Integrating Data Layers with Machine Learning to Predict Yield for Irrigated Grain Crops within Management Zones

L. G. G. Sterle[1], R. Canal Filho[1], V. A. Ferraz[1], M. S. Gelain[1], J. P. Molin[1]

[1]Luiz de Queiroz College of Agriculture (ESALQ), University of São Paulo (USP),
Av. Pádua Dias, 11, 13418-900, Piracicaba, São Paulo, Brazil.

A paper from the Proceedings of the
17th International Conference on Precision Agriculture and
11th Brazilian Congress on Precision and Digital Agriculture
13-15 July 2026
Porto Alegre, Brazil

Abstract.

Precision agriculture requires spatially explicit data for site-specific crop management, yet prediction models frequently assume uniform field conditions, overlooking substantial within-field variability driven by soil properties, terrain, and production history. This study evaluated whether management zone-based machine learning improves grain yield prediction over whole-field approaches in a 52.3 ha irrigated field in Itaí, São Paulo, Brazil. Apparent soil electrical conductivity at two depths (ECa75 and ECa150), elevation, and slope were used as stable delineation variables. Raw yield data from six consecutive harvests, sorghum (2020), soybean (2021, 2023, and 2025), wheat (2021), and barley (2023), were filtered using MapFilter 2.0 and interpolated to a 3 m grid by ordinary kriging. Principal Component Analysis retained three components explaining 97.6% of total variance, serving as input for fuzzy c-means clustering across cluster numbers (k = 2-6) and fuzziness exponents (m = 1.5, 2.0, and 2.5). The optimal solution was selected by maximizing a composite normalized score from five indices: Calinski-Harabasz, Davies-Bouldin, Silhouette coefficient, Fuzzy Partition Coefficient, and Variance Reduction. Four management zones were identified (k = 4, m = 1.5), containing 38.0%, 2.8%, 25.7%, and 33.5% of total pixels. Six modelling scenarios were built by progressively adding historical yield maps to stable predictors, from a soil-features-only baseline (SF) to five accumulated maps (SF+5). Four regression algorithms were evaluated: Random Forest (RF), XGBoost (XGB), Support Vector Regression (SVR), and Linear Regression (LR). Spatial GroupKFold cross-validation with coordinate-based blocking minimized autocorrelation bias; performance was assessed using out-of-fold R^2, RMSE, MAE, and normalized RMSE. Historical yield maps were the most impactful addition: RF whole-field R^2 increased from 0.623 (SF, sorghum 2020) to 0.841 (SF+5, soybean 2025), while SVR achieved the highest accuracy under SF+5 (R^2 = 0.877, nRMSE = 1.72%). The management zone strategy improved XGB performance in barley 2023 (R^2 from 0.721 to 0.765) and reduced prediction errors for wheat across zones, while whole-field SVR outperformed zone-specific models for soybean 2025, indicating scenario-dependent benefits of spatial stratification. LR consistently underperformed across all scenarios, confirming the nonlinear nature of yield response. Feature importance analysis revealed a progressive shift from terrain and ECa dominance in early scenarios toward historical yield maps as primary predictors in later ones, with elevation retaining higher importance within zones than in whole-field models. These findings confirm that integrating stable environmental variables, accumulated yield history, fuzzy management zones, and spatially validated machine learning provides a robust framework for site-specific yield prediction in irrigated grain systems.

Keywords. Artificial Intelligence, Fuzzy C-Means, Apparent Soil Electrical Conductivity, Terrain Attributes, Spatiotemporal Variability, Spatial cross-validation, Random Forest, XGBoost, Support Vector Regression

Introduction

Global food demand continues to increase while agricultural expansion becomes progressively constrained by environmental, economic, and climatic limitations. Under this scenario, precision agriculture (PA) has emerged as a strategy to improve yield, optimize resource allocation, and reduce environmental impacts through site-specific crop management (Gebbers and Adamchuk, 2010; Mulla, 2013). The fundamental principle of PA is that agricultural fields are spatially heterogeneous, meaning that soil properties, water availability, topography, and crop performance vary considerably within short distances. Consequently, uniform management practices often fail to maximize productivity and input-use efficiency across entire fields.

Yield maps generated from combine harvesters equipped with Global Navigation Satellite Systems (GNSS) and flow sensors are among the most valuable datasets in PA because they integrate the cumulative effects of soil, climate, management, and landscape conditions over time (Blackmore, 1999; Whelan and McBratney, 2000). Multi-year yield datasets are particularly relevant because they reveal stable spatial patterns associated with persistent soil and terrain characteristics rather than temporary environmental fluctuations (Maestrini and Basso, 2018). However, raw yield monitor data frequently contain substantial noise caused by sensor delays, GPS inaccuracies, machine acceleration and deceleration, partial header filling, and operational errors during harvesting. These inconsistencies can severely compromise subsequent spatial analyses and predictive modelling if not properly filtered (Sudduth and Drummond, 2007; Vega et al., 2019).

Recent advances in machine learning (ML) have significantly expanded the capacity to predict crop yield using large and complex agricultural datasets. Unlike traditional statistical approaches, ML algorithms can model non-linear relationships and high-order interactions among soil, terrain, and historical yield variables without requiring strong assumptions regarding data distribution (Liakos et al., 2018; van Klompenburg et al., 2020). Among the most widely adopted algorithms in agricultural prediction studies are Random Forest (RF), Extreme Gradient Boosting (XGBoost, XGB), and Support Vector Regression (SVR). Random Forest is an ensemble-based algorithm capable of handling noisy and high-dimensional datasets while maintaining robust predictive performance (Breiman, 2001). XGBoost has demonstrated excellent predictive capacity due to its gradient boosting framework, computational efficiency, and ability to capture complex non-linear patterns (Chen and Guestrin, 2016). Support Vector Regression has also been successfully applied in crop prediction studies because of its effectiveness in high-dimensional feature spaces and its flexibility through kernel functions (Smola and Schölkopf, 2004).

