PERFORMANCE OF SPATIAL PREDICTION MODELS UNDER DIFFERENT SAMPLE DENSITIES IN CORN YIELD MAPS

Matheus H. B. Vieira[1]*, Leandro M. Gimenez[1], Lara M. G. Santos[2]

1.Department of Biosystems Engineering, "Luiz de Queiroz" College of Agriculture (ESALQ), University of São Paulo (USP), 13418900 Piracicaba, São Paulo, Brazil

2.Fatec Pompeia - Shunji Nishimura, 17580-000 Pompeia, São Paulo, Brazil

**Corresponding author: matheus.hass27@gmail.com*

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.

Yield maps are essential for the consistent management of spatial variability in agricultural fields. However, the accuracy of these maps is directly affected by the data density collected by combine harvesters and the choice of the interpolation method, whether deterministic or statistical. This study aimed to evaluate the performance of four spatial prediction methods across decreasing sample densities for generating corn yield maps. The study area comprised an 11-hectare commercial field, utilizing a dataset from three consecutive growing seasons (2015, 2016, and 2017), each containing approximately 10,000 original sampling points. Data processing included two deterministic approaches: Inverse Distance Weighting (IDW) and Triangulated Irregular Network (TIN); one geostatistical approach: Kriging; and one machine learning algorithm: Random Forest (RF). To ensure unbiased model validation, an external cross-validation technique was adopted. Consequently, 500 points from each growing season were randomly selected and reserved exclusively for error verification. The remaining data volume was divided into three density scenarios: the total remaining dataset, a random subset of 5,000 points, and another of 2,500 points. The generated outputs resulted in raster files with a spatial resolution of 10 × 10 meters. Accuracy assessment consisted of overlaying the 500 validation points onto the rasters, extracting the predicted values for comparison with the actual harvested yield. Method performance in each scenario was quantified using Mean Error, Standard Deviation, and Root Mean Square Error (RMSE). Analyses indicated that spatiotemporal variation influences prediction errors. In high-point-density scenarios, the deterministic methods (TIN and IDW) satisfactorily represented crop yield. Conversely, as the dataset density was reduced, Kriging and Random Forest demonstrated a greater capacity to maintain spatial prediction stability. These results suggest that the application of machine learning and sample size optimization can be effective strategies to reduce the computational workload in precision agriculture projects. Based on these findings, the authors suggest that further studies on this topic should be conducted.

Keywords.
Precision Agriculture; Interpolation; Machine Learning; Yield Maps.

Introduction

As one of the main pillars of the Brazilian Gross Domestic Product (GDP), agriculture contributes approximately 25.13% to the country's economy (CEPEA. 2026). This value is closely correlated with global population growth in recent decades, which is projected to reach 8.5 billion people by the year 2030 (United Nations. 2024). Consequently, the demand for an increased food supply has risen, positioning Brazil as one of the largest food producers and exporters in the world (FAO. 2025). In this context, the adoption of Precision Agriculture (PA) serves as a critical tool to close the yield gap and optimize agricultural input management, aiming to increase both productivity and profitability (Cherubin et al. 2022).

Precision Agriculture (PA) is a management strategy that utilizes advanced technologies to monitor the spatial and temporal variability of crops. Its objective is to optimize input application, maximize yield, and promote socioeconomic and environmental sustainability (Cherubin et al. 2022; Oliveira et al. 2015). In Brazil, the trajectory of PA spans 25 years, evolving from an initial phase focused primarily on agricultural machinery and soil mapping into a robust digital ecosystem that integrates sensors, remote sensing, and artificial intelligence (Cherubin et al. 2022). Currently, PA is considered the foundation for Agriculture 4.0 or “Smart Farming", which seeks full automation and total connectivity in the field (Karunathilake et al. 2023).

Within this framework, yield maps are considered the fundamental pillar of PA, as they materialize the actual crop response to variations in soil, climate, and applied management practices (Martins et al. 2023). They serve as the starting point for variability management, enabling the quantification of localized production and the identification of limiting factors (Oliveira et al. 2015). Despite their importance in supporting strategic decisions, such as variable-rate input application and the delineation of Management Zones (MZs**)**, the widespread adoption of these maps still faces barriers in Brazil, including the high cost of yield monitors and the demand for technical training in raw data processing (Cherubin et al. 2022).

