the use of predictors from the three previous seasons helped reduce the RMSE and MAE and increase R2 (Table 2). The prediction quality for soybean yield using RefNIR as an input variable was superior to that obtained using NDVI. The use of RF as the prediction method showed superior performance to using MLR.

Table 2. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Rsquared (R2) for predicting soybean production (2020/21) using three boundary conditions (M1, M2 and M3) based on multiple linear regression (MLR) and the Random Forest technique (RF).

Model	MAE			RMSE			R2		
	Z1	Z2	Z3	Z1	Z2	Z3	Z1	Z2	Z3
M1.RF: NDVI + HB	0.212	0.236	0.241	0.298	0.306	0.321	0.176	0.129	0.205
M1.RF: NIR + HB	0.179	0.218	0.230	0.259	0.289	0.308	0.387	0.255	0.252
M1.MLR: NDVI + HB	0.218	0.236	0.247	0.312	0.319	0.319	0.143	0.080	0.129
M1.MLR: NIR + HB	0.221	0.240	0.243	0.306	0.317	0.318	0.096	0.072	0.161
M2.RF: NDVI + HB	0.200	0.221	0.235	0.284	0.292	0.307	0.210	0.230	0.257
M2.RF: NIR + HB	0.168	0.203	0.217	0.244	0.273	0.296	0.462	0.351	0.322
M2.MLR: NDVI + HB	0.215	0.244	0.246	0.309	0.322	0.321	0.164	0.035	0.126
M2.MLR: NIR + HB	0.221	0.239	0.247	0.310	0.320	0.325	0.081	0.060	0.105
M3.RF: NDVI + HB	0.179	0.195	0.223	0.249	0.264	0.291	0.402	0.379	0.303
[1]M3.RF: NIR + HB	***0.164**	**0.185**	**0.212**	**0.233**	**0.243**	**0.287**	**0.531**	**0.472**	**0.353**
M3.MLR: NDVI + HB	0.209	0.237	0.249	0.296	0.316	0.316	0.199	0.094	0.125
M3.MLR: NIR + HB	0.210	0.234	0.243	0.288	0.316	0.315	0.217	0.094	0.163

[1]M3.RF:NIR+HB="Yield2020/21=NIR$_{\text{(nov16+dec16+jan17+feb17+mar17+dec17+jan18+mar18+nov18+dec18+mar19+dec19+jan20+feb20+mar20+nov20+dec20+jan21+feb21+mar21)}}$+WB[DEF$_{\text{(oct16+nov16+dec16+jan17+feb17+mar17+oct17+nov17+dec17+jan18+feb18+mar18+oct18+nov18+dec18+jan19+feb19+mar19+oct19+nov19+dec19+jan20+feb20+mar20+oct20+nov20+dec20+jan21+feb21+mar21)}}$+EXC$_{\text{(oct16+nov16+dec160+jan17+feb17+mar17+oct17+nov17+dec17+jan18+feb18+mar18+oct18+nov18+dec18+jan19+feb19+mar19+oct19+nov19+dec19+jan20+feb20+mar20+oct20+nov20+dec20+jan21+feb21+mar21)}}$]".*black are the most accurate model.

3.3.4. RF as an alternative for predicting soybean yield in wet or dry seasons

The results of the RF method for estimating soybean yield in a dry year (2019/20 season) and a rainy year (2018/19 season), based on the DEF, EXC and RefNIR monthly predictors from the three previous seasons are shown in Fig 7. Z3 showed different behavior to Z1 and Z2, where there was a substantial reduction in RMSE and MAE during the rainy year (20% and 25%, respectively) compared to the dry year (9% and 15%, respectively). In regions Z1 and Z2, the performance of the yield prediction was lower for the rainy year compared to the dry year, which may be related to the lower water retention capacity in these regions compared to Z3.

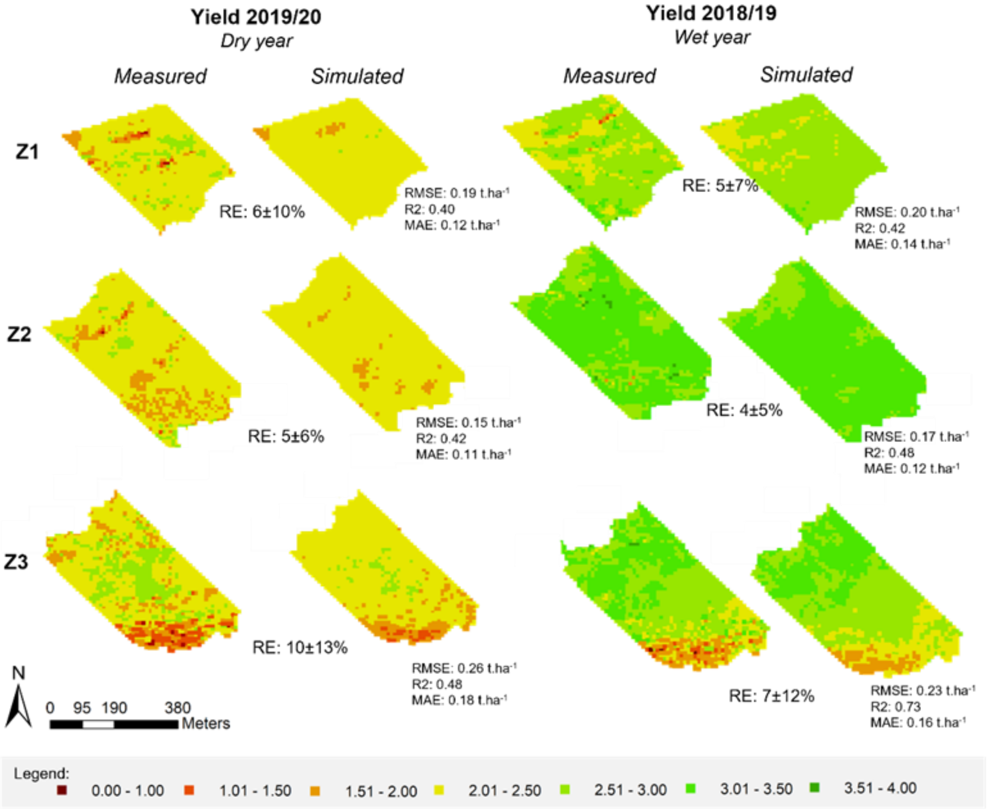


Figure 7. Spatial and temporal variability of soybean yield by MZs as a function of soil the water balance and NIR reflectance for the 2019/20 and 2018/19 seasons. *RE = Relative Error.

Fig. 8 shows the result of the prediction of soybean yield as a function of the MZs and the stability of the observed data. The simulated data showed a smaller range of variation than the observed data. One of the factors associated with this result may have been due to the drought intensifying over the previous three harvests (2019/20, 2020/21 and 2021/22), which possibly affected the drop in yield in relation to the historical series, and even the performance of the model. The longer time taken for the collections and manual evaluations justifies the need to use harvest maps to carry out the study. On the other hand, the values for RMSE found in this study ranged from 0.15 to 0.60 t ha^{-1}, very similar to the estimates of soybean yield supported by agroecosystem models, such as Aquacrop 0.27-0.35 t ha^{-1} (CAMPOS et al., 2018). The same was seen for MAE, with results ranging from 0.11 to 0.21 t ha^{-1}, and in the literature, from 0.11 to 0.33 t ha^{-1} (da SILVA et al., 2017).

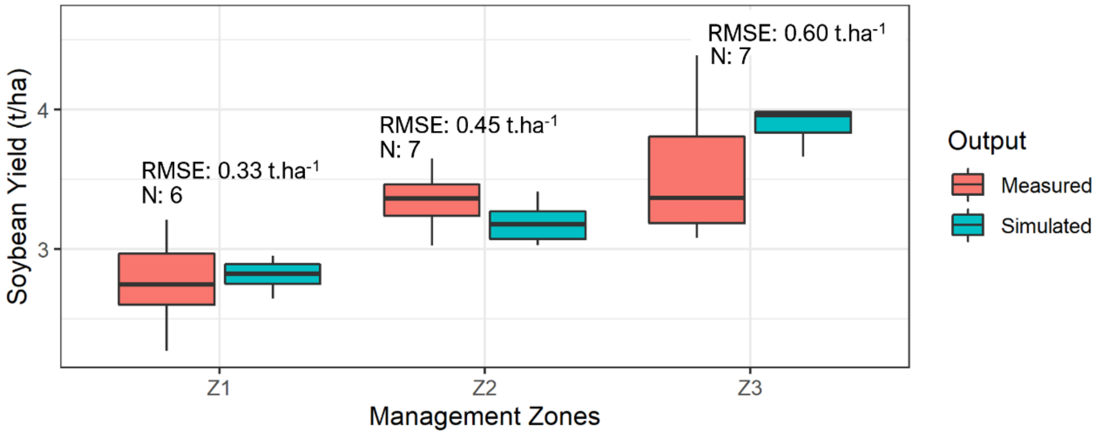


Figure 8. Boxplot of measured and simulated soybean yield (2021/22) in the management zones (Z1, Z2 and Z3) as a function of the soil water balance and variability in NIR reflectance.

