IMPACT OF SAMPLING DENSITY ON THE SPATIAL PREDICTION OF SOIL ATTRIBUTES USING GEOSTATISTICS AND MACHINE LEARNING

Sarah Abade Ribeiro[1], Helena Fantin Gebler[1], José Paulo Molin[1], Roberto Fray da Silva[1]

[1] Department of Biosystems Engineering, Luiz de Queiroz College of Agriculture, University of São Paulo, 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.

Soil fertility maps are strongly affected by the sampling strategy used to represent spatial variability. Although high-density sampling improves spatial diagnosis, it also increases field and laboratory costs, limiting the operational adoption of detailed soil mapping. This study evaluated how reducing soil sampling density affects the spatial prediction of soil attributes and compared the performance of ordinary kriging (OK), random forest (RF), and random forest regression kriging (RFRK) in preserving the interpolation quality obtained from a high-density reference grid. The study used 1,139 georeferenced soil sampling points collected across six commercial fields at an initial density of approximately 1 ha point^{-1}. Reduced sampling scenarios were simulated from 2 to 10 ha point^{-1}. For each field, sampling density, and soil attribute, interpolated surfaces were generated using OK, RF, and RFRK. The 1 ha point^{-1} surface was used as the reference condition, and reduced-density surfaces were compared with the corresponding reference surface using common grid cells. Model performance was evaluated using normalized error metrics, including RMSE/SD and MAE/SD. Sampling density had a clear effect on interpolation performance. The 2 ha point^{-1} density showed the lowest normalized error and was identified as the most reliable reduced sampling strategy. Densities of 3 and 4 ha point^{-1} showed moderate increases in error and may be considered secondary alternatives, whereas densities from 5–6 ha point^{-1} onward showed greater divergence from the reference surface. Among interpolation methods, RF showed the lowest mean normalized RMSE, while OK was the most frequent best-performing method across individual field–density–attribute combinations. RFRK did not provide a consistent improvement over RF or OK. Field-level analysis also showed that interpolation performance varied substantially among fields, indicating that local field conditions affected the reliability of reduced-density sampling. These results indicate that reducing sampling density can substantially affect soil attribute mapping, and that the most appropriate sampling strategy depends not only on the interpolation method, but also on the target attribute and field-level spatial complexity. Overall, 2 ha point^{-1} was the most stable reduced sampling density for preserving the spatial patterns obtained from the 1 ha point^{-1} reference condition.

Keywords. Precision Agriculture; Ordinary Kriging; Random Forest; Soil sampling optimization.

Introduction

Soil fertility maps play an important role in precision agriculture, supporting the spatial diagnosis of soil chemical attributes and the recommendation of fertilizers and lime at variable rates. In Brazil, soil sampling in regular grids is one of the most common practices used to generate these maps, with sampling grid sizes varying among agricultural regions. Molin (2017) reported that producers who mapped soil fertility used different grid sizes across Brazilian agricultural regions, indicating that sampling grid is often defined under practical and regional conditions rather than by a technical standard (Figure 1).