To effectively utilize these maps, spatial interpolation is required. This mathematical process converts discrete sampling points into a continuous surface, estimating values for unmeasured locations (Oliveira et al. 2015; Sobjak et al. 2023). This procedure is indispensable in agronomy and geosciences, given that sampling every square meter of a field is operationally and economically unfeasible (Oliveira et al. 2015). Interpolation methods can be classified into deterministic approaches, which are based on mathematical proximity functions, or **stochastic** (geostatistical) approaches, which model the spatial dependence structure to minimize estimation errors (Betzek et al. 2017; Sobjak et al. 2023).

In PA, the precision of the generated map directly affects the accuracy of agronomic recommendations and site-specific interventions. There is no universal consensus regarding which interpolator yields the best results, as the performance of each technique is highly sensitive to sampling density and the inherent nature of the studied variable (Betzek et al. 2017; Sobjak et al. 2023). Currently, tools such as the AgDataBox platform and the Smart-Map plugin seek to automate this selection process based on rigorous statistical criteria (Pereira et al. 2022; Sobjak et al. 2023).

The Inverse Distance Weighting (IDW) method is a deterministic interpolator widely used due to its computational simplicity and fast processing speed (Betzek et al. 2017; Pereira et al. 2022). It estimates values based on the premise that the influence of a sampled point decreases as the distance to the estimated location increases (Pereira et al. 2022). Although effective for high-density datasets, IDW does not account for the spatial autocorrelation structure. Consequently, this can result in less precise maps for variables with complex spatial dependence when compared to geostatistical methods (Sobjak et al. 2023).

The Triangulated Irregular Network (TIN) is a direct estimation method that connects sampling points to form a network of triangles, where each edge is linearly interpolated (Adedapo & Zurqani. 2024). Its primary advantage in PA is its strict fidelity to the observed data, as the method does

not extrapolate values outside the known data range and restricts its outputs to the perimeter of the sampled area (Oliveira et al. 2015). Recent studies indicate that TIN is particularly efficient for generating digital elevation models and representing uniform surfaces in high-point-density scenarios (Adedapo & Zurqani. 2024).

Ordinary Kriging (OK) is widely recognized as the "Best Linear Unbiased Estimator" (BLUE) (Martins et al. 2023; Oliveira et al. 2015). Unlike simpler methods, Kriging utilizes variographic analysis to model the spatial dependence structure of the variable. This allows not only for the estimation of localized values but also for the quantification of the uncertainty associated with the predictions through the estimation variance (Oliveira et al. 2015). Thus, it remains the benchmark method for soil fertility mapping and for estimating yield gains following biological interventions (Betzek et al. 2017; Martins et al. 2023).

The integration of Machine Learning (ML) algorithms, such as Random Forest (RF), represents the current frontier of Agriculture 4.0 (Filippi et al. 2025; Karunathilake et al. 2023). RF is an ensemble learning method that constructs multiple decision trees to identify complex, non-linear correlations between the target variable and various environmental covariates (Cherubin et al. 2022; Sekulić et al. 2020). Its application in PA stands out for its robustness in data-intensive scenarios and its high predictive capacity, frequently outperforming traditional interpolators by integrating multiple environmental factors into the spatial mapping process (Filippi et al. 2025; Sekulić et al. 2020).

Accordingly, this study aims to evaluate the performance of four spatial prediction methods (IDW, TIN, Ordinary Kriging, and Random Forest) tested under different decreasing sampling densities, providing a comprehensive comparison for the development of high-accuracy corn crop yield maps.

Material and methods

The data are derived from three corn crop growing seasons (2015, 2016, and 2017), from an 11 ha agricultural area under No-Tillage, in a clay-textured Oxisol (Soil Survey Staff. 2022), located in the municipality of Cândido Mota – SP, (22°53'57.10"S, 50°23'58.11"W), (Figure 1), Cfa climate (Köppen. 1931).