3.4. Discussion

3.4.1. Performance of the digital mapping of soil water as a function of MZs versus the entire area

Understanding the spatial variability of the SAW in each of the MZs made it possible to identify distinct patterns between and within regions. Under local field conditions, the accuracy of SAW estimated by MLR was considered high, corroborating other research under different climate conditions (VEREECKEN et al., 2010, Szabó et al., 2019). The use of digital soil mapping was a quick and effective alternative in this research, which ensured an understanding of the soil and its different behaviors in each management zone, similar to the studies of Cousin et al. (2022). It was possible to validate the first hypothesis of this study, in which the soil water balance in the MZs varies compared to the entire area, justifying the importance of the heterogeneous management of these regions.

For the SAW prediction, the granulometric fraction, bulk density and organic matter surfaces were used, which favored an increase in estimation accuracy in our study. According to Dongli et al. (2017) and Amorim et al. (2022), the use of primary soil variables in models for estimating soil water can be an alternative method for reducing the number of analyses and collections necessary to determine the PWP and FC, which are considered time-consuming.

In region Z1, bulk density proved to be more important for the SAW, possibly as the region is sandier, with naturally higher densities (REINERT et al., 2008). The interactions of SAW with the carbon content of the soil may be associated with the predominance of the type of MOS fraction in each region. It’s important to note that the MOP and MOM affect both the structure and adsorption properties of the soil, and as a result, water retention. These MOS fractions are sensitive to changes that occur due to climate change and the changes in management practices that correspond to regions Z1, Z2 and Z3 defined in this study (DAMIAN et al., 2016). In Z3, the increase in MOM was more significant. MOM corresponds to a more stable fraction of the MOS and plays a fundamental role in the stability of microaggregates (CAMBARDELLA and ELLIOT, 1992). Z1 showed high levels of MOP, considered a labile fraction, and may be more sensitive to anthropogenic or natural variation (BAYER et al., 2002; ROSSI et al., 2012). Region Z2 appeared transitory in relation to the SAW and most of the soil predictors.

3.4.2. Machine learning as an alternative for predicting multi-year soybean yield by MZ

Several studies have predicted soybean yield on a regional scale in tropical areas, considering vegetation and soil water indices (MERCANTE et al., 2010; GRZEGOZEWSKI et al., 2017; SONG et al., 2022). However, few studies have predicted soybean yield at the level of management zones. Machine learning and remote sensing tools may be able to speed up the efficient use of inputs on a local scale to ensure higher yield (LEO et al., 2023). Prediction accuracy using the RF method was superior to the MLR method, as seen in other studies (FILGUEIRAS et al., 2020; SONG et al., 2022). The use of ML is promising for predicting soybean yield as a function of MZs, validating our hypothesis. This is related to the ability of the RF method to reduce multicollinearity, which helps resolve sensitive interactions between the predictors and the simulated variables (BROKAMP et al., 2017).

The temporal and spatial seasonality of water deficits, water surpluses, and the NIR increased the prediction accuracy of soybean yield in the management zones. This increase resulted from the addition of these temporal predictors from the three previous seasons. The estimate of soybean yield in the zone of low potential (Z1) was mainly influenced by variations in the water and carbon, a result of the multivariate analysis. The largest contributors seen in the region were related to the periods during which the crop was at stage R5, i.e. during the grain filling stage. An increase in water deficit during the grain filling phase drastically reduces crop yield (Neumaier et al., 2000). The low-potential zone is a sandier region, where the impact of drought on the crops is greater when there is inadequate protection of the soil or crop residue (SCUDIERO et al., 2018), coupled with the fact that with the increased water deficit, plants tend not to produce a large amount of photoassimilates.

The regions of high and medium potential (Z2 and Z3, respectively) were classified as more clayey, with greater water retention and stable carbon, and were more dependent on crop-related factors. Due to these characteristics, the impact of drought on the soil may be less intense compared to the zone of low potential. Clay soils and a high carbon content contribute to the resilience of the soil to events of extreme water shortage. Lizumi and Wagai (2019) estimated that small increases in the organic carbon content of the soil in the 0–30 cm layer are sufficient to increase the drought tolerance of agricultural systems that operate in more than 70% of the global harvested area. On the other hand, a reduction in soybean yield was seen in the southern part of the zone with medium potential, possibly due to the transition from an Oxissol, a highly weathered soil, to an Entissol, a less developed soil that is stonier with shallower profiles.

Using the NIR vegetation index resulted in higher performance than using NDVI, validating our hypothesis and justifying the potential use of reflectance surfaces as an aid to agriculture. NIR reflectance is characterized by the strong absorption of red light by chlorophyll and low absorption by green leaves, showing a correlation with the water balance in the plant (Liu et al., 2014). In this research, the use of reflectance, representing development of the canopy, contributed to the local prediction of soybean yield, possibly due to the correlation of the spectral range with such parameters as water and yield, as also seen in studies under controlled conditions (FILGUEIRAS et al., 2020). The use of high-resolution satellite data on a monthly scale was one of the factors that made this study possible. On the other hand, for estimating soybean yield, the characteristics derived from satellite sensors should vary at most by an interval of between two and three months (SKAKUN et al., 2021). In the case of this study, the temporal aspect of the predictors (one month) may have guaranteed an increase in prediction accuracy.

Spatial statistics make it possible to evaluate the variability of soybean production in different regions, together with spatial and temporal aspects made possible by geographic information systems, as seen in this and other studies (GRZEGOZEWSKI et al., 2017). On the other hand, validation on a temporal scale requires obtaining predictive surfaces over time and needs at least a monthly history. One of the main challenges of this study was to obtain historical data from reliable sources with representation on a local scale, as is the case of harvest maps. Alternatives should become possible through research and development, supporting quick decision-making by the farmer, seeking to increase yield in any one area, and considering the economic and environmental aspects. This will only be possible with the practical application of modelling under actual field conditions, as is the case of our research.

The use of agroecosystem models has the advantage of predicting environmental factors more efficiently and on a larger scale, which can assist in the expansion and agility of research and marketing decisions. More and more, machine models are being used to estimate and predict these factors, as shown in this research, advancing towards a new frontier with the use of AI in digital and precision agriculture.

3.5. Conclusions

This research combined digital soil mapping techniques incorporating the spatial and temporal variability of water and vegetation to map soybean yield. This was only possible through the use of remote sensing, machine learning techniques, and high-resolution machine data. The digital mapping of soil water based on management zones proved to be an adequate alternative compared to evaluating the entire area. The use of near-infrared reflectance surfaces to predict soybean development resulted in higher performance than when using NDVI. The Random Forest machine learning technique showed greater accuracy in estimating soybean yield in each region compared to the conventional method. The strategy presented in this study can help decision-making in agricultural management as it allows the most important characteristics that condition yield to be recognized in each management zone, and also based on the temporal conditions of the crop, climate and soil during each season. The current approach guarantees rapid results for management at the field level, allowing the use of inputs and differentiated management to be planned on a local scale, season by season. The present research corroborates traditional approaches that employ high-density, sporadic, or systematic sampling, where the use of MZs allows fewer samples to be collected, which are allotted based on the performance of the crop, and from which it is possible to observe a greater number of variables and establish more-significant relationships from an agronomic point of view. In particular, this study showed the relevance of including variables related to water availability among those used as inputs for the models.

References

ADEBOYE OB, SCHULTZ B, ADEBOYE AP, ADEKALU KO, OSUNBITAN JA. Application of the AquaCrop model in decision support for optimization of nitrogen fertilizer and water productivity of soybeans. **Information Processing in Agriculture** 2021;8:419–36. https://doi.org/10.1016/j.inpa.2020.10.002.

ALABI TR, Abebe AT, Chigeza G, Fowobaje KR. Estimation of soybean grain yield from multispectral high-resolution UAV data with machine learning models in West Africa. **Remote Sensing Applications: Society and Environment** 2022;27:100782. https://doi.org/10.1016/j.rsase.2022.100782.

ALI A, RONDELLI V, MARTELLI R, FALSONE G, LUPIA F, BARBANTI L. Management Zones Delineation through Clustering Techniques Based on Soils Traits, NDVI Data, and Multiple Year Crop Yields. **Agriculture** 2022;12:231. https://doi.org/10.3390/agriculture12020231.

ALMEIDA, B.G ET AL. **Standardization of Methods for Granulometric Analysis in Brazil.** Technical Notice. ISSN 1517-5685. Rio de Janeiro, RJ: Embrapa, 2012.

AMIRI M, SALEM A, GHZAL M. Spatial-Temporal Water Balance Components Estimation Using Integrated GIS-Based Wetspass-M Model in Moulouya Basin, Morocco. **IJGI** 2022;11:139. https://doi.org/10.3390/ijgi11020139.