Despite the increasing use of ML approaches in agriculture, many yield prediction studies still assume that the entire field behaves as a homogeneous production unit (Liakos et al., 2018; van Klompenburg et al., 2020). In practice, agricultural fields commonly present substantial spatial variability associated with soil texture, elevation, drainage conditions, compaction, and nutrient distribution. Ignoring this heterogeneity may reduce model accuracy because a single predictive model must simultaneously represent multiple environmental conditions and yield responses.

Management zones (MZs) have been proposed as a practical framework for addressing within-field variability in PA. MZs are subregions of a field that exhibit relatively homogeneous combinations of soil, terrain, and yield characteristics and therefore may respond similarly to management interventions (Fridgen et al., 2004; Moral et al., 2010). By partitioning fields into more homogeneous areas, MZ-based approaches can support localized agronomic recommendations and improve predictive modelling performance (Maestrini and Basso, 2018).

Among the variables commonly used for MZs delineation, apparent soil electrical conductivity (ECa) has received considerable attention due to its association with several soil properties, including clay content, moisture, salinity, cation exchange capacity, and organic matter. In addition, ECa exhibits temporal stability, meaning that repeated measurements are not required on an annual basis. (Corwin and Lesch, 2005; Doolittle and Brevik, 2014). Terrain attributes such as elevation and slope are also important because they directly influence water redistribution, erosion dynamics, soil formation processes, and crop development (Moore et al., 1993; Schepers et al., 2004). These variables are particularly suitable for delineating MZs because they are relatively stable over time and represent persistent patterns of field variability.

Several clustering approaches have been proposed for MZ delineation, with fuzzy c-means (FCM) clustering being one of the most widely used methods in PA. Unlike hard clustering approaches, such as k-means, FCM allows gradual transitions among zones by assigning membership probabilities to each observation rather than imposing rigid boundaries (Bezdek, 1981). This characteristic is particularly important in agricultural systems, where soil transitions are typically continuous rather than abrupt. Principal Component Analysis (PCA) is frequently combined with FCM to reduce multicollinearity among variables and condense the main variability structure into a smaller number of orthogonal components before clustering (Jolliffe, 2002; Córdoba et al., 2013).

Although numerous studies have evaluated MZs delineation and ML-based yield prediction independently, relatively few investigations have integrated both approaches under spatial validation frameworks using multi-year datasets. Moreover, the contribution of historical yield maps as predictor variables needs to be more explored. Historical yields integrate the cumulative influence of environmental conditions, soil variability, and management practices over time and may contain information not fully represented by static soil or terrain attributes alone (Filippi et al., 2019; Maestrini and Basso, 2018; Wei et al., 2026). Understanding how successive yield maps improve prediction performance is essential for designing practical and scalable predictive systems in PA.

Another critical issue concerns model validation. Random cross-validation procedures often overestimate predictive performance in spatial datasets because neighbouring observations tend to be spatially autocorrelated (Roberts et al., 2017; Habibi et al., 2024). Spatial cross-validation strategies are therefore necessary to produce more realistic assessments of model generalization capacity under operational field conditions (Filippi et al., 2025).

Therefore, the objective of this study was to evaluate whether MZs-based machine learning models improve grain yield prediction compared with conventional whole-field models in a center-pivot irrigated grain production system in Brazil. Specifically, the study aimed to: (i) delineate MZs using stable soil and terrain variables through PCA and FCM clustering; (ii) compare whole-field and MZs-based prediction strategies using RF, XGB, SVR, and Linear Regression models; and (iii) evaluate the progressive contribution of historical yield maps as predictor variables across six crop scenarios between 2020 and 2025.

Material and Methods

The study was conducted in a 52.3 ha center-pivot irrigated commercial field located in Itaí, São Paulo, Brazil (23° 32' 49" S and 48° 55' 11" W), situated in the Alto Paranapanema region. According to Climate Data, 2024. The climate in the area is classified as Cfa (humid subtropical) according to Köppen and Geiger (1928). The soils in the studied fields are classified as Rhodic Eutrudox (USDA, 2014). The slope is predominantly classified as flat (< 3%), with an average elevation of 675 meters (EMBRAPA, 1979) (Figure 1).

The production system operates under no-tillage management with crop succession involving soybean, sorghum, wheat, and barley between 2020 and 2025. Six harvest datasets were evaluated, including sorghum (2020), soybean (2021, 2023, and 2025), wheat (2021), and barley (2023). Yield data were collected using a commercial yield monitoring system (9.1 m cutting width, 1 Hz frequency) coupled with GNSS positioning, resulting in density from 673 to 1087 points ha^{-1} Soil apparent electrical conductivity (ECa) was measured using the EMP-400 EMI sensor (Geophysical Survey Systems, Inc.), at two depths (0 0.75 m and 0 1.50 m), with RTK-GNSS positioning. Data were collected in March 2017 (4878 points). Elevation and slope were derived from SRTM DEM using the OpenTopography QGIS plugin (Win, 2024).

Raw yield data were filtered using MapFilter 2.0 (Maldaner et al., 2021) to remove outliers through statistical (median-based) and spatial filtering. For yield, accepted variation (v) was 50% (statistical) and 75% (local), with 91 m spatial dependence. Upper (LimS) and lower (LimI) limits were established using the median for filtering, defined as:

$$LimS = MK + MK \cdot v \tag{1}$$

$$LimI = MK - MK \cdot v \tag{2}$$

Where LimS: upper limit; LimI: lower limit; MK: median of all values in the dataset; v: maximum accepted variation for the median.

Filtered data were interpolated by ordinary kriging using the Smart-Map QGIS plugin (Pereira et al., 2022), generating 3 m resolution rasters. Model selection was based on the highest R² and lowest RMSE from cross-validation.

A 3 m regular grid was used to extract raster values for ECa, yield, and terrain attributes. Exploratory analysis included mean, standard deviation, coefficient of variation, minimum and maximum, skewness, and kurtosis. Normality was tested using the Anderson Darling test at 5%. Spearman's rank correlation was used to assess monotonic relationships between variables, as it is non-parametric and does not assume normality or linearity and classified following Rumsey (2023).