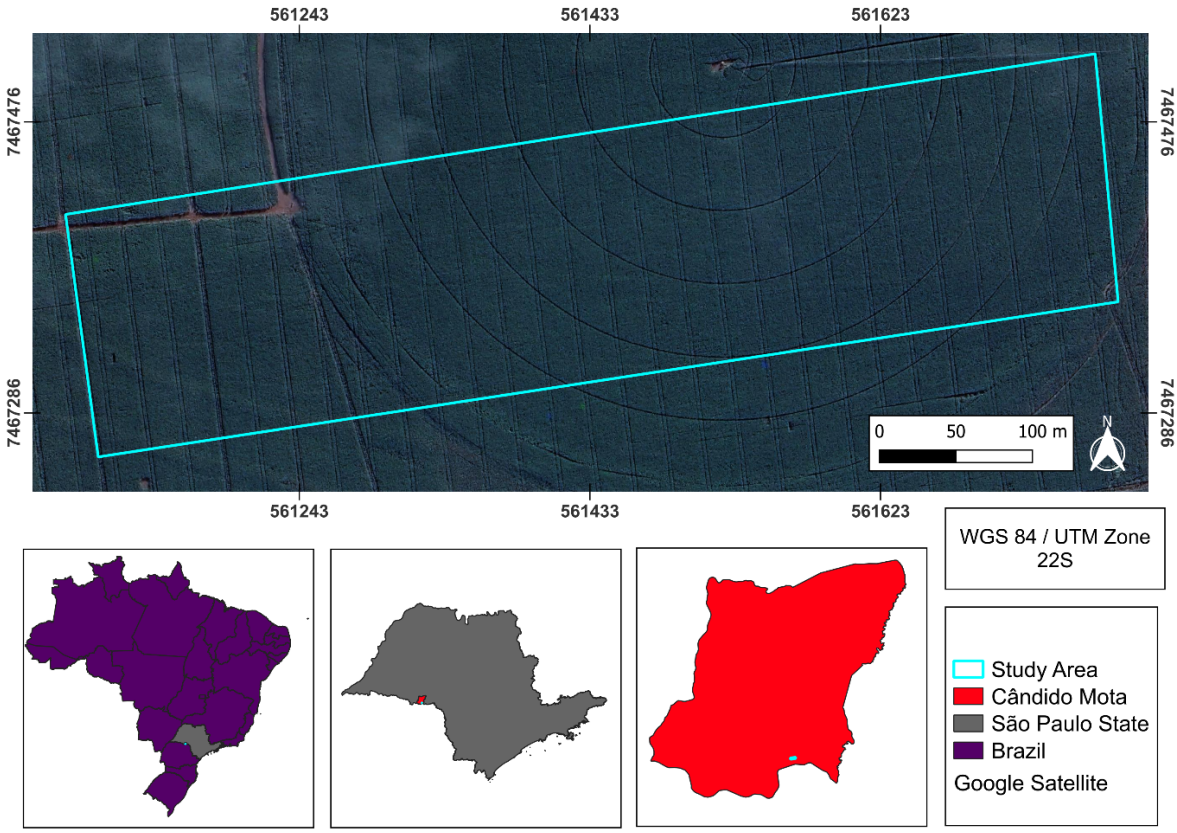

Figure 1: Study area

The data were obtained using a grain harvester from the manufacturer John Deere, model STS 9750, equipped with an impact plate-type yield sensor; a capacitance moisture sensor; a Hall effect-type speed sensor installed on the drive wheels; a gyroscope-type tilt sensor; and a John Deere GS4® model monitor that displays, records information, and provides the saved data in an electronic spreadsheet format. These data were obtained prior to the installation of the center-pivot irrigation mechanism.

The yield measured by the sensor was corrected to 13% moisture content using the method suggested by Aguiar (1977). Subsequently, the data were processed by the method presented by Guanais Santos (2020). Then, a standardized identification was established for these data. The identifications and their respective yield values were organized by growing season in an electronic spreadsheet.

For model verification, the external cross-validation technique was adopted. Thus, 500 points from each year were randomly sampled and reserved solely for error verification. The remaining data volume was divided into three density scenarios: the total remaining set, a random subset of 5,000 points, and another of 2,500 points. The generated products resulted in raster files with a spatial resolution of 10x10 meters. The processing considered two deterministic approaches, Inverse Distance Weighting (IDW) and Triangulated Irregular Network (TIN), executed in QGIS software version 3.40.15. To perform the geostatistical approach, Kriging, the data were processed in Vesper software version 1.62 (Minasny et al. 2005), and a machine learning algorithm, Random Forest (RF), structured in Python version 3.11 through the scikit-learn library (Pedregosa et al. 2011).

To perform the accuracy assessment, an overlay of the 500 validation points onto the rasters was carried out, extracting the predicted value for comparison with the actual harvested yield. The performance of the methods in each scenario was quantified using the Mean Error, Standard Deviation, and Root Mean Square Error (RMSE).