AMORIM RSS, ALBUQUERQUE JA, COUTO EG, KUNZ M, RODRIGUES MF, SILVA L DE CM DA, ET AL. Water retention and availability in Brazilian Cerrado (neotropical savanna) soils under agricultural use: Pedotransfer functions and decision trees. **Soil and Tillage Research** 2022;224:105485. https://doi.org/10.1016/j.still.2022.105485.

BAYER, C. et al Stocks and humification degree of organic matter fractions as afeccted by no-tillage on a subtropical soil. **Plant Soil,** v. 238, n. 01, p. 133-140, 2002.

BALL, B.C., HUNTER, R., 1988. The determination of water release characteristics of soil cores at low suctions. **Geoderma** 43, 195–212.

BREUNIG FM, GALVÃO LS, DALAGNOL R, SANTI AL, DELLA FLORA DP, CHEN S. Assessing the effect of spatial resolution on the delineation of management zones for smallholder farming in southern Brazil**. Remote Sensing Applications: Society and Environment** 2020;19:100325. https://doi.org/10.1016/j.rsase.2020.100325.

BROKAMP C, JANDAROV R, RAO MB, LEMASTERS G, RYAN P. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. **Atmospheric Environment** 2017;151:1–11. https://doi.org/10.1016/j.atmosenv.2016.11.066.

BRUCE, PETER, AND ANDREW BRUCE. 2017. **Practical Statistics for Data Scientists.** O'Reilly Media.

CAMBARDELLA, C. A.; ELLIOTT, E. T. Particulate soil organic-matter changes across a grassland cultivation sequence. **Soil Science Society of America Journal**, v. 56, n. 03, p. 777-783, 1992.

CAMBARDELLA, C.A.; MOORMAN, T.B.; PARKIN, T.B.; KARLEN, D.L.; NOVAK, J.M.; TURCO, R.F.; KONOPKA, A.E. **Field-scale variability of soil properties in central Iowa soils**. Soil Sci. Soc. Am. J. 1994, 58, 1501–1511.

CAMPOS MANTOVANELLI B, PETRY MT, BROETTO WEILER E, CARLESSO R. Geostatistical interpolation based ternary diagrams for estimating water retention properties in soils in the Center-South regions of Brazil. **Soil and Tillage Research** 2021;209:104973. https://doi.org/10.1016/j.still.2021.104973.

CAMPOS I, NEALE CMU, ARKEBAUER TJ, SUYKER AE, GONÇALVES IZ. Water productivity and crop yield: A simplified remote sensing driven operational approach. **Agricultural and Forest Meteorology** 2018;249:501–11. https://doi.org/10.1016/j.agrformet.2017.07.018.

COUSIN, I., BUIS, S., LAGACHERIE, P. et al. Available water capacity from a multidisciplinary and multiscale viewpoint. A review. **Agron. Sustain. Dev.** 42, 46 (2022). https://doi.org/10.1007/s13593-022-00774-8

CUEFF S, COQUET Y, AUBERTOT J-N, BEL L, POT V, ALLETTO L. Estimation of soil water retention in conservation agriculture using published and new pedotransfer functions. **Soil and Tillage Research** 2021;209:104967. https://doi.org/10.1016/j.still.2021.104967.

DA SILVA EHFM, HOOGENBOOM G, BOOTE KJ, GONÇALVES AO, MARIN FR. Predicting soybean evapotranspiration and crop water productivity for a tropical environment using the CSM-CROPGRO-Soybean model. **Agricultural and Forest Meteorology** 2022;323:109075. https://doi.org/10.1016/j.agrformet.2022.109075.

DONGLI S, QIAN C, TIMM LC, BESKOW S, WEI H, CALDEIRA TL, et al. Multi-scale correlations between soil hydraulic properties and associated factors along a Brazilian watershed transect. **Geoderma** 2017;286:15–24. https://doi.org/10.1016/j.geoderma.2016.10.017.

DAMIAN JM, PIAS OHC, SANTI AL, DI VIRGILIO N, BERGHETTI J, BARBANTI L, MARTELLI R. Delineating management zones for precision agriculture applications: a case study on wheat in sub-tropical Brazil. **Italian Journal of Agronomy** 2016; 11:171-179. https://doi.org/10.4081/ija.2016.713

ESRI 2011. **ArcGIS Desktop: Release 10.** Redlands, CA: Environmental Systems Research Institute.

FERINA J, VUČETIĆ V, BAŠIĆ T, ANIĆ M. Spatial distribution and long-term changes in water balance components in Croatia. **Theor Appl Climatol** 2021;144:1311–33. https://doi.org/10.1007/s00704-021-03593-1.

FILGUEIRAS R, ALMEIDA TS, MANTOVANI EC, DIAS SHB, FERNANDES-FILHO EI, DA CUNHA FF, et al. Soil water content and actual evapotranspiration predictions using regression algorithms and remote sensing data. **Agricultural Water Management** 2020;241:106346. https://doi.org/10.1016/j.agwat.2020.106346.

GLOBAL MODELING AND ASSIMILATION OFFICE (GMAO) (2015), inst3_3d_asm_Cp**: MERRA-2 3D IAU State, Meteorology Instantaneous 3-hourly (p-coord, 0.625x0.5L42),** version 5.12.4, Greenbelt, MD, USA: Goddard Space Flight Center Distributed Active Archive Center (GSFC DAAC), Accessed Enter User Data Access Date at doi: 10.5067/VJAFPLI1CSIV.

GRZEGOZEWSKI DM, URIBE-OPAZO MA, JOHANN JA, GUEDES LPC. Spatial correlation of soybean productivity, enhanced vegetation index (evi) and agrometeorological variables. **Agric Eng** 2017;37:541–55. https://doi.org/10.1590/1809-4430-eng.agric.v37n3p541-555/2017.

IIZUMI, T., WAGAI, R. Leveraging drought risk reduction for sustainable food, soil and climate via soil organic carbon sequestration**. Sci Rep** 9, 19744 (2019). https://doi.org/10.1038/s41598-019-55835-y

KUHN, M. CARET: **Classification and Regression Training.** R package version 6.0-90. (2021) https://github.com/topepo/caret/

LEO S, DE ANTONI MIGLIORATI M, NGUYEN TH, GRACE PR. Combining remote sensing-derived management zones and an auto-calibrated crop simulation model to determine optimal nitrogen fertilizer rates. **Agricultural Systems** 2023;205:103559. https://doi.org/10.1016/j.agsy.2022.103559.

LIU L, HUANG W, PU R, WANG J. Detection of Internal Leaf Structure Deterioration Using a New Spectral Ratio Index in the Near-Infrared Shoulder Region. **Journal of Integrative Agriculture** 2014;13:760–9. https://doi.org/10.1016/s2095-3119(13)60385-8.

MERCANTE E, LAMPARELLI RAC, URIBE-OPAZO MA, ROCHA JV. Linear regression models to estimate soybean productivity in western Paraná, using spectral data. **Agricultural Engineer** 2010;30:504–17. https://doi.org/10.1590/s0100-69162010000300014.

MINASNY B, MCBRATNEY ALEXB. Digital soil mapping: A brief history and some lessons. **Geoderma** 2016;264:301–11. https://doi.org/10.1016/j.geoderma.2015.07.017.

MCBRATNEY AB, MENDONÇA SANTOS ML, MINASNY B. On digital soil mapping. **Geoderma** 2003;117:3–52. https://doi.org/10.1016/s0016-7061(03)00223-4.

MURUGANANTHAM P, WIBOWO S, GRANDHI S, SAMRAT NH, ISLAM N. A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing. **Remote Sensing** 2022;14:1990. https://doi.org/10.3390/rs14091990.

NEUMAIER, N., NEMOPUCENO, A.L., FARIAS, J.R., OYA, T. **Soybean development stages.** In: BONATO, E. R. (Ed.). Stresses in soy. Passo Fundo: Embrapa Wheat, 2000.

OTTONI FILHO, T.B., OTTONI, M.V., OLIVEIRA, M.B., MACEDO, J.R. Estimation of field capacity from ring infiltrometer-drainage data. **R. Bras. S. Sci.,** 38:1765-1771, 2014. https://www.redalyc.org/articulo.oa?id=180232852011

PLANET. **Planet Imagery Product Specifications**—June 2020; Planet Labs, Inc.: San Francisco, CA, USA, 2020.

RAO P, ZHOU W, BHATTARAI N, SRIVASTAVA AK, SINGH B, POONIA S, et al. Using Sentinel-1, Sentinel-2, and Planet Imagery to Map Crop Type of Smallholder Farms. **Remote Sensing** 2021;13:1870. https://doi.org/10.3390/rs13101870.

REINERT, D.J., ALBUQUERQUE, J.A., REICHERT, M., AITA, C., ANDRADA, M.M. Critical limits of soil density for the growth of roots of plants of coverage in red argisol. **R. Bras. S. Sci.,** 32:1805-1816, 2008.

ROLIM, G.S., SENTELHAS, P.C., BARBIERI, V. Spreadshets in excel environment to calculation of water balance: normal, sequential, culture and potential, real productivity. **J. Braz. Agrom.** V.6, n.1, p.133-137, 1998.