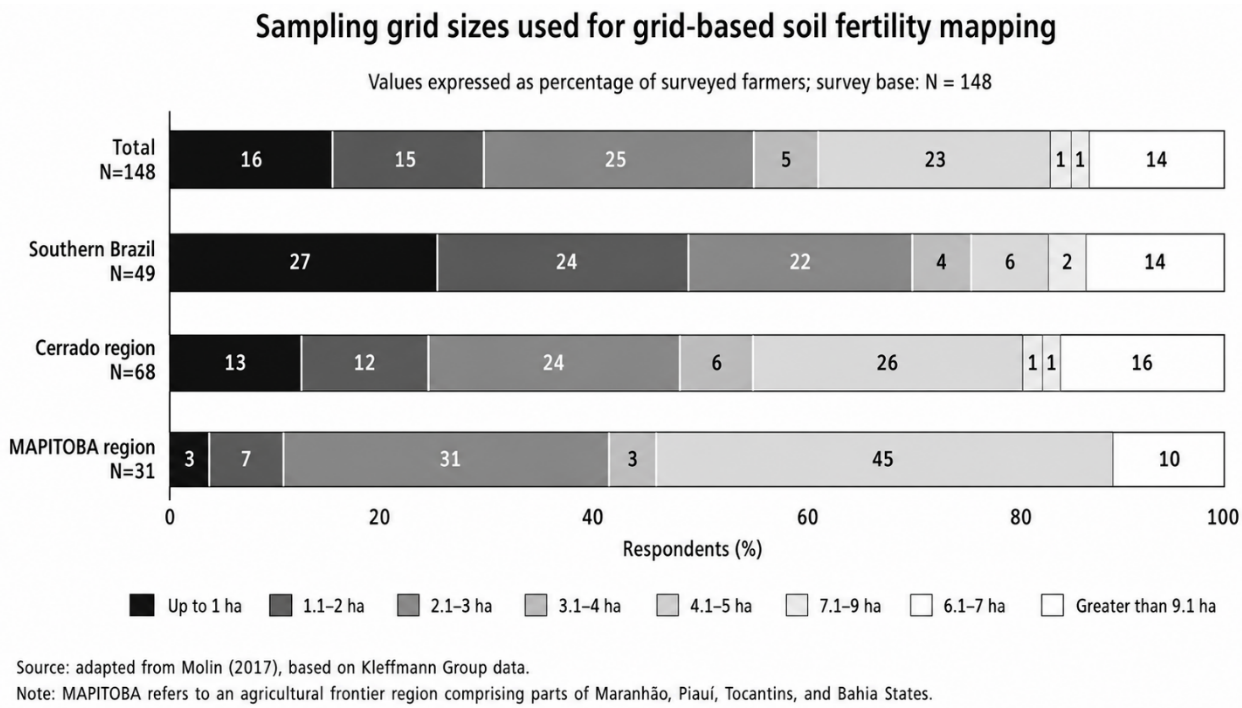

Figure 1. Sampling grid sizes used for soil fertility mapping by region in Brazil. Source: adapted from Molin (2017), based on Klefmann Group data.

The lack of clear technical criteria makes choosing a sampling grid highly complex. Denser grids better represent soil spatial variability but increase sampling and laboratory costs, requiring a high initial investment that many farmers are reluctant to make due to uncertain financial returns (Mattoso and Garcia, 2006). Conversely, sparse grids reduce diagnostic costs but may fail to capture crucial spatial patterns, generating inaccurate recommendation maps that impair input efficiency and reduce the overall economic and environmental benefits of precision agriculture (Valente et al., 2024). Thus, the primary challenge lies in identifying a suitable balance between data acquisition costs and the efficiency of site-specific management.

While the expansion of Digital Agriculture Technologies has increased the capacity to process spatial data and support site-specific management, the effect of reducing soil sampling density on interpolation quality remains a practical challenge. Therefore, this study evaluated how reduced sampling densities affect the spatial prediction of soil attributes and compared ordinary kriging, random forest, and random forest regression kriging in preserving the interpolation quality obtained from a 1 ha point^{-1} reference grid.

Methodology

The study was conducted in an agricultural production area of approximately 1,300 ha located in southern Maranhão State, Brazil. The area was divided into six fields based on operational boundaries and management history. The soil dataset was obtained from a high-density reference sampling grid of approximately 1 ha point⁻¹, comprising 1,139 georeferenced sampling points.

Soil samples were submitted for laboratory chemical and physical analyses. The evaluated soil attributes included aluminum (Al), boron (B), calcium (Ca), iron (Fe), potassium (K), magnesium (Mg), phosphorus (P), pH, sulfur (S), zinc (Zn), copper (Cu), sum of bases (SB), base saturation (V%), cation exchange capacity (CEC), clay content, soil organic matter (SOM), and potential acidity (H+Al). All coordinates were standardized to WGS 84 / UTM zone 23S, EPSG:32723.

Reduced sampling grids were generated from the original 1 ha point^{-1} reference grid to simulate lower-density soil sampling scenarios commonly used in commercial field conditions (Figure 2). The objective of the downsampling procedure was to select subsets of points that were not only numerically compatible with each target density, but also spatially plausible from an operational sampling perspective, rather than using a purely random selection of points.