Results and discution

Visual Analysis of Spatial Yield Variability

The spatial characterization of corn yield for the 2015, 2016, and 2017 growing seasons, generated by the four prediction methods, Inverse Distance Weighting (IDW), Triangulated Irregular Network (TIN), Ordinary Kriging, and Random Forest (RF), under different sampling density scenarios (Full dataset, 5,000 points, and 2,500 points), is presented comparatively in Figure 2. The standardized spatial representation with a pixel resolution of 10x10 m enabled a direct visual comparison among the different predicted surfaces.

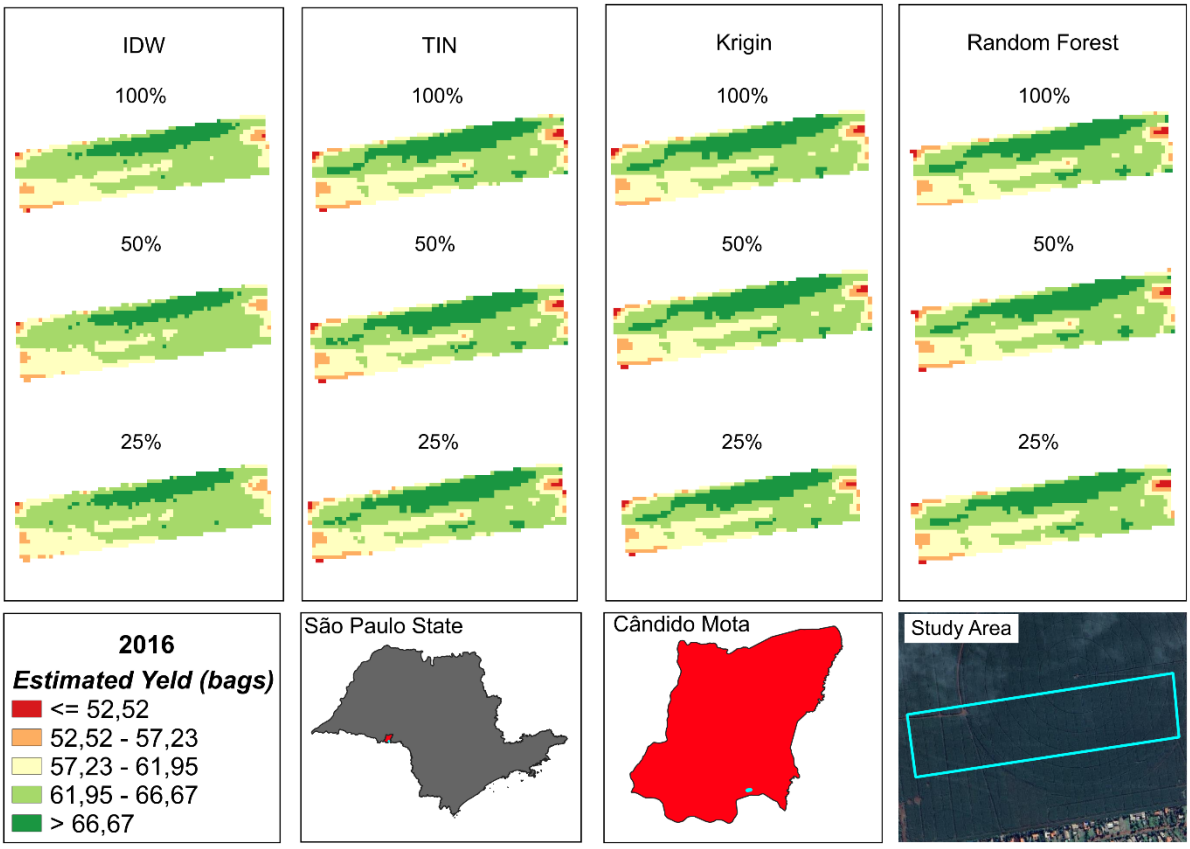

Figure 2. Corn yield maps for the 2016 growing season using IDW, TIN, Kriging, and Random Forest methods under three sampling densities.

The visual analysis of the maps reveals that spatial yield variability was consistently mapped by all algorithms. However, distinct differences in the delineation of spatial variability are observed. The TIN interpolator presented surfaces with abrupt transitions marked by a faceted triangulation effect. IDW generated the classic "bull's-eye" patterns (concentric islands around the sampling points). In contrast, Kriging and the Random Forest algorithm generated smoother, more continuous surfaces, delineating yield clusters and transition zones in a manner more consistent with the variability typically observed in fields (Betzek et al. 2017; Oliveira et al. 2015; Sobjak et al. 2023).