ROSSI, C.Q., PEREIRA, M.G., GIÁCOMO, S.G., BETTA, M., POLIDORO, J.C. Labile fractions of organic matter in cropping system with straw of brachiaria and sorghum. **J. Agron. Sci.** 43 (1) • Mar 2012. https://doi.org/10.1590/S1806-66902012000100005

R CORE TEAM (2022). **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

SAGAN V, MAIMAITIJIANG M, BHADRA S, MAIMAITIYIMING M, BROWN DR, SIDIKE P, et al. Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning. ISPRS **Journal of Photogrammetry and Remote Sensing** 2021;174:265–81. https://doi.org/10.1016/j.isprsjprs.2021.02.008.

SILVA V DE PR DA, SILVA RA E, MACIEL GF, BRAGA CC, SILVA JÚNIOR JLC DA, SOUZA EP de, et al. Calibration and validation of the AquaCrop model for the soybean crop grown under different levels of irrigation in the Motopiba region, Brazil. **Rural Sci.** 2017;48. https://doi.org/10.1590/0103-8478cr20161118

SPEKKEN, M., ANSELMI, A.A., MOLIN, J.P. A simple method for filtering spatial data.

In: 9th European conference on precision agriculture, 2013, Lleida.

SOUZA, S.A., RODRIGUES, L.N. & DA CUNHA, F.F. Assessing the precision irrigation potential for increasing crop yield and water savings through simulation. **Precision Agric** (2022). https://doi.org/10.1007/s11119-022-09958-4

SONG X-P, LI H, POTAPOV P, HANSEN MC. Annual 30 m soybean yield mapping in Brazil using long-term satellite observations, climate data and machine learning. **Agricultural and Forest Meteorology** 2022;326:109186. https://doi.org/10.1016/j.agrformet.2022.109186.

SOIL SURVEY STAFF, 2014 **Soil Survey Staff Keys to Soil Taxonomy** (12th), USDA (2014), pp. 1-410

SCHWALBERT RA, AMADO TJC, REIMCHE GB, GEBERT F. Fine-tuning of wheat (Triticum aestivum, L.) variable nitrogen rate by combining crop sensing and management zones approaches in southern Brazil. **Precision Agric** 2018;20:56–77. https://doi.org/10.1007/s11119-018-9581-6.

SCUDIERO E, TEATINI P, MANOLI G, BRAGA F, SKAGGS T, MORARI F. Workflow to Establish Time-Specific Zones in Precision Agriculture by Spatiotemporal Integration of Plant and Soil Sensing Data. **Agronomy** 2018;8:253. https://doi.org/10.3390/agronomy8110253.

SKAKUN S, KALECINSKI NI, BROWN MGL, JOHNSON DM, VERMOTE EF, ROGER J-C, et al. Assessing within-Field Corn and Soybean Yield Variability from WorldView-3, Planet, Sentinel-2, and Landsat 8 Satellite Imagery. **Remote Sensing** 2021;13:872. https://doi.org/10.3390/rs13050872.

STACKHOUSE PW, et al. (2017). **Prediction of worldwide energy resource (POWER), agroclimatology methodology, (1° Latitude by 1° Longitude Spatial Resolution).** Accessed in november, 2019. Link: https://power.larc.nasa.gov/documents/Agroclimatology_Methodology.pdf. Accessed 02 January 2020.

SZABÓ B, SZATMÁRI G, TAKÁCS K, LABORCZI A, MAKÓ A, RAJKAI K, et al. Mapping soil hydraulic properties using random-forest-based pedotransfer functions and geostatistics. **Hydrol Earth Syst Sci** 2019;23:2615–35. https://doi.org/10.5194/hess-23-2615-2019.

TEIXEIRA PC, DONAGEMMA GK, FONTANA A, TEIXEIRA WG. **Manual of soil analysis methods**. 3rd ed. rev. and amp. – Brasilia, DF: Embrapa, 2017

TORSONI, GB, DE OLIVEIRA APARECIDO, LE, DOS SANTOS, GM et al.

Soybean yield prediction by machine learning and climate. **Theor Appl Climatol.** 151 , 1709–1725 (2023). https://doi.org/10.1007/s00704-022-04341-9

THORNTHWAITE, C.W.; MATHER, J.R. **The water balance Centerton**, NJ: Drexel Institute of Technology - Laboratory of Climatology, 1955. 104p. (Publications in Climatology, vol. VIII, n.1)

VEREECKEN H, WEYNANTS M, JAVAUX M, PACHEPSKY Y, SCHAAP MG, VAN GENUCHTEN MTh. Using Pedotransfer Functions to Estimate the van Genuchten-Mualem Soil Hydraulic Properties: A Review. **Vadose Zone Journal** 2010;9:795–820. https://doi.org/10.2136/vzj2010.0045.

YEOMANS, A.; BREMNER, J. M. A rapid and precise method for routine determination of organic carbon in soil. **Communication Soil Science Plant Analysis**, v.19, p.1467- 1476, 1988. DOI: https://doi.org/10.1080/00103628809368027

ZENG L, WARDLOW BD, WANG R, SHAN J, TADESSE T, HAYES MJ, et al. A hybrid approach for detecting corn and soybean phenology with time-series MODIS data. **Remote Sensing of Environment** 2016;181:237–50. https://doi.org/10.1016/j.rse.2016.03.039.

4. PERFORMANCE OF CROP MODELS TO PREDICT SOYBEAN YIELD ON PHYSICAL SOIL FACTORS IN MANAGEMENT ZONES

Abstract

Crop models are widely used to estimate crop yield in uniform areas, considering constant levels of soil factors as water retention. On the other hand, the available soil water varies depending on different factors at the plot level. The aim of this chapter was to evaluate the performance of two models for estimating soybean yield in different established regions based on the physical factors of the soil. The Aquacrop-FAO and DSSAT-CROPGRO models were used, and 44 soil profiles (Z1 = 8, Z2 = 18 and Z3 = 18) were simulated, considering texture, carbon and soil water as input for the crop models. The simulations were carried out over three consecutive seasons, including one wet year (2018/19) and two dry years (2019/20 and 2020/21). There were significant variations in available soil water (SAW) in the different regions. The more clayey region Z3 (44 mm) had a higher SAW than the sandier region Z1 (30 mm). The Aquacrop model performed better in estimating yield in the management zones for the wet year (2018/19: RMSE < 1172 kg ha^{-1}, MAE < 1093 kg ha^{-1}) and the CROPGRO for the dry years (2019/20: RMSE < 865 kg ha^{-1}, MAE < 815 kg ha^{-1}; 2020/21: RMSE < 1137 kg ha^{-1}, MAE < 997 kg ha^{-1}). The Aquacrop model overestimated yield compared to CROPGRO. It can be concluded that changes in soil factors influence yield in regions within the same plot in both crop models. More studies should be carried out in order to understand the performance of yield prediction by crop models at a spatial and temporal scale.

Keywords: Aquacrop, FAO, DSSAT, Management Zones, Soybean, SAW

4.1. Introduction

Understanding the spatial variability and temporal determinants of soybean yield is fundamental to decision-making. Crop growth models are widely used for estimating yield; on the other hand, understanding the sensitivity of such methods at the farm level are necessary as management alternatives in precision agriculture (AHMADPOUR et al., 2022; SINGH et al., 2023). The use of crop models to predict yield can help in the intelligent management of production environments (SINGH et al., 2023).

Among the models, there are those that use the relationship between carbon and radiation to estimate yield and crop development, as is the case of DSSAT (Decision Support System for Agrotechnology) which uses the CSM-CROPGRO module (MULAZZANI et al., 2022; AKUMAGA et al., 2023; SINGH et al., 2023). DSSAT-CSM-CROPGRO is one of the most used models for crop simulation, and employs soil, climate, management and cultivar information to predict crop yields (AKUMAGA et al., 2023).

Singh et al. (2023) evaluated the use of the CROPGRO model to estimate soybean yield at a spatial resolution of 25 km in a humid subtropical climate, with a variation of between 8% and 24% compared to the observed results. The authors stated that the methodology was promising, but the time taken in preparing the input data was high. Akumaga et al. (2023), also in a humid subtropical climate, used remote sensing data and field observations as inputs to the DSSAT model to predict soybean yield, obtaining an RMSE of between 200 and 700 kg ha-1, and considered the predictions obtained by the model as reasonable.

Other research is based on models that include the relationship between available water and biomass development based on the concept of water yield (STEDUTO et al., 2007; ADEBOYE et al., 2019). The FAO (Food and Agriculture Organisation of the United Nations) created the Aquacrop model, which can predict the development and yield of agricultural crops based on water efficiency and availability. According to Adeboye et al. (2019), the model is relatively simple, with few input parameters and intuitive algorithms for obtaining the evaluated outputs. Battisti and Sentelhas (2015), using Aquacrop to evaluate drought-tolerant soybean cultivars, found prediction errors in soybean yield of less than 280 kg ha-1. Adeboye et al. (2017), evaluating the performance of Aquacrop in predicting soybean yield under rainfed conditions and different types of ground cover, obtained an RMSE of 30 kg ha-1. For the authors, the model proved to be suitable for determining yield. Gimenez et al. (2017) used Aquacrop to estimate soybean yield under rainfed conditions in the south of Brazil, obtaining variations of between 4% and 23% of the actual values. According to the same authors, in seasons of high water stress, the model did not respond well to the estimates. According to Adeboye et al. (2019), models that use processes related to water variation to estimate crop yield are better applied over space and time than models that are based on energy, due to the greater ease of normalising the processes using the dynamics of water evapotranspiration and gas exchange.