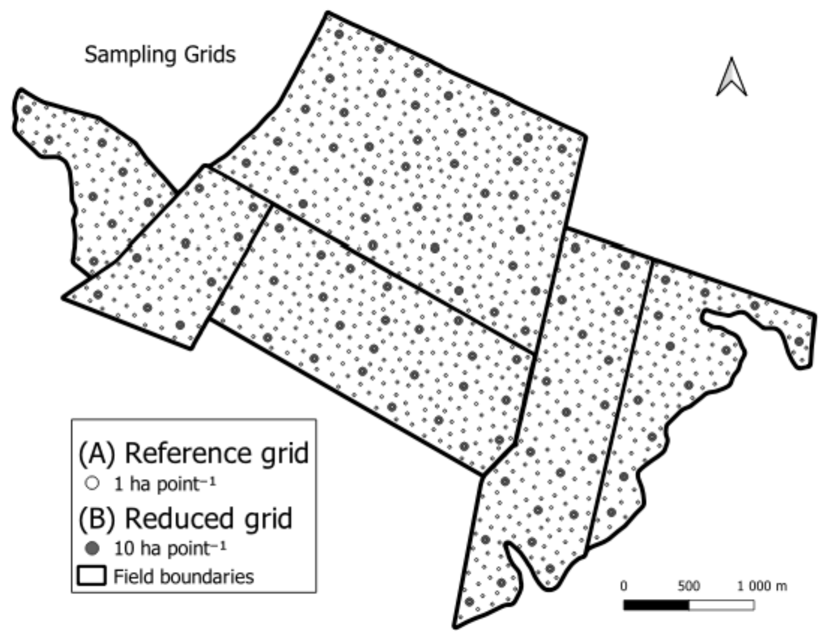

Fig. 2. Spatial distribution of soil sampling points in the study area for the reference and reduced sampling grids: (A) reference grid with 1 ha point^{-1} and (B) simulated reduced grid with 10 ha point^{-1}.

For each target density, from one point every 2 ha to one point every 10 ha, the expected number of sampling points was calculated based on the total field area and the desired sampling density (Table 1). Points from the reference grid were then selected to preserve a regular spatial distribution across the six fields, avoiding excessive clustering, large unsampled gaps, and uneven representation among field subdivisions.

The geometric consistency of each reduced grid was checked before the interpolation analyses. This assessment considered the final number of selected points, the effective sampling grid, the spatial distribution of points within each field, nearest-neighbor distances, and the visual identification of sampling gaps.

Grids (ha point $^{-1}$)	Total points	Absolute reduction vs. 1 ha point $^{-1}$	Relative reduction vs. 1 ha point $^{-1}$ (%)	Additional reduction vs. previous grid (%)
1	1139	0	0	0
2	571	568	50	50
3	381	758	67	33
4	286	853	75	25
5	229	910	80	20
6	192	947	83	16
7	165	974	86	14
8	144	995	87	13
9	129	1010	89	10
10	116	1023	90	10

Table 1. Point reduction by sampling density, relative to 1 ha point^{-1}.

To reduce instability in variogram estimation under sparse sampling conditions, field–density–attribute combinations were screened according to the number of valid observations. Combinations with at least 30 points were retained for variogram fitting and final interpolation comparison. Combinations with 15 to 29 points were treated only as exploratory diagnostics, whereas those with fewer than 15 points were excluded from modeling due to limited support for estimating spatial dependence.

Experimental semivariograms were fitted using spherical, exponential, Gaussian, and Matérn models. For the Matérn model, the smoothness parameter was fixed at $\kappa = 0.5$ to ensure consistent fitting across combinations. For each model, nugget, partial sill, total sill, and range were estimated. Spatial dependence was described using the nugget-to-sill ratio and classified as strong, moderate, or weak following Cambardella et al. (1994), as a variographic descriptor rather than a direct measure of interpolation accuracy. The best model for each field–density–attribute combination was selected by cross-validation, prioritizing the lowest root mean square prediction error, with mean absolute error, mean error, residual sum of squares, and model parsimony used as secondary criteria.

Spatial prediction was performed using ordinary kriging (OK), random forest (RF), and random forest regression kriging (RFRK). In OK, the selected semivariogram model and its parameters were used directly for kriging. In RF, easting and northing coordinates were used as predictors of each soil attribute. In RFRK, RF predictions were first generated, the residuals from the RF model were interpolated by ordinary kriging, and the kriged residual surface was added to the RF prediction surface.