Regarding the effect of reducing sampling density, downsampling the original database to 5,000 (50%) and 2,500 (25%) points maintained the spatial pattern of the field. This factor visually indicates a high spatial redundancy in the raw data from the yield monitor.

Accuracy Assessment and Sampling Density Effect

In order to quantify the statistical rigor of the predictions, the yield estimated by each model was cross-referenced with the actual yield from the 500 external validation points. The quantitative

evaluation based on Mean Error, Standard Deviation, and Root Mean Square Error (RMSE) revealed differences in the predictive capacity of each model and the impact of data volume.

Table 1. Statistical error indicators of spatial predictions for the 2015, 2016, and 2017 growing seasons under different interpolation methods and sampling densities.

		2015			2016			2017		
		Mean Error*	Standard Deviation*	RMSE*	Mean Error*	Standard Deviation*	RMSE*	Mean Error*	Standard Deviation*	RMSE*
IDW	100%	-0.005	0.453	0.452	-0.021	1.188	1.187	0.021	0.56	0.559
	50%	-0.022	0.483	0.483	-0.005	1.34	1.339	0.032	0.591	0.592
	25%	-0.008	0.519	0.518	0.015	1.487	1.486	0.029	0.655	0.655
TIN	100%	-0.05	0.582	0.584	0.02	1.671	1.669	0.093	0.609	0.616
	50%	-0.048	0.558	0.559	0.034	1.641	1.64	0.087	0.608	0.614
	25%	-0.036	0.56	0.561	0.045	1.613	1.612	0.085	0.628	0.633
Krig	100%	0.01	0.302	0.302	-0.065	0.832	0.834	0.014	0.334	0.334
	50%	0.004	0.304	0.304	-0.062	0.833	0.834	0.011	0.332	0.332
	25%	0.004	0.306	0.306	-0.056	0.852	0.853	0.016	0.343	0.343
RF	100%	0.015	0.415	0.415	-0.038	0.918	0.918	0.019	0.347	0.347
	50%	0.003	0.396	0.395	-0.067	0.901	0.903	0.013	0.346	0.346
	25%	0.022	0.365	0.366	-0.075	0.918	0.92	0.005	0.362	0.362

*** All error metrics (Mean Error, Standard Deviation, and RMSE) are expressed in bags ha^{-1})**

The consolidated results demonstrated the technical superiority of Ordinary Kriging in representing field variability. Kriging consistently achieved the lowest error indices among all tested methods. In the 2015 growing season, for instance, the Kriging RMSE operated in the range of 0.302 to 0.306 bags ha^{-1}, also presenting the lowest dispersion of residuals (standard deviation).

On the other hand, the strictly deterministic and geometric methods (IDW and TIN) showed the highest error levels, confirming their high sensitivity to the spatial distribution of samples and their inability to extrapolate trends or efficiently handle harvester logging anomalies.

The Inverse Effect of Sampling Density

There was a divergent behavior among the models when subjected to the reduction of sampling density. Contrary to the common assumption that maximum point density always generates the best map, the data proved opposite dynamics between Machine Learning and traditional approaches.

When reducing the training data from 100% to only 2,500 samples, the deterministic methods were severely penalized. IDW, for example, saw its RMSE rise from 0.452 (with all data) to 0.518 (with 2,500 points) in the 2015 growing season (Table 1). Because these algorithms depend exclusively on the Euclidean distance between available samples, increasing the spacing between points degraded the accuracy of the map (Betzek et al. 2017; Oliveira et al. 2015; Sobjak et al. 2023).

In contrast, the same data reduction benefited the performance of the Random Forest algorithm. In the 2015 growing season, the Random Forest RMSE dropped progressively from 0.415 (with 100% of the data) to 0.395 (with 5,000 points) and reached its best performance, 0.366, with only 2,500 points. This phenomenon demonstrates that decision-tree-based algorithms can suffer from overfitting or over-assimilation. The random sampling likely acted as a redundancy filter, simplifying machine learning and generating predictions that were not only computationally faster but also more accurate (Pereira et al. 2022).