To estimate water stress, DSSAT uses field capacity and permanent wilting point variables that have a direct or indirect relationship with root growth and plant development. Aquacrop also uses the variables to estimate crop yield by means of processes that influence the availability of water in the soil. Few studies use crop models to predict yield at the management-zone level, mainly considering the variation in such soil factors as available water, generally considered constant. Understanding the influence of these factors on crop models used to predict soybean yield is therefore important. This study sought to investigate the efficiency of estimating soybean yield using two crop models (Aquacrop and CROPGRO) and their correlation with the changes in available soil water. The aim of this study is to understand the importance of evaluating soil factors at the management-zone level (Z1, Z2 and Z3), and the ability of crop models to detect these variations on soybean yield. Given the relevance of temporal variability in crop models, three different growing seasons (2018/19, 2019/20 and 2020/21) were considered.

4.2. Materials and Methods

4.2.1. Study site and Management Zones (MZs)

The study was carried out in a 10-hectare area of grain production in the district of São Jorge do Ivaí, Paraná, in the south of Brazil (23°24'S, 52°15'W, altitude 492). The soil is classified as a Red Oxisol transitioning to an Entisol (USDA, 2014). The mean historical annual rainfall is 1600 mm and the mean annual temperature is 21°C. The region is classified as humid subtropical with hot summers (ALVARES et al., 2013). The management zones were defined using the fuzzy c-means clustering technique, considering five soybean harvest maps (2017-2021), apparent electrical conductivity (Eca), soil maps referring to the 0 to 30 cm and 0 at 90 cm layers and an elevation map. The data were filtered using the Mapfilter® software (SPEKKEN, ANSELMI and MOLIN, 2013) and normalised by the range in variation. Spatial principal component analysis was used to define the input layers for clustering. Four management zones were defined, based on an evaluation of zoning quality, the fourth being disregarded in this study due to the presence of gravel, which might affect the physical analysis. The Vesper 1.62® software was used to interpolate by block kriging, considering a resolution of 25m^2. Region Z1 was considered to have low historical yields with a low ECa; Z2 to have high historical yields with a non-significant ECa, and Z3 to have a moderate ECa with non-significant yields.

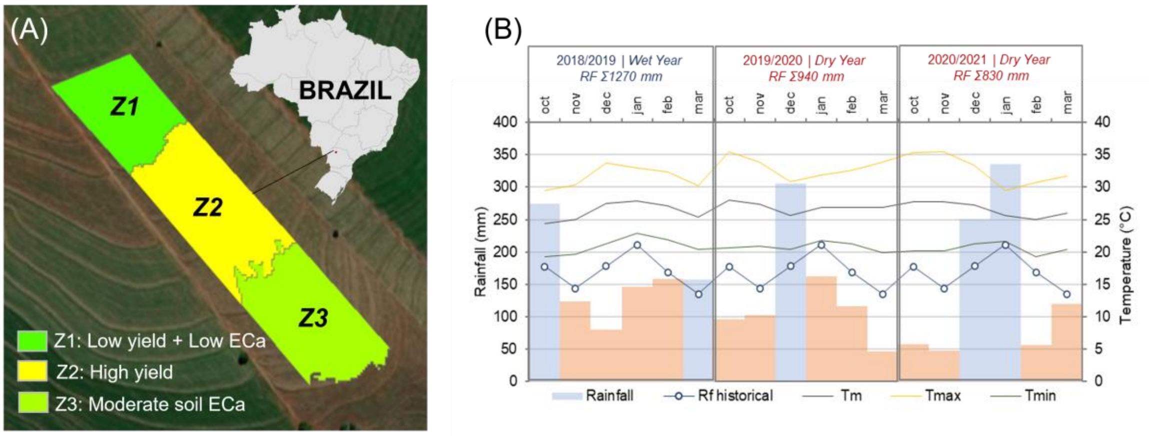


Figure 1. Spatial location of the study and management zones (A). Accumulated and monthly variation of rainfall (mm), maximum, average and minimum temperature for the 2028/19, 2019/20 and 2020/21 crops seasons (B).

4.2.2. Soil and weather databases

Soil samples were collected on an irregular grid at a density of 5 points per hectare at three depths (0-20, 20-40 and 40-60 cm). The levels of clay, silt and organic carbon (MOS), field capacity (FC) and permanent wilting point (PWP) were determined for each management zone (Table 1). Particle size was determined by mechanical and chemical dispersion as per Teixeira et al. (2017). Field capacity at -6kPa was estimated according to Ball and Hunter (1988). The PWP was determined using the WP4-T potentiometer (Meter Group Inc.) at -1.5 MPa. The MOS content was obtained from the sum of the particle-size fractionation of the soil organic matter using mineral-associated organic matter (MOM) and particulate organic matter (MOP), as per Cambardella and Elliot (1992). A geostatistical analysis of the soil variables was then carried out using the ArcMap software (ESRI, 2011) and kriging to determine high resolution surfaces (25m^2) for the three layers.

The soybean harvest is carried out in the summer, from October to March, therefore, daily meteorological data for the 2018/19, 2019/20 and 2020/21 seasons were considered. Rainfall, temperature, solar radiation, net radiation and wind-speed data were obtained from the Nasa Power platform (STACKHOUSE et al., 2017) at a spatial resolution of 50 km and temporal resolution of one day. Evapotranspiration data were calculated using the method of Thornthwaite and Mather (1955).

Table 1. Mean values for clay, silt, total organic carbon (MOS), field capacity (-6kPa) and permanent wilting point (-1.5 MPa) at different depths (0-20, 20-40 and 40-60 cm) for the management zones under consideration (Z1, Z2 and Z3).

Soil Factor	Clay	Silt	COT	Field Capacity	Wilting Point
			(%)		
0-0.20 cm					
Z1	36.8	5.9	5.1	14.3	11.7
Z2	53.3	8.7	5.0	17.1	13.4
Z3	61.7	12.1	4.8	20.8	15.9
20 – 40 cm					
Z1	45.2	4.8	5.1	15.4	12.2
Z2	59.6	7.2	4.9	18.0	14.7
Z3	66.6	12.1	4.8	22.7	17.7
40 – 60 cm					
Z1	48.7	7.0	5.1	17.0	15.2
Z2	63.9	5.4	4.9	18.1	14.2
Z3	72.8	8.4	4.7	21.8	18.3

4.2.3. Crop models development

To estimate soybean yield as a function of the management zones during the 2018/19, 2019/20 and 2020/21 seasons, two crop models were considered: Aquacrop and DSSAT-CROPGRO (HOOGEBOOM et al., 2015). Aquacrop is a platform with different crop models developed by the FAO that offers one approach to simulating growth and yield based on variations in soil and plant water using quality or stress coefficients, as per Raes et al. (2012). Climate, soil and crop data in the region were used as input for the model. The following were considered: sowing dates of 1 October 2018, 5 October 2019 and 20 October 2020, a cultivar with a cycle of 130 days, and a density of 33 plants per m^2. Forty-four virtual soil profiles were created in the 0-20, 20-40 and 40-60 cm layers considering values for soil texture, permanent wilting point and field capacity, both estimated and obtained in the field.

DSSAT-CROPGRO is a platform with simulation models for crop growth and production that consider a uniform area as a function of the variations in climate, water, soil, nitrogen and carbon. For crop management, a cultivar from maturity group 6 was considered, as recommended for the south of Brazil. A plant population of 33 plants per m^2, row spacing of 50 cm and planting depth of 5 cm were used. Forty-four virtual profiles were generated (Z1 = 8, Z2 = 18 and Z3 = 18) considering the data on clay, silt, MOS, PWP, FC in the 0-20, 20-40 and 40-60 cm layers. The saturated water content was estimated based on the pedotransfer functions of the platform, as was the saturated hydraulic conductivity. Figure 2 describes the process, from generating the management zones to estimating the performance of the best model.