All interpolation methods were applied to the same regular prediction grid within each field, allowing direct cell-by-cell comparison among sampling densities. The 1 ha point^{-1} sampling density was used as the high-density reference condition. Therefore, the analysis evaluated agreement with the reference surface rather than absolute prediction accuracy against independent observations. For each method, surfaces generated from reduced sampling densities were compared with the corresponding 1 ha point^{-1} reference surface using common grid cells.

Prediction performance was evaluated using mean error (ME), mean absolute error (MAE), root mean squared error (RMSE), Pearson correlation, coefficient of determination, regression slope, and intercept. Because soil attributes differed in units and numerical ranges, MAE and RMSE were also normalized by the standard deviation of the reference surface. The final sampling-density recommendation was based on a balanced comparison including only combinations available. Normalized RMSE, the 75th and 90th percentiles of normalized RMSE, and normalized MAE were used to compare OK, RF, and RFRK. Attribute-specific results were treated as exploratory when the balanced dataset retained few fields per attribute.

The global efficiency of the interpolation methods was assessed using the average normalized error and the frequency with which each method produced the lowest RMSE/SD for each field–density–attribute combination. To identify the main sources of variation in interpolation performance, linear models were fitted using RMSE/SD as the response variable and interpolation method, sampling density, field, and soil attribute as explanatory factors. Approximate partial R^2 values were used to describe the relative importance of each factor. A field-balanced analysis was also performed, with field identity interpreted as a proxy for local management history and spatial heterogeneity.

All analyses were conducted in R. Data handling was performed using dplyr, tidyr, readr, readxl, and janitor; spatial data processing was carried out with sf; geostatistical analyses were implemented with gstat; and RF models were fitted using ranger.

RESULTS AND DISCUSSION

The comparison among sampling densities showed a clear increase in normalized interpolation error as sampling became more sparse (Figure 3). The lowest normalized error was observed at 2 ha point^{-1}, which presented the smallest mean RMSE/SD and the smallest 75th percentile of RMSE/SD relative to the 1 ha point^{-1} reference surface. Densities of 3 and 4 ha point^{-1} showed moderate increases in error and can be considered secondary alternatives. From 5–6 ha point^{-1} onward, the divergence from the reference surface became more pronounced, especially in the upper error distribution, indicating greater instability in preserving the spatial pattern of the reference maps.

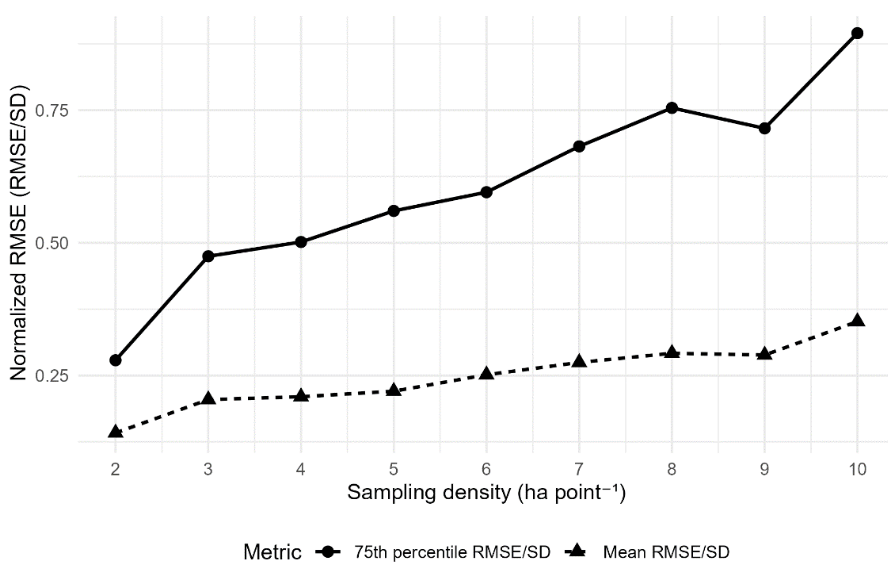

Fig. 3. Effect of sampling density on interpolation error relative to the 1 ha point^{-1} reference surface. Values represent the mean and 75th percentile of normalized RMSE (RMSE/SD) across interpolation methods, fields, and soil attributes.

When the interpolation methods were compared using mean normalized RMSE, RF showed the lowest global error, followed by OK and RFRK (Figure 4). This indicates that RF was slightly more efficient when all field–density–attribute combinations were pooled and evaluated by average error magnitude. However, this result should be interpreted together with the frequency of best performance at the individual-combination level.