Scatter Analysis and Linear Correlation

To graphically understand the behavior of residuals and smoothing phenomena, Figures 3, 4, and 5 detail the linear regression between the observed yield at the validation points and the estimated yield for each growing season.

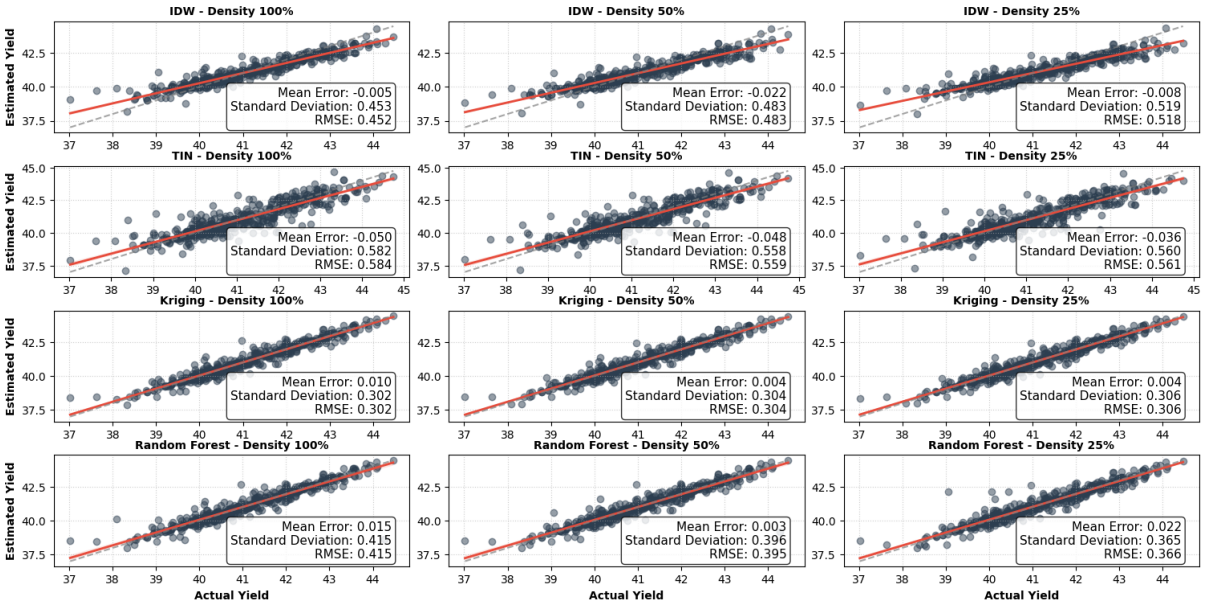

Figure 3. Scatter plots and linear regression between observed and predicted yield under densities of 100%, 50%, and 25% for the 2015 growing season.