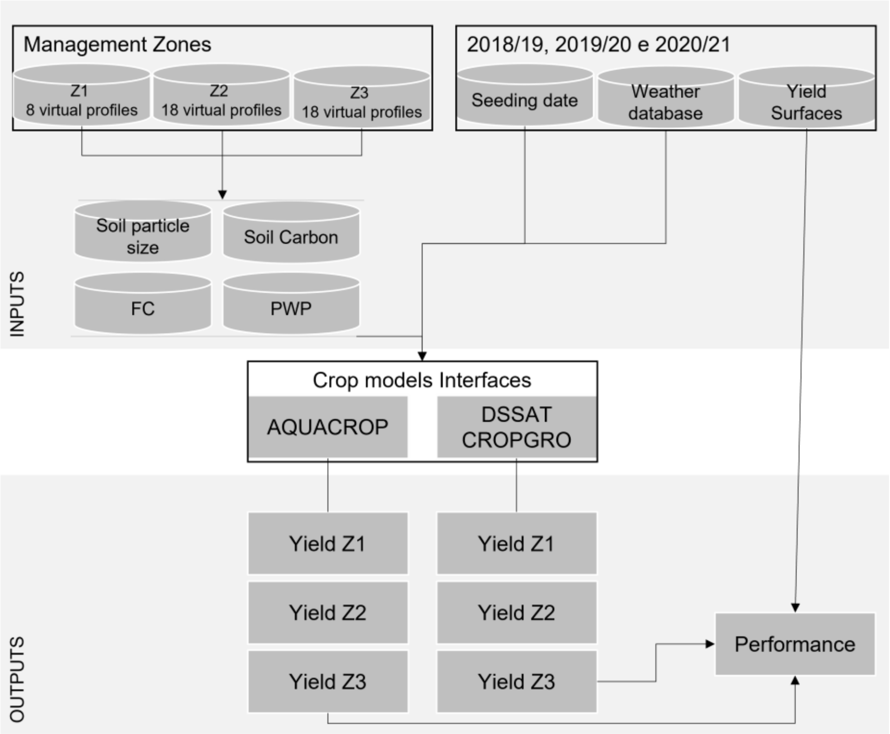


Figure 2. Flowchart of the development of the steps for the estimation of yield in the management zones (Z1, Z2 and Z3) as a function of two crop models (DSSAT-CROPGRO and Aquacrop) in three harvests (2018/19, 2019/20 and 2020/21).

4.2.4. Data analysis

The correlation between the surface-soil data and the observed data in each of the management zones was determined using the Spearman method. The homoscedasticity of soybean yield between the management zones was verified using the Kruskal-Wallis ($p<0.05$) and Man-Whitney-Wilcoxon ($p<0.05$) tests. The Spearman correlation and boxplots were also generated to compare the estimated behavior of soybean yield with the observed data. The performance of the methods for estimating soybean yield for the management zones in different seasons was evaluated based on the RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) and R2 (Coefficient of Determination), as per equations (1-3) below.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{n-k}} \qquad (1)$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{\sum_{i=1}^{n}(yi-\bar{y})^2} \qquad (2)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|yi-\hat{y}i| \qquad (3)$$

where: n represents the number of observations, k the number of estimated parameters, yi the observed value, y ̂i the estimated value, and y ̅is the sum of the observed values.

4.3. Results and Discussion

4.3.1. Correlation between soil properties and yield in MZs

The results of the correlations between the observed soil values and those of the surfaces in each of the regions can be seen in Table 2. In region Z1, the levels of clay (20-60 cm) and silt (20-40 cm) had the highest correlation with the observed data (>0.95). In Z2, significant correlations were also seen for the levels of clay (20-60 cm), silt (20-40 cm) and PWP (20-40 cm). In Z3 there were also significant, high correlations (p>0.95) with clay (20-60cm), silt (20-40 cm) and PWP (20-40cm). Regardless of the MZ, the MOS variables for some of the layers gave the worst performance. Zhang et al. (2010), comparing different kriging methods for estimating MOS, also found low correlation values for ordinary kriging (r = 0.38) due to lower sensitivity of the local variations and smoothing, which resulted in an overall representation of MOS that masked the local variations. Similar behavior was seen in our study. The use of soil data extracted from surfaces estimated by kriging is one alternative in the case of missing data or lost samples. In this study, it was possible to use these data and complete the database under evaluation. On the other hand, due to the quality of the kriging, the representativeness of the textural data was better than that of the MOS results.

Table 2. Correlation between the observed soil data and the data extracted from the kriged surfaces for the layers (0-20, 20-40 and 40-60 cm) and management zones (Z1, Z2 and Z3).

Soil Layer (cm)	Spearman Correlation				
	COT	Clay	Silt	Field Capacity	Wilting Point
		(kg kg^{-1})		(m^3 m^{-3})	
			Z1		
0-20	0.66	0.92	0.85	0.94	0.82
20-40	0.85	0.96	0.99	0.85	0.94
40-60	0.44	0.99	0.92	0.82	0.63
			Z2		
0-20	0.44	0.83	0.30	0.86	0.80
20-40	0.30	0.99	0.99	0.75	0.99
40-60	0.59	0.99	0.79	0.61	0.57
			Z3		
0-20	0.71	0.81	0.15	0.73	0.75
20-40	0.45	0.99	0.99	0.66	0.99
40-60	0.58	0.97	0.46	0.52	0.64

The results of the mean, maximum and minimum variability, and standard deviation of available soil water in each of the management zones are shown in Table 3. For region Z1, there was an average SAW of 30 mm. In region Z2, the mean SAW did not differ significantly from Z1 or Z3, with a mean of 39 mm. Region Z3, on the other hand, had the highest mean SAW (43 mm), differing significantly from region Z1. It can be seen that region Z2 presented transition characteristics with a higher mean standard deviation compared to the other regions. According to Van Lier et al. (2022), SAW is related to soil texture, with sandier soils having a lower SAW, and more-clayey soils a higher SAW. The results of the authors corroborate the present study, with a higher sand content in Z1 compared to Z3. Table 3 shows the importance of considering variations in soil water in different regions of the same area, as they can significantly affect crop yield under rainfed conditions.

Table 3. Average, maximum, minimum and standard deviation values observed (2020/21) of available water in the soil (SAW) by management zones (Z1, Z2 and Z3).

MZs	Total SAW (mm)			
	Average	Maximum	Minimum	Standard Deviation
Z1*	30.11 B	41.19	18.16	8.18
Z2	39.06 AB	76.90	26.22	12.01
Z3	43.46 A	56.94	29.30	7.66

*Diferença entre letras são significativas pelo teste de Wilcoxcon ($p<0.05$).

The results of SAW correlations (p) between the observed yield and estimated by the crop models can be seen in Table 4. For the observed data, regardless of the season, there was an increase in yield due to the increase in SAW, especially in region Z1 (p = 0.22 to 0.40). Less marked values were found for region Z3 (p = 0.05 to 0.27). For the Aquacrop model, the increase in yield was more marked, showing an increase in SAW in the less clayey regions: Z1 (p = 0.65 to 0.86) and Z2 (p = 0.53 to 0.60). According to van Lier et al. (2022), the Aquacrop model uses SAW values in calculating the water balance, these being widely used in modelling. On the other hand, the authors claim that it is necessary to adjust the input parameters using local measurement data, which allows the level of uncertainty to be reduced, corroborating the present study.

For the DSSAT model, there was an increase in yield in the clayey Z3 region for the increase in SAW (p = 0.27 to 0.50). In the sandier Z1 region, the yield showed a negative correlation with SAW (p = -0.49 to -0.44). Nogueira et al. (2001) used CROPGRO to predict the impact of changes in the soil water retention parameters when predicting simulated production in the soybean for different seasons. According to the authors, an increase in the values for water retention capacity in the model can reflect in increased performance when predicting yield. This may be related to the negative correlation between SAW and yield in Z1, which has lower values for CC and PWP compared to region Z3.

The variation in SAW at the management-zone level influenced the yield prediction of the crop models. The degree of correlation between the linear relationship of SAW and yield also varied from one season to another, albeit showing no specific pattern. As seen in this study, calibrating the SAW using actual field data is extremely important to ensure lower prediction uncertainty.

Table 4. Spearman's correlation between available soil water (SAW) and observed and estimated yields (CROPGRO) in each management zone (Z1, Z2 and Z3).

Management Zones (MZ)	Yield observed (Maps)			Aquacrop			CROPGRO/DSSAT		
	2018/19	2019/20	2020/21	2018/19	2019/20	2020/21	2018/19	2019/20	2020/21
Z1	0.64	0.40	0.22	0.65	0.81	0.86	-0.48	-0.49	-0.44
Z2	0.02	-0.07	0.07	0.53	0.59	0.60	-0.02	0.04	-0.03
Z3	0.10	0.27	0.05	-0.15	-0.09	-0.03	0.30	0.50	0.27

4.3.2. Soybean yield correlation and variation in MZs

Box-plots and statistical analyses for soybean yield in the MZs for the observed data and the crop models are shown in Figure 3. For the 2018/19 season, region Z2 had the highest values for observed yield (Z1 = 2773 kg ha^{-1}, Z2 = 3141 kg kg ha^{-1} and Z3 = 2663 kg kg ha^{-1}); for the same crop, soybean yield was overestimated by Aquacrop and underestimated by CROPGRO. Battisti et al. (2017), calibrating five crop models to estimate the soybean in the south of Brazil using Aquacrop and CROPGRO, found that both overestimated yield (~3%). Morales-Santos et al. (2023) also found that the Aquacrop model overestimated soybean yield under rainfed conditions. This overestimation by Aquacrop is a result of the low sensitivity of the model when estimating the transpiration rate of plants, which is related to the basal crop coefficient that must be properly calibrated (ADEBOYE et al., 2019). With Aquacrop, there were no significant differences in yield as a function of the MZs for the 2019/20 season, whereas the yield estimated by the CROPGRO method showed significant differences between the MZs (Z2 = 2999 kg kg ha^{-1}, Z3= 2632 kg kg ha^{-1} and Z1= 2546 kg kg ha^{-1}). Ejaz et al. (2022) state that the CROPGRO model shows sensitivity in predicting soybean yield as a function of variations in temperature, rainfall and CO2, which may explain the significant sensitivity between MZs compared to Aquacrop for the dry year. Furthermore, according to the same method, for the 2020/21 season, soybean yield in Z2 and Z3 was markedly similar (Z2 = 1950 kg kg ha^{-1} and Z3 = 2032 kg kg ha^{-1}), while in region Z1 the yield was low (Z1 = 1491 kg kg ha^{-1}). For this season, unlike CROPGRO, the Aquacrop model showed the same significant differences seen in the observed yield. Gimenez et al. (2017), evaluating soybean yield estimated by Aquacrop for different seasons also found significant differences for ANOVA. The Aquacrop model overestimated yield compared to the other models regardless of the MZ.