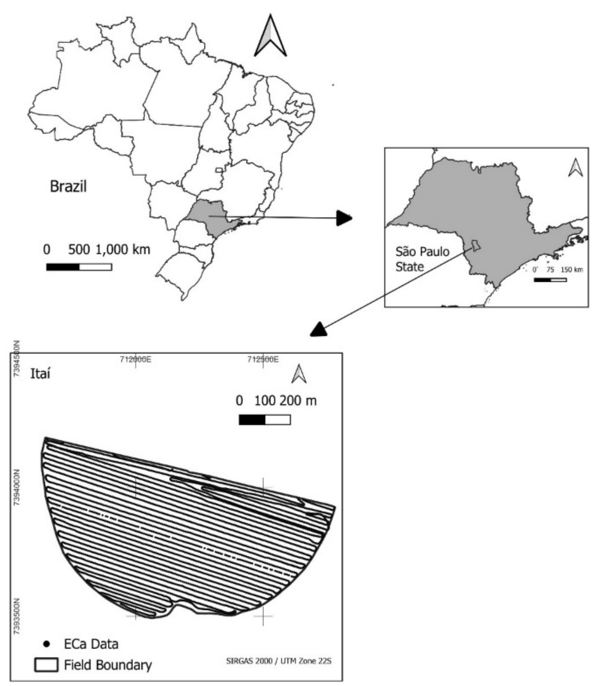

Fig. 1 Study area location and distribution of ECa sampling rows.

Four stable environmental variables were used for MZ delineation: ECa measured at 0–0.75 m depth (ECa75), ECa conductivity measured at 0–1.50 m depth (ECa150), elevation, and slope. These variables were selected because they represent relatively stable soil and terrain attributes

associated with long-term crop productivity patterns. Before multivariate analysis, all variables were standardized to zero mean and unit variance using z-score normalization.
Principal Component Analysis (PCA) was applied to the standardized variables to reduce dimensionality and minimize multicollinearity effects. The minimum number of principal components required to explain at least 80% of cumulative variance was retained for subsequent clustering analyses. The retained PCA scores were used as input variables for fuzzy c-means (FCM) clustering.
MZs were delineated using a grid-search strategy combining fuzziness exponent values ($m = 1.5$, 2.0, and 2.5) and cluster numbers ranging from $k = 2$ to $k = 6$. Each clustering solution was evaluated using five complementary validation metrics: Calinski-Harabasz index (CH), Davies-Bouldin index (DB), Silhouette coefficient (SC), Fuzzy Partition Coefficient (FPC), and Variance Reduction (VR). All metrics were normalized to a 0–1 scale to allow combined evaluation. Because lower Davies-Bouldin values indicate superior clustering performance, this metric was inverted during normalization. The final clustering solution was selected based on the highest average normalized score across all validation metrics.
Six predictive modelling scenarios were evaluated to investigate the contribution of historical yield maps as predictor variables. The first scenario used only stable soil and terrain variables (SF). The remaining scenarios sequentially incorporated historical yield maps as additional predictors, resulting in the following configurations: SF, SF + one yield map, SF + two yield maps, SF + three yield maps, SF + four yield maps, and SF + five yield maps, presenting in table 1. This cumulative strategy simulated the progressive accumulation of historical information commonly available in precision agriculture systems.

Table 1. Modelling scenarios: target crop, predictor features, and number of historical yield maps.

Scenario	*Target crop*	*Predictor features*	*Historical yield maps (n)*
SF	Sorghum 2020	ECa75, ECa150, Elevation, Slope	0 (soil features)
SF+1	Soybean 2021	SF + Sorghum 2020	1
SF+2	Wheat 2021	SF + Sorghum 2020 + Soybean 2021	2
SF+3	Soybean 2023	SF + Sorghum 2020 + Soybean 2021 + Wheat 2021	3
SF+4	Barley 2023	SF + Sorghum 2020 + Soybean 2021 + Wheat 2021 + Soybean 2023	4
SF+5	Soybean 2025	SF + all five preceding crops	5

Four regression algorithms were evaluated: Random Forest (RF), XGBoost (XGB), Support Vector Regression (SVR), and Linear Regression (LR). RF models were implemented using 300 trees with a minimum leaf size of three observations. XGB models were implemented using a learning rate of 0.05, maximum tree depth of five, subsampling rate of 0.8, and column subsampling rate of 0.8. SVR models employed a radial basis function kernel with standardized predictors, while Linear Regression was included as a baseline linear model. Hyperparameter tuning using exhaustive GridSearchCV was not performed because of the substantial computational burden associated with the experimental design, which included multiple algorithms, six historical-yield scenarios, spatial GroupKFold cross-validation, and zone-specific training routines. Instead, hyperparameters were predefined according to values consistently reported in the literature for agricultural yield prediction applications. This strategy substantially reduced computational time while preserving methodological consistency and reproducibility across scenarios. All analyses were performed in Python using the scikit-learn, xgboost, and scikit-fuzzy libraries.
Model validation was performed using spatial cross-validation to minimize bias associated with spatial autocorrelation. Spatial groups were generated using k-means clustering based on geographic coordinates. Subsequently, GroupKFold cross-validation was applied using five folds, ensuring that spatially adjacent observations remained within the same fold during model training and testing. This approach provided a more realistic estimate of model generalization capacity under operational conditions.
Two prediction strategies were compared. In the whole-field strategy, a single model was calibrated using all observations within the training dataset regardless of MZs membership. In the

MZs strategy, independent models were calibrated separately for each MZ using only observations belonging to the corresponding zone. When a given zone presented insufficient observations for local model calibration, predictions were generated using the whole-field model trained within the same fold.
Model performance was evaluated using out-of-fold predictions generated during spatial cross-validation. Performance metrics included coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and normalized RMSE relative to the observed mean (nRMSE%). Feature importance analyses were additionally performed for RF and XGB models using impurity-based importance scores averaged across folds.