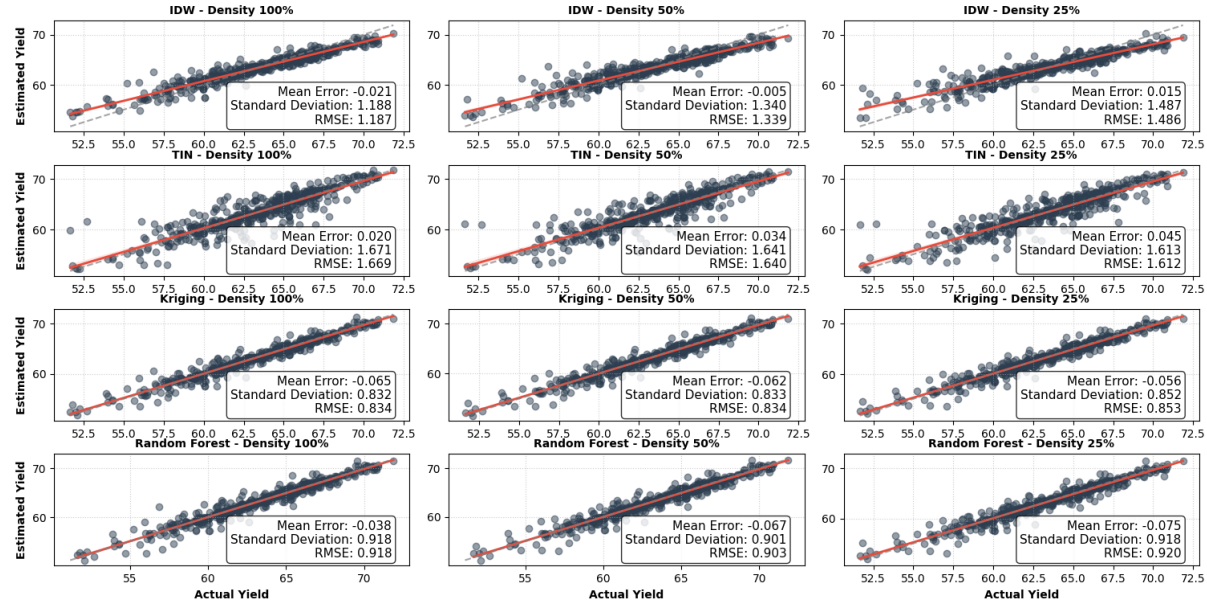

Figure 4. Scatter plots and linear regression between observed and predicted yield under densities of 100%, 50%, and 25% for the 2016 growing season.

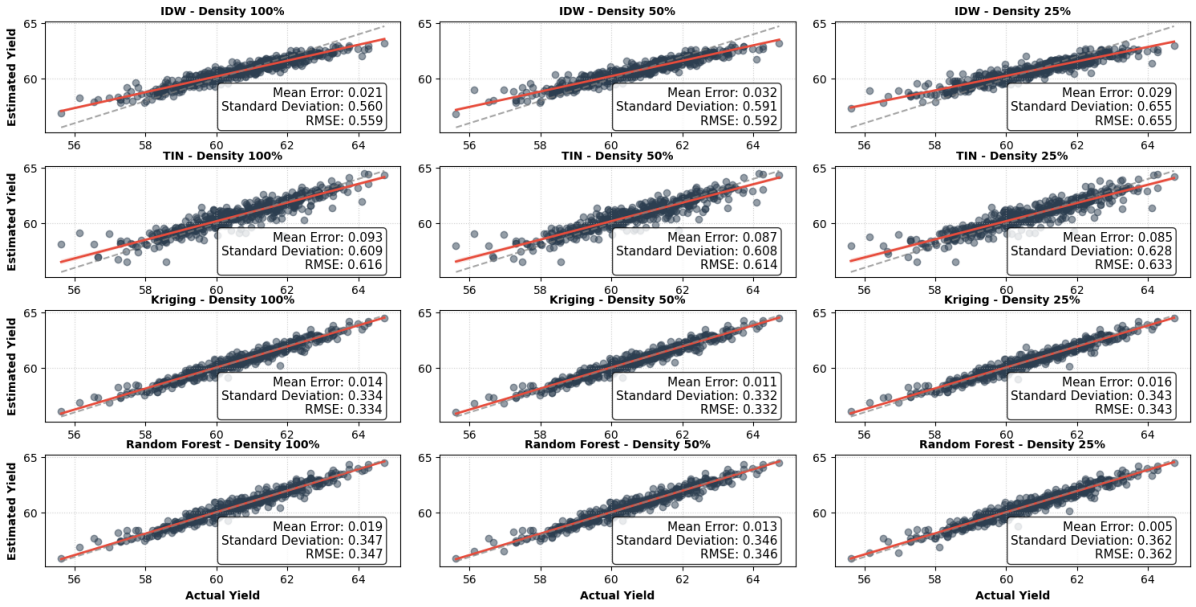

Figure 5. Scatter plots and linear regression between observed and predicted yield under densities of 100%, 50%, and 25% for the 2017 growing season.

The analysis of the scatter clouds corroborates the quantitative metrics. Kriging displays the tightest scatter widths relative to the trend line (1:1), confirming its high structural reliability.

In the case of Random Forest, the plots show a slight flattening trend at the extremes (overestimating very low values and underestimating absolute peaks). However, the progression of the panels (from the 100% density columns to the 25% column) shows a cleaner consolidation of the scatter cloud, visually reflecting the reduction in error (RMSE) achieved through the sampling filter.

Conclusion

It is concluded that Ordinary Kriging remains the most accurate methodology for mapping corn crop yield. However, the use of Machine Learning algorithms can be considered a potential technique, as it does not require assumptions of stationarity or the modeling of theoretical semivariograms. Random Forest constitutes an alternative for massive data processing. The use of reduced sample sizes enables not only a reduction in the computational cost in precision agriculture, but likely serves as a noise management mechanism, improving the accuracy of artificial intelligence for predicting yield maps. Based on these findings, the authors suggest that further studies on this topic should be conducted.