From the results, it can be seen that both models showed sensitivity in estimating soybean yield, affording significant differences between the MZs for the different seasons. Several factors should be considered when estimating yield, and each crop model has its own specific characteristics, with Aquacrop being more related to variations in the water, and CROPGRO to variations in the photoperiod. Mulazzani et al. (2022), evaluating the effect of soil compaction on the historical estimate of soybean yield by CROPGRO, stated that even using the model, interactions between the soil-plant system and the climate should be considered. According to the authors, although the model can be used to predict soybean yield, field experiments should be carried out to validate the estimates. This corroborates the present study and the importance of assessments at the management-zone level, which are still little used when employing crop models.

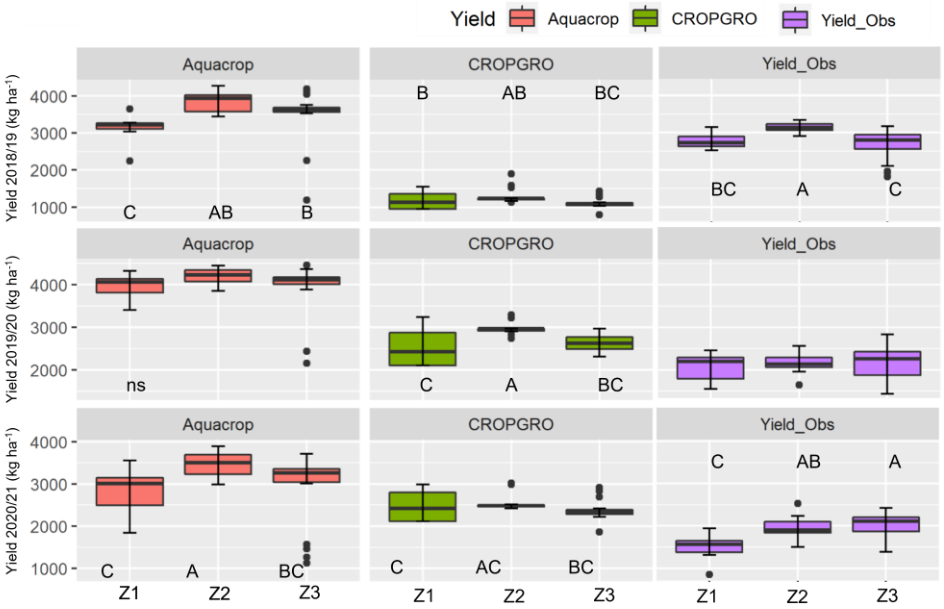


Figure 3. Box-plots of soybean yields estimated by Aquacrop and CROPGRO methods and observed by harvest maps for each of the management zones (Z1, Z2 and Z3). *Letters differ significantly between management zones by the Wilconxon test ($p<0.05$). ns There are no significant differences between management zones by the Wilcoxcon test ($p<0.05$).

4.3.3. Crop models performance to predict soybean yield based on MZs

O The performance (RMSE, MAE and R2) of Aquacrop and CROPGRO in estimating soybean yield based on the management zones are shown in Table 5. For the wet year (2018/19), Aquacrop showed superior performance (RMSE < 1172 kg kg ha^{-1}, MAE < 1093 kg kg ha^{-1}) to CROPGRO. For the dry years (2019/20 and 2020/21), CROPGRO performed better in estimating soybean production in each of the management zones (2019/20: RMSE < 865 kg kg ha^{-1}, MAE < 815 kg kg ha^{-1}; 2020/21: RMSE < 1137 kg kg ha^{-1}, MAE < 997 kg kg ha^{-1}). The values of R2, regardless of the model under evaluation, were mostly around 0.1. Battisti et al. (2017) also obtained similar errors for Aquacrop (RMSE= 536 to 2010 kg kg ha^{-1}; MAE = 458 to 1824 kg kg ha^{-1}) and CROPGRO (RMSE= 548 to 1397 kg kg ha^{-1}; MAE = 444 to 1061 kg kg ha^{-1}). According to the authors, the CROPGRO model performed better than Aquacrop.

Gimenez et al. (2017), using Aquacrop to estimate soybean yield for different seasons, found high variability in the results with high deviations (RMSE = 1010 kg kg ha^{-1}), corroborating the present study. According to the authors, the high level of error in the results are related to the high levels of water stress and low sensitivity of Aquacrop in representing transpiration. Morales-Santos et al. (2023) also stated that under rainfed conditions, the performance of Aquacrop was reduced due to the water limitations. For Battisti et al. (2017), crop models such as Aquacrop are limited in their predictions as the model does not consider the photoperiod, which can result in an increase in yield values, unlike models such as CROPGRO. The crop models are also sensitive to changes between MZs, even when the crop is not calibrated for actual field conditions.

Battisti et al. (2017) state that the use of a single crop model may be less efficient than when they are combined, in addition to highlighting the importance of calibrating the soil-related coefficients that affect the rate of water absorption by the roots. In this study, the variables related to soil texture, carbon and water were used as input for the crop models, which may have helped to achieve similar accuracy as in the studies by Battisti et al. (2017); Adeboye et al. (2019) and Gimenez et al. (2017).

Table 5. Performance of soybean yield estimates by Aquacrop and CROPGRO methods for the three management zones (Z1, Z3 and Z3) in the different harvests considered (2018/19, 2019/20 and 2020/21).

Year	MZ	Aquacrop			DSSAT/CROPGRO		
		RMSE	MAE	R2	RMSE	MAE	R2
		kg ha^{-1}			kg ha^{-1}		
2018/19	Z1	437	425	0.6	1615	1589	0.1
	Z2	769	719	0.2	1846	1834	0.1
	Z3	1172	1093	0.1	1625	1553	0.1
2019/20	Z1	1956	1910	0.1	771	628	0.1
	Z2	2039	2019	0.1	864	815	0.1
	Z3	1959	1857	0.1	638	495	0.1
2020/21	Z1	1446	1331	0.1	1137	997	0.1
	Z2	1574	1542	0.1	722	646	0.1
	Z3	1246	1188	0.1	492	407	0.1

4.4. Conclusions

The present study sought to investigate the efficiency of the Aquacrop and CROPGRO models in estimating soybean yield at the management-zone level as a function of the available water and soil variables. The crop models showed sensitivity in estimating yield for the different regions. There were differences between the seasons for both models, Aquacrop overestimating the results compared to CROPGRO; there were also differences in the results between wet and dry years. As a suggestion, future studies should be carried out on the spatial and temporal prediction of soybean yield under field conditions and in high resolution.

References

ALVARES, C; STAPE, J; SENTELHAS, P; DE MORAES GONÇALVES, J; SPAROVEK, G. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift,** vol. 22, nº 6, p. 711–728, 1 dez. 2013.

AKUMAGA U, GAO F, ANDERSON M, DULANEY WP, HOUBORG R, RUSS A, et al. Integration of Remote Sensing and Field Observations in Evaluating DSSAT Model for Estimating Maize and Soybean Growth and Yield in Maryland, USA. **Agronomy** 2023;13:1540. https://doi.org/10.3390/agronomy13061540.

AHMADPOUR A, FARHADI BANSOULEH B, AZARI A. Proposing a combined method for the estimation of spatial and temporal variation of crop water productivity under deficit irrigation scenarios based on the AquaCrop model. **Appl Water Sci** 2022;12. https://doi.org/10.1007/s13201-022-01666-8.

BATTISTI R, SENTELHAS PC, BOOTE KJ. Inter-comparison of performance of soybean crop simulation models and their ensemble in southern Brazil. **Field Crops Research** 2017;200:28–37. https://doi.org/10.1016/j.fcr.2016.10.004.

BATTISTI, R., & SENTELHAS, P. (2015). Drought tolerance of brazilian soybean cultivars simulated by a simple agrometeorological yield model. **Experimental Agriculture**, 51(2), 285-298. doi:10.1017/S0014479714000283

BALL, B.C., HUNTER, R., 1988. The determination of water release characteristics of soil cores at low suctions. **Geoderma** 43, 195–212.