Results and Discussion

Data filtering was a necessary step to remove inconsistencies and outliers commonly present in raw yield monitor data. The reduction in the total number of yield points ranged from 4.8% (Soybean 2023) to 36.1% (Sorghum 2020), reflecting differences related to field conditions, crop types, and harvest years. Most datasets showed reductions between 5% and 15%, with sorghum experiencing the most substantial data loss due to the greater variability observed in this crop. Data filtering consistently improved overall data quality, resulting in a substantial reduction in standard deviation, by approximately 45-65% across most datasets, with the coefficient of variation decreasing to below 0.25 in all filtered datasets. The filtering process also increased mean yield values by 15-40% across the datasets, indicating effective removal of erroneous low-yield points.
The analyzed attributes exhibit distinct structures of spatial dependence (Table 2), as reflected in the fitted semivariogram models and the DSD (Degree of Spatial Dependence) values. The electrical conductivity measurements (ECa75 and ECa150) show strong spatial dependence, particularly ECa150, indicating well-defined spatial continuity and strong potential for use in targeted soil sampling. Crop yield variables display more heterogeneous behavior: Sorghum 2020, Wheat 2021, and Soybean 2021 present strong or moderate spatial dependence with considerable ranges (295 to 554 m), suggesting that soil properties and management factors influence the spatial structure of these yields. In contrast, Soybean 2023 and Soybean 2025 show moderate dependence and a higher C0/(C0 + C1) ratio, indicating greater random variability or the influence of unaccounted factors. The large ranges observed (445 to 692 m) suggest that spatial variability occurs over broad distances within the field. Overall, the results indicate that soil physical–chemical attributes tend to exhibit stronger spatial dependence than yield, reinforcing their value as effective indicators for delineating MZs (Table 2).

Table 2. Semivariogram models and fitting parameters used for spatial interpolation in SmartMap software (version 1.4.2) for irrigated and rainfed fields.

Attribute	Model	Max_dist (m)	lag	C0	C0 + C1	Range (m)	RMSE	Radius	DSD	DSD Class
ECa_{75} (mS m^{-1})	Linear with sill	450.00	6.72	8.83	32.26	317.21	39.22	1160.48	27.36%	Moderate
ECa_{150} (mS m^{-1})	Linear with sill	450.00	6.72	25.91	120.34	342.75	445.14	1160.48	21.53%	Strong
Sorghum 2020 (Mg ha^{-1})	Exponential	650.00	12.50	0.22	1.73	295.49	0.03	295.49	12.75%	Strong
Soybean 2021 (Mg ha^{-1})	Exponential	699.56	12.51	0.16	0.36	554.34	0.01	554.34	43.41%	Moderate
Wheat 2021 (Mg ha^{-1})	Linear with sill	692.81	12.50	0.20	1.43	539.87	0.06	539.87	14.21%	Strong
Soybean 2023 (Mg ha^{-1})	Spheric	698.61	12.50	0.12	0.23	692.14	0.00	692.14	51.09%	Moderate
Barley 2023 (Mg ha^{-1})	Exponential	699.22	12.51	0.23	0.72	445.32	0.02	445.32	31.57%	Moderate
Soybean 2025 (Mg ha^{-1})	Spheric	640.00	12.46	0.19	0.41	637.51	0.00	637.51	47.17%	Moderate

* Max_dist: maximum distance between point pairs; C0: nugget effect; C1: structural component; C0 + C1: sill.

The descriptive statistics of the dataset (Table 3) show considerable variability among the variables. Soil electrical conductivity (ECa75 and ECa150) and elevation presented low coefficients of variation (CV < 11%), while slope exhibited high variability (CV = 68.8%). Crop yields varied moderately, with soybean 2025 showing the lowest CV (4.9%) and sorghum 2020 and wheat 2021 the highest (near to 23%). Skewness and kurtosis values indicate asymmetrical and heavy tails for some variables, particularly slope. All variables were classified as non-normal according to the Anderson-Darling test, justifying the use of non-parametric analyses or machine learning approaches.

Table 3. Descriptive statistics of variables.

	Mean	*SD*	*Min*	*Max*	*CV (%)*	*S*	*K*
Slope (%)	2.26	1.56	0.31	13.61	68.76	3.46	16.76
Elevation (m)	674.70	2.98	666.23	682.20	0.44	-0.09	-0.14
ECa_{150} ($mS\ m^{-1}$)	100.22	10.48	78.36	130.97	10.45	-0.24	-0.54
ECa_{75} ($mS\ m^{-1}$)	57.80	5.49	46.40	72.95	9.50	-0.43	-0.57
Soybean 2025 ($Mg\ ha^{-1}$)	7.95	0.39	6.25	8.74	4.90	-1.07	1.48
Soybean 2023 ($Mg\ ha^{-1}$)	5.74	0.26	4.71	6.35	4.56	-0.97	0.98
Soybean 2021 ($Mg\ ha^{-1}$)	7.46	0.42	5.88	8.39	5.64	-0.62	0.27
Sorghum 2020 ($Mg\ ha^{-1}$)	5.18	1.21	2.74	7.69	23.25	0.07	-1.07
Wheat 2021 ($Mg\ ha^{-1}$)	4.19	0.93	2.33	6.02	22.28	0.13	-0.92
Barley 2023 ($Mg\ ha^{-1}$)	8.66	0.64	6.03	11.23	7.37	0.35	1.93

* Min: minimum;Mean: average;Max: maximum; SD: standard deviation; S: skewness; K: kurtosis.

Based on the correlation analysis, several relationships between terrain attributes, ECa, and crop yields were found to be statistically significant ($p < 0.001$) but demonstrated varying strengths of association. The strongest correlation was observed between the two ECa measurements (ECa150 vs ECa75, $r = 0.82$), indicating a strong linear relationship. Moderate positive correlations were identified between elevation and soybean yield in 2025 ($r = 0.51$) and between soybean yields across different years, particularly between 2025 and 2023 ($r = 0.56$).
Most variable pairs showed either weak or no meaningful linear relationships despite statistical significance. Notably, slope demonstrated weak negative correlations with soybean yields across multiple years ($r = -0.32$ to -0.41), while elevation showed weak to moderate positive associations with several crops. Most relationships between crop yields and soil ECa measurements, as well as among different crop types across seasons, were not statistically significant or exhibited only negligible correlations, suggesting limited linear dependence between these agricultural variables in the studied system (Figure 2).