Acknowledgments

We thank Professor Leandro for his guidance, assistance in the process of idealizing this work, and for facilitating our attendance at the congress. To Professor Lara for her friendship, for providing the data, and for her assistance in the review process. To the Precision Agriculture Laboratory (LAP) for funding the transport and accommodation required for my participation in the congress.

References

Adedapo, S. M., & Zurqani, H. A. (2024). Evaluating the performance of various interpolation techniques on digital elevation models in highly dense forest vegetation environment. *Ecological Informatics*, *81*, 102646. https://doi.org/10.1016/j.ecoinf.2024.102646

Betzek, N. M., Souza, E. G. de, Bazzi, C. L., Sobjak, R., Bier, V. A., & Mercante, E. (2017).

Interpolation methods for thematic maps of soybean yield and soil chemical attributes. *Semina: Ciências Agrárias*, *38*(2), 1059. https://doi.org/10.5433/1679-0359.2017v38n2p1059

Centro de Estudos Avançados em Economia Aplicada (CEPEA). (2026). PIB do agronegócio brasileiro. CEPEA/Esalq/USP, CNA. https://cepea.org.br

Cherubin, M. R., Damian, J. M., Tavares, T. R., Trevisan, R. G., Colaço, A. F., Eitelwein, M. T., et al. (2022). Precision Agriculture in Brazil: The Trajectory of 25 Years of Scientific Research. *Agriculture*, *12*(11), 1882. https://doi.org/10.3390/agriculture12111882

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*(1), 8. https://doi.org/10.1007/s11119-024-10212-2

Food and Agriculture Organization (FAO). (2025). FAOSTAT: Suite of Food Security Indicators. FAO. https://www.fao.org/faostat/

Guanais Santos, L. M. (2020). Tratamento e interpretação de dados de produtividade para agricultura de precisão [Dissertação de Mestrado]. Universidade Estadual de Londrina.

Karunathilake, E. M. B. M., Le, A. T., Heo, S., Chung, Y. S., & Mansoor, S. (2023). The Path to Smart Farming: Innovations and Opportunities in Precision Agriculture. *Agriculture*, *13*(8), 1593. https://doi.org/10.3390/agriculture13081593

Köppen, W. P. (1931). Grundriss der Klimakunde. Walter de Gruyter.

Martins, G. D., Xavier, L. C. M., de Oliveira, G. P., de Lourdes Bueno Trindade Gallo, M., de Abreu Júnior, C. A. M., Vieira, B. S., et al. (2023). Using Geospatial Information to Map Yield Gain from the Use of Azospirillum brasilense in Furrow. *Agronomy*, *13*(3), 808. https://doi.org/10.3390/agronomy13030808

Minasny, B., McBratney, A.B., and Whelan, B.M. (2005). VESPER version 1.62. Australian Centre for Precision Agriculture, McMillan Building A05, The University of Sydney, NSW 2006. (http://www.usyd.edu.au/su/agric/acpa)

Oliveira, R. P. de., Grego, C. Regina., & Brandão, Z. Neiva. (2015). *Geoestatística aplicada na agricultura de prescisão utilizando o Vesper*. Embrapa.

Pereira, G. W., Valente, D. S. M., de Queiroz, D. M., Santos, N. T., & Fernandes-Filho, E. I. (2022). Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. *Precision Agriculture*, *23*(4), 1189–1204. https://doi.org/10.1007/s11119-022-09880-9

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research (Vol. 12). http://scikit-learn.sourceforge.net.

Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random Forest Spatial Interpolation. *Remote Sensing*, *12*(10), 1687. https://doi.org/10.3390/rs12101687

Sobjak, R., de Souza, E. G., Bazzi, C. L., Opazo, M. A. U., Mercante, E., & Aikes Junior, J. (2023). Process improvement of selecting the best interpolator and its parameters to create thematic maps. *Precision Agriculture*, *24*(4), 1461–1496. https://doi.org/10.1007/s11119-023-09998-4

Soil Survey Staff. (2022). Keys to Soil Taxonomy, 13th edition. USDA Natural Resources Conservation Service.

United Nations, Department of Economic and Social Affairs, Population Division. (2024). World Population Prospects 2024: Summary of Results (UN DESA/POP/2024/TR/NO. 1). https://www.un.org