CAMBARDELLA, C. A.; ELLIOTT, E. T. Particulate soil organic-matter changes across a grassland cultivation sequence. **Soil Science Society of America Journal**, v. 56, n. 03, p. 777-783, 1992.

DE JONG VAN LIER Q, LOGSDON SD, PINHEIRO EAR, GUBIANI PI. Plant available water**. Reference Module in Earth Systems and Environmental Sciences** 2022. https://doi.org/10.1016/b978-0-12-822974-3.00043-4.

EJAZ, M., ABBAS, G., FATIMA, Z. et al. Modelling Climate Uncertainty and Adaptations for Soybean-Based Cropping System. **Int. J. Plant Prod.** 16, 235–250 (2022). https://doi.org/10.1007/s42106-022-00190-8

ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

GIMÉNEZ L, PAREDES P, PEREIRA LS. Water Use and Yield of Soybean under Various Irrigation Regimes and Severe Water Stress. Application of AquaCrop and SIMDualKc Models. **Water** 2017;9:393. https://doi.org/10.3390/w9060393.

HOOGENBOOM, G., C.H. PORTER, V. SHELIA, K.J. BOOTE, U. SINGH, J.W. WHITE et al. 2017. **Decision Support System for Agrotechnology Transfer (DSSAT) Version 4.7.** https://DSSAT.net. DSSAT Foundation, Gainesville, FL.

MORALES-SANTOS A, GARCÍA-VILA M, NOLZ R. Assessment of the impact of irrigation management on soybean yield and water productivity in a subhumid environment. **Agricultural Water Management** 2023;284:108356. https://doi.org/10.1016/j.agwat.2023.108356.

MULAZZANI RP, GUBIANI PI, ZANON AJ, DRESCHER MS, SCHENATO RB, GIRARDELLO VC. Impact of soil compaction on 30-year soybean yield simulated with CROPGRO-DSSAT. **Agricultural Systems** 2022;203:103523. https://doi.org/10.1016/j.agsy.2022.103523.

RAES, D., STEDUTO, P., HSIAO, T. C. & FERERES, E. (2012). **AquaCrop Reference Manual**, AquaCrop version 4.0. Rome, Italy: FAO.

RUÍZ-NOGUEIRA B, BOOTE KJ, SAU F. Calibration and use of CROPGRO-soybean model for improving soybean management under rainfed conditions [Internet]. Vol. 68, **Agricultural Systems**. Elsevier BV; 2001. p. 151–73. Available from: http://dx.doi.org/10.1016/S0308-521X(01)00008-7

SINGH RS, SINGH KK, GOHAIN GB. Simulating crop yield using the DSSAT v4.7-CROPGRO-soyabean model with gridded weather and soil data. **Model Earth Syst Environ**. 2023. https://doi.org/10.1007/s40808-023-01807-1.

STACKHOUSE PW, et al. (2017). **Prediction of worldwide energy resource (POWER), agroclimatology methodology,** (1° Latitude by 1° Longitude Spatial Resolution). Acesso em 25 de novembro de 2019. Link: https://power.larc.nasa.gov/documents/Agroclimatology_Methodology.pdf. Accessed 02 January 2020.

STEDUTO P, HSIAO TC, FERERES E. On the conservative behavior of biomass water productivity. **Irrig Sci** 2007;25:189–207. https://doi.org/10.1007/s00271-007-0064-1.

SOIL SURVEY STAFF, 2014 **Soil Survey Staff Keys to Soil Taxonomy** (12th), USDA (2014), pp. 1-410

SPEKKEN, M., ANSELMI, A.A., MOLIN, J.P. A simple method for filtering spatial data. **In: 9th European conference on precision agriculture**, 2013, Lleida.

TORSONI, G.B., DE OLIVEIRA APARECIDO, L.E., DOS SANTOS, G.M. et al. Soybean yield prediction by machine learning and climate. **Theor Appl Climatol** 151, 1709–1725 (2023). https://doi.org/10.1007/s00704-022-04341-9

THORNTHWAITE, C.W.; MATHER, J.R. **The water balance Centerton,** NJ: Drexel Institute of Technology - Laboratory of Climatology, 1955. 104p. (Publications in Climatology, vol. VIII, n.1)

ZHANG Z, YU D, SHI X, WARNER E, REN H, SUN W, et al. Application of categorical information in the spatial prediction of soil organic carbon in the red soil area of China. **Soil Science and Plant Nutrition** 2010;56:307–18. https://doi.org/10.1111/j.1747-0765.2010.00457.x.

5. FINAL CONSIDERATIONS

One of the main challenges to increasing yield is predicting the factors that influence the production cycle. A gap in knowledge that led to this study was the prediction of soybean yield considering spatial variability. The water balance or water cycle of the soil-plant-atmosphere relationship was one of the fundamental pillars for the study, and at field level was only possible through the use of remote sensing and sensors-machine data. Machine learning was useful in analysing the data and allowed estimates close to the actual field results. Among the benefits of the results of this study are support for decision-making during the planning stages, ensuring inputs are rationalised and reducing environmental impact.

As a result of the first chapter, it was seen that the management of agricultural areas should not be carried out homogeneously, given the influence of the spatial-temporal variability of the physical factors of the soil on soybean yield, as in the case of the sandiest region of the study area. The use of ECa surfaces and the historical yield-map series ensured the MZs were generated, and the regions classified based on their most prominent characteristics. Defining the proper number of MZ groups was helped by the use of principal component analysis, ensuring that the accuracy indices converged for three or four regions. One of the main findings in this chapter was understanding the specific characteristics of each region together with the transitory aspects, which are due to textural factors and soil formation. As a practical result, specific management practices can be suggested based on the use of cover crops that have more-aggressive roots in regions of lower yield. This can be one way of increasing organic matter and thereby water storage capacity by reducing evaporation. Use of the classification tree helped to validate the differences between the MZs based on soil water retention and textural variation in each region. The use of digital maps of water retention and soil texture can be one alternative together with yield maps for generating MZs. Although methods for generating MZs are easily found in the literature, predicting soybean yield from these regions remains a challenge: the present study also sought to understand and address this need.

As a result of Chapter 2, it was possible to demonstrate the challenges of digitally mapping soil water retention and especially its variability at the MZ level. This once again justified the importance of managing agricultural areas considering spatial-temporal variability. This chapter showed how soil water availability also varies depending on the texture of the soil, where the degree of water retention correlated differently in each region. In situations where information on field capacity and the permanent wilting point was not available, the impact of generating digital soil water retention maps based on primary soil data such as texture, apparent density and organic matter was evaluated in the different fractions and layers, highlighting the specific characteristics of each region, and of the factors that differed between the regions. The multivariate principal component analysis also demonstrated the relationship between the variability of the vegetation indices and the water balance in each MZ. The region characterised as having low productive potential showed greater correlation with the factors of water deficit and water surplus in the soil, while the other regions showed a correlation with the vegetation indices and crop development. The interactions seen in the MZs occurred mainly during certain phenological stages of the soybean and were different for wet and dry years.

Use of the Random Forest method in this study resulted in high accuracy compared to the conventional MLR method. However, seasonality and the number of digital maps were essential for increased quality, requiring the use of three seasons of information to guarantee the quality of the predictions. Further studies are needed in order to use the technique, taking observed yield data into consideration. In this study, the sample yield database was small in relation to the resolution of the harvest maps. The learning performance of the model was therefore poor due to yield being measured in the field; on the other hand, the results showed high stability. These results show the importance of the quality of the input data and of the available historical information. The end of the second chapter demonstrated the importance of vegetation indices for inferring plant development. The use of NIR reflectance surfaces proved to be an alternative to the use of NDVI, widely used in field evaluations. Further research should be developed to investigate the potential of this information.

Chapter 3 explored the performance of the crop models in estimating soybean yield at the management-zone level and as a function of the variations in soil water. The crop models showed different behavior for the different management zones and crops under evaluation. The performance of the Aquacrop model was higher for the wet year, while CROPGRO showed better performance for the dry years. Aquacrop tended to overestimate yield in the management zones, while with CROPGRO, yield was underestimated under the same conditions. The use of crop models with simulations of virtual soil profiles related to water, texture and carbon retention supported the importance of considering the differences between regions at the plot level.

Finally, this study contributed to an understanding of the use of learning techniques and crop models at the field level, and recognition of how soil, plant and climate data might support decision-making. Another aspect is a combination of techniques such as digital mapping, clustering, and predictive models for crop yield as an alternative for generating actions and understanding specific characteristics at the local level. The joint use of this information is part of the digital transformation in agriculture and helps to increase yield by rationalising inputs based on the productive potential of each MZ. One suggestion for further studies is the possibility of understanding whether current predictive models of crop yield are superior to machine learning methods that use sensing data. From the point of view of process automation, it is suggested that future studies investigate the creation of software that combines all these techniques and that allows decision-making on the part of technicians or farmers.