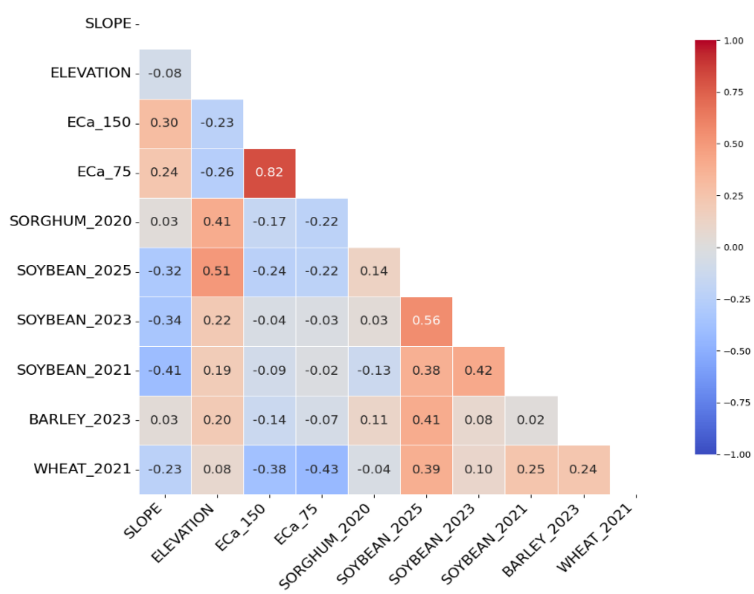

Fig 2 Spearman's correlation between variables.

The FCM clustering procedure identified an optimal configuration with four MZs ($k = 4$) and a fuzziness coefficient $m = 1.5$, based on the highest average normalized score among the CH, SC, DB, FPC, and VR metrics (Figure 3), in agreement with previous work using fuzzy clustering for MZ delineation (Bansod and Pandey, 2013; Saleh and Belal, 2014; Sterle and Molin, 2025). The PCA retained three principal components, which explained 97.6% of the total variance, indicating that the selected stable variables effectively captured the spatial variability of the field, as also reported in other PCA-based zoning studies (Bansod and Pandey, 2013; Saleh and Belal, 2014). PC1 was mainly associated with ECa variables, whereas PC2 and PC3 were more strongly influenced by terrain attributes such as slope and elevation, consistent with findings that ECa and topography are key drivers of within-field variability (Saleh and Belal, 2014; Sterle and Molin, 2025).

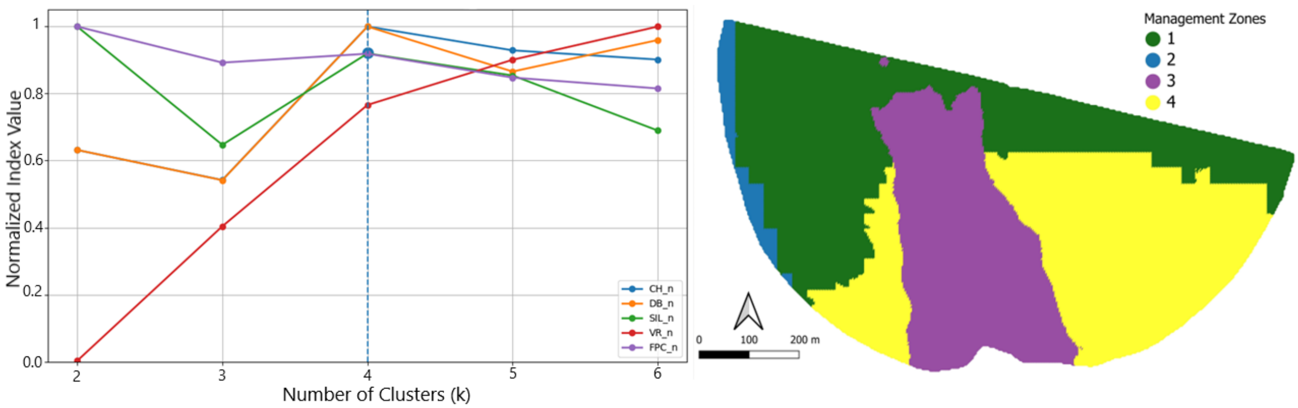

Fig 3 Normalized clustering validation indices for the FCM solutions using a fuzziness exponent of m = 1.5.

ML performance varied according to crop season, input scenario, and modelling strategy, which is in line with other PA and yield-modelling studies (Burdett and Wellen, 2022; Filippi et al., 2025). Models trained exclusively with stable soil and terrain attributes (SF scenario) achieved moderate predictive performance for sorghum yield, with R^2 values ranging from 0.16 for LR to 0.62 for RF under whole-field modelling, like the superiority of RF over linear models reported by Burdett and Wellen (2022). The incorporation of historical yield maps progressively improved prediction accuracy across all algorithms and crops, in agreement with multi-year and temporal-yield studies (Filippi et al., 2019; Morales and Villalobos, 2023).

Figure 4 shows that the highest predictive performance was observed for soybean yield in 2025, where models incorporated five historical yield maps (SF+5). Under whole-field modelling, SVR achieved the best overall performance with $R^2 = 0.877$ and nRMSE = 1.72%, followed closely by RF ($R^2 = 0.841$) and XGB ($R^2 = 0.836$). These results indicate that temporal yield history substantially increased the predictive capacity of ML models, an effect also highlighted by Filippi et al. (2019) and Cao et al. (2021).

Differences between whole-field and MZ approaches depended on the crop and algorithm. For barley in 2023, the MZs strategy improved XGB performance from $R^2 = 0.721$ in the whole-field model to $R^2 = 0.765$ in the MZs approach. Similar gains were observed for wheat, where MZ models reduced RMSE values in several zones, reflecting how MZs can reduce within-zone variance and aid prediction (Gavioli et al., 2019; Sterle and Molin, 2025). However, improvements were not universal. In soybean 2025, whole-field SVR outperformed the zone-based strategy, suggesting that the benefit of MZ delineation depends on the degree of spatial heterogeneity and temporal yield stability (Maestrini and Basso, 2018; Gavioli et al., 2019).