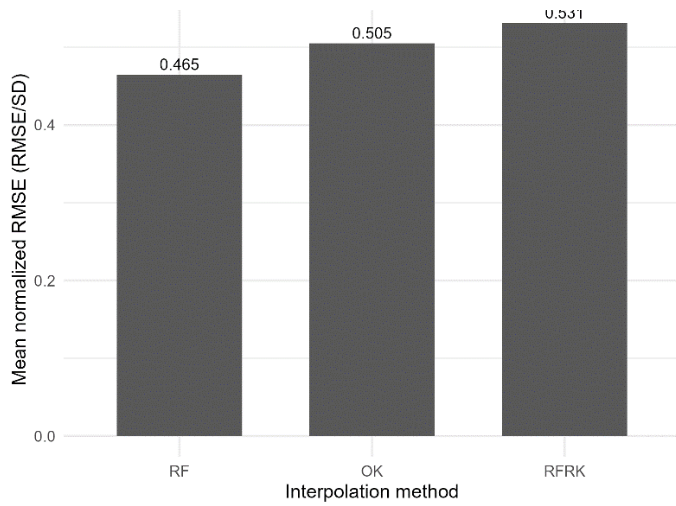

Fig. 4. Global efficiency of the interpolation methods based on mean normalized RMSE (RMSE/SD). Lower values indicate better overall agreement with the 1 ha point^{-1} reference surface.

Although RF had the lowest mean normalized RMSE, OK was the most frequent best-performing method, producing the lowest RMSE/SD in 66.5% of the field–density–attribute combinations (Figure 5). RF was the best method in 31.1% of the combinations, whereas RFRK was selected as the best method in only 2.4%. Therefore, RF reduced the overall average error, but OK showed greater consistency across individual combinations. The hybrid RFRK approach did not provide a consistent improvement over RF or OK under the conditions evaluated in this study.

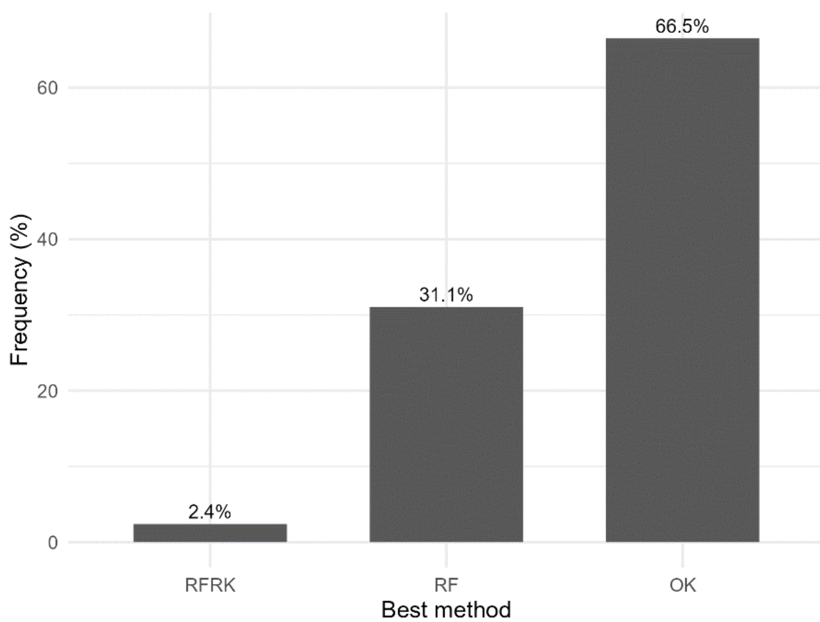

Fig. 5. Frequency of best-performing interpolation methods across field–density–attribute combinations. The best method was defined as the method with the lowest RMSE/SD for each combination

The field-balanced analysis also revealed substantial differences in interpolation performance among fields (Figure 6). Field 3 showed the greatest stability, with the lowest mean RMSE/SD, followed by field 5. Fields 4 and 1 showed the highest normalized errors, indicating greater difficulty in reproducing the reference surfaces under reduced sampling density. These results suggest that interpolation performance was not controlled only by sampling density or interpolation method, but also by local field conditions.