Linear Regression consistently showed the lowest predictive performance across all scenarios, with several zone-specific models producing near-zero or negative R^2 values. This behaviour indicates that yield response was predominantly nonlinear and better captured by ensemble and kernel-based approaches, as widely documented in yield-prediction literature (Burdett and Wellen, 2022; Jabed and Azmi Murad, 2024).

SVR models achieved competitive predictive accuracy but required substantially longer computational times, frequently exceeding 300 s per cross-validation routine, whereas XGB produced similar predictive performance with execution times below 5 s in most scenarios. In practical PA applications, this difference is relevant because computational efficiency directly affects operational scalability, a point also raised in comparative ML studies (Jabed and Azmi Murad, 2024; Joshi et al., 2023; Badshah et al., 2024).

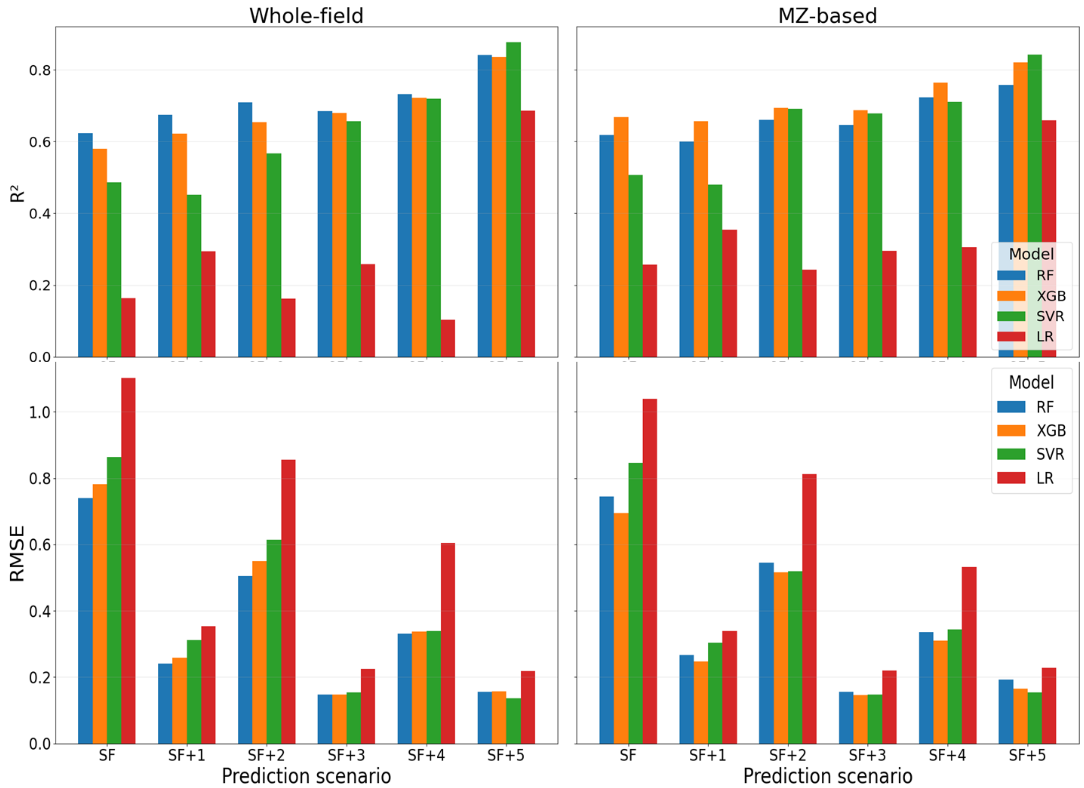

Fig 4 Comparison of prediction metrics between whole-field and MZ-based modelling approaches. Top panels show R^2 values and bottom panels show RMSE values for different prediction scenarios using RF, XGB, SVR, and LR models.

Feature importance analysis revealed differences between whole-field and MZ-based modelling strategies (Figure 5). In the whole-field approach, elevation consistently appeared among the most influential predictors in both RF and XGB models, suggesting that broad-scale terrain effects strongly controlled spatial yield variability. Historical yield layers from soybean seasons also showed high importance, reinforcing the temporal stability importance in grain production systems.

Under the MZ-based strategy, predictor importance became more redistributed among variables. In RF models, barley yield from 2023 substantially increased in importance within the MZ framework, while XGB models showed greater contribution from ECa and wheat-related layers. This behaviour suggests that spatial stratification reduced within-zone variability and allowed localized relationships between soil, terrain, and crop response to emerge more clearly. Similar effects have been reported in studies showing that MZ delineation can enhance the sensitivity of ML models to site-specific patterns by reducing background spatial noise (Gavioli et al., 2019; Sterle and Molin, 2025).

The results indicate that MZ-based modelling not only affected prediction accuracy but also altered how ML algorithms interpreted the relative importance of environmental and historical yield variables across the field.

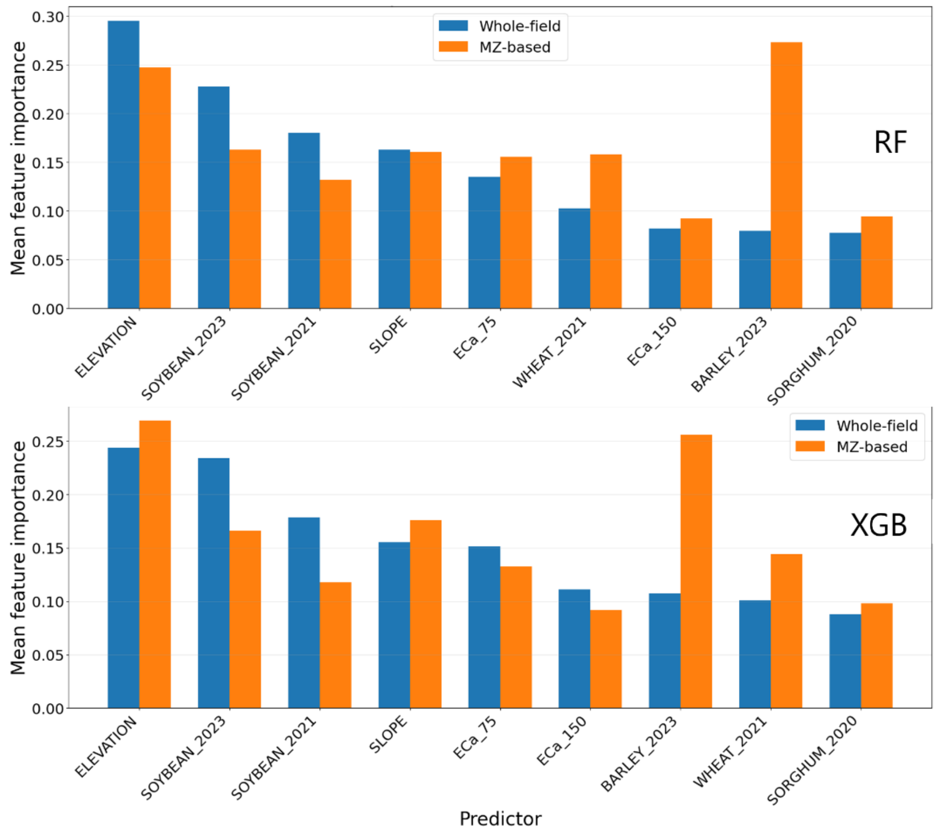

Figure 5. Feature importance obtained from RF and XGB models comparing whole-field and MZ-based approaches. Mean feature importance values indicate the relative contribution of terrain attributes, ECa, and historical crop yield layers to model predictions. The MZ-based approach increased the importance of specific spatial predictors, particularly barley and wheat yield layers, suggesting that stratification by MZs enhanced the ability of ML models to capture localized spatial relationships.

The spatially constrained cross-validation adopted in this study likely produced more conservative and realistic estimates than conventional random validation approaches. This explains why the reported R^2 values, although high in later scenarios, remained lower than values sometimes reported in the literature using random train–test partitioning (Filippi et al., 2025). Spatial validation reduces information leakage caused by spatial autocorrelation and provides a more reliable assessment of model generalization capability under real-field conditions.

The results demonstrate that combining stable soil–terrain attributes, historical yield information, fuzzy MZs, and ML algorithms provides a robust framework for yield prediction in irrigated grain production systems, consistent with previous integrated-data and ML frameworks (Filippi et al., 2025; Paudel et al., 2021, Li et al., 2025). The progressive improvement observed with additional historical yield maps also reinforces the importance of temporal stability in PA datasets (Filippi et al., 2025; Morales and Villalobos, 2023; Wei et al., 2026).

Conclusion

The integration of stable soil and terrain attributes, historical yield maps, fuzzy MZs, and ML algorithms improved within-field crop yield prediction under irrigated grain production conditions. Historical yield information was the main factor increasing predictive performance, with accuracy progressively improving as additional yield maps were incorporated into the models. MZ-based modelling enhanced prediction performance for some crops and algorithms, particularly under conditions of greater spatial heterogeneity, although whole-field approaches remained competitive in highly stable scenarios. Among the evaluated algorithms, RF, XGB, and SVR outperformed LR, confirming the nonlinear nature of crop yield response. Spatially constrained cross-validation provided more realistic model assessment by minimizing spatial autocorrelation effects. The proposed framework demonstrates potential for supporting site-specific management and decision-making in PA systems.

References

Badshah, A., Alkazemi, B. Y., Din, F., Zamli, K. Z., & Haris, M. (2024). Crop classification and yield prediction using robust machine learning models for agricultural sustainability. *IEEE Access*, 12, 162799–162813. doi:10.1109/ACCESS.2024.3486653

Bansod, B. S., & Pandey, O. P. (2013). An application of PCA and fuzzy C-means to delineate management zones and variability analysis of soil. *Eurasian Soil Science*, 46, 556–564. doi:10.1134/S1064229313050165

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York, USA: Plenum Press. doi:10.1007/978-1-4757-0450-1

Blackmore, S., & Moore, M. (1999). Remedial correction of yield map data. *Precision Agriculture*, 1(1), 53–66. doi:10.1023/A:1009969601387

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. doi:10.1023/A:1010933404324

Burdett, H., & Wellen, C. (2022). Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. *Precision Agriculture*, 23(5), 1553–1574. doi:10.1007/s11119-022-09897-0

Cao, J., Zhang, Z., Luo, Y., Zhang, L., Zhang, J., Li, Z., & Tao, F. (2021). Wheat yield predictions at a county and field scale with deep learning, machine learning, and Google Earth Engine. *European Journal of Agronomy*, 123, 126204. doi:10.1016/j.eja.2020.126204

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, USA: Association for Computing Machinery. doi:10.1145/2939672.2939785

Climate-Data.org. (n.d.). Climate data for cities worldwide. Available at: https://en.climate-data.org. Accessed 16 February 2026.

Córdoba, M., Bruno, C., Costa, J., & Balzarini, M. (2013). Subfield management class delineation using cluster analysis from spatial principal components of soil variables. *Computers and Electronics in Agriculture*, 97, 6–14. doi:10.1016/j.compag.2013.06.005

Corwin, D. L., & Lesch, S. M. (2005). Apparent soil electrical conductivity measurements in agriculture. *Computers and Electronics in Agriculture*, 46(1–3), 11–43. doi:10.1016/j.compag.2004.10.005

Doolittle, J. A., & Brevik, E. C. (2014). The use of electromagnetic induction techniques in soils studies. *Geoderma*, 223–225, 33–45. doi:10.1016/j.geoderma.2014.01.027

Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA). (1979). *Súmula da 10ª reunião técnica de levantamento de solos*. Rio de Janeiro, Brazil: Serviço Nacional de Levantamento e Conservação de Solos.

Filippi, P., Han, S. Y., & Bishop, T. F. A. (2025). On crop yield modelling, predicting, and forecasting and addressing the common issues in published studies. *Precision Agriculture*, 26, Article 8. doi:10.1007/s11119-024-10212-2

Filippi, P., Jones, E. J., Wimalathunge, N. S., Somarathna, P. D. S. N., Pozza, L. E., Ugbaje, S. U., et al. (2019). An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, 20, 1015–1029. doi:10.1007/s11119-018-09628-4

Fridgen, J. J., Kitchen, N. R., Sudduth, K. A., Drummond, S. T., Wiebold, W. J., & Fraisse, C. W. (2004). Management zone analyst (MZA): Software for subfield management zone delineation. *Agronomy Journal*, 96(1), 100–108. doi:10.2134/agronj2004.0100

Gavioli, A., de Souza, E. G., Bazzi, C. L., Guedes, L. P. C., & Schenatto, K. (2019). Optimization of management zone delineation by using spatial principal components. *Computers and Electronics in Agriculture*, 164, 104859. doi:10.1016/j.compag.2019.104859

Gebbers, R., & Adamchuk, V. I. (2010). Precision agriculture and food security. *Science*, 327(5967), 828–831. doi:10.1126/science.1183899

Habibi, L. N., Matsui, T., & Tanaka, T. S. T. (2024). Critical evaluation of the effects of a cross-validation strategy and machine learning optimization on the prediction accuracy and transferability of a soybean yield prediction model using UAV-based remote sensing. *Journal of Agriculture and Food Research*, 16, 101096. doi:10.1016/j.jafr.2024.101096

Jabed, M. A., & Azmi Murad, M. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*, 10(24), Article e40836. doi:10.1016/j.heliyon.2024.e40836

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York, USA: Springer. doi:10.1007/b98835

Joshi, A., Mohanty, S. P., & Kougianos, E. (2023). Artificial intelligence and machine learning approaches for smart agriculture: A review. *Computers and Electronics in Agriculture*, 209, 107836. doi:10.1016/j.compag.2023.107836

Köppen, W., & Geiger, R. (1928). *Klimate der Erde*. Gotha, Germany: Verlag Justus Perthes.

Li, Y., Miao, Y., Rawat, A., Stueve, K., & Cao, W. (2025). Identifying key soil and landscape factors influencing yield spatial patterns for management zone delineation using ensemble machine learning. *Computers and Electronics in Agriculture*, 237, 110572. doi:10.1016/j.compag.2025.110487

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 2674. doi:10.3390/s18082674

Maestrini, B., & Basso, B. (2018). Predicting spatial patterns of within-field crop yield variability. *Field Crops Research*, 219, 106–112. doi:10.1016/j.fcr.2018.01.028

Maldaner, L. F., Molin, J. P., & Spekken, M. (2021). Map Filter 2.0. Available at: https://www.agriculturadeprecisao.org.br/softwares. Accessed 22 February 2026.

Moore, I. D., Gessler, P. E., Nielsen, G. A., & Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, 57(2), 443–452. doi:10.2136/sssaj1993.03615995005700020026x

Moral, F. J., Terrón, J. M., & Marques da Silva, J. R. (2010). Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. *Soil and Tillage Research*, 106(2), 335–343. doi:10.1016/j.still.2009.12.002

Morales, A., & Villalobos, F. J. (2023). Using machine learning for crop yield prediction in the past or the future. *Frontiers in Plant Science*, 14, Article 1128388. doi:10.3389/fpls.2023.1128388

Mulla, D. J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering*, 114(4), 358–371. doi:10.1016/j.biosystemseng.2012.08.009

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187, 103016. doi:10.1016/j.agsy.2020.103016

Pereira, G. W., Valente, D. S. M., Queiroz, D. M., Coelho, A. L. F., Costa, M. M., & Grift, T. (2022).

Smart-Map: An open-source QGIS plugin for digital mapping using machine learning techniques and ordinary kriging. *Agronomy*, 12(6), Article 1350. doi:10.3390/agronomy12061350

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. doi:10.1111/ecog.02881

Rumsey, D. J. (2023). What is r value correlation? *Dummies*. Available at: https://www.dummies.com/article/academics-the-arts/math/statistics/what-is-r-value-correlation-289445. Accessed 15 January 2026.

Saleh, A. M., & Belal, A. A. (2014). Delineation of management zones by fuzzy clustering techniques using soil and topographic attributes. *Egyptian Journal of Remote Sensing and Space Sciences*, 17(2), 181–190. doi:10.1016/j.ejrs.2014.03.001

Schepers, A. R., Shanahan, J. F., Liebig, M. A., Schepers, J. S., Johnson, S. H., & Luchiari, A. (2004). Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years. *Agronomy Journal*, 96(1), 195–203. doi:10.2134/agronj2004.0195

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222. doi:10.1023/B:STCO.0000035301.49549.88

Sterle, L. G. d. G., & Molin, J. P. (2025). Management zones for irrigated and rainfed grain crops based on data layer integration. *Agronomy*, 15(8), Article 1864. doi:10.3390/agronomy15081864

Sudduth, K. A., & Drummond, S. T. (2007). Yield editor: Software for removing errors from crop yield maps. *Agronomy Journal*, 99(6), 1471–1482. doi:10.2134/agronj2006.0326

United States Department of Agriculture (USDA). (2014). *Keys to soil taxonomy* (12th ed.). Washington, DC, USA: USDA.

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

Vega, A., Córdoba, M., Castro-Franco, M., & Balzarini, M. (2019). Protocol for automating error removal from yield maps. *Precision Agriculture*, 20(5), 1030–1044. doi:10.1007/s11119-018-09632-8

Wei, M. C. F., Longchamps, L., Colaço, A. F., & Molin, J. P. (2026). Integrating stability zones and machine learning for enhanced crop management. *Precision Agriculture*, 27(2), 38. doi:10.1007/s11119-026-10318-9

Whelan, B. M., & McBratney, A. B. (2000). The "null hypothesis" of precision agriculture management. *Precision Agriculture*, 2(3), 265–279. doi:10.1023/A:1011838806489

Win, K. N. (2024). OpenTopography DEM Downloader — QGIS plugin. Available at: https://plugins.qgis.org/plugins/OpenTopography-DEM-Downloader. Accessed 5 January 2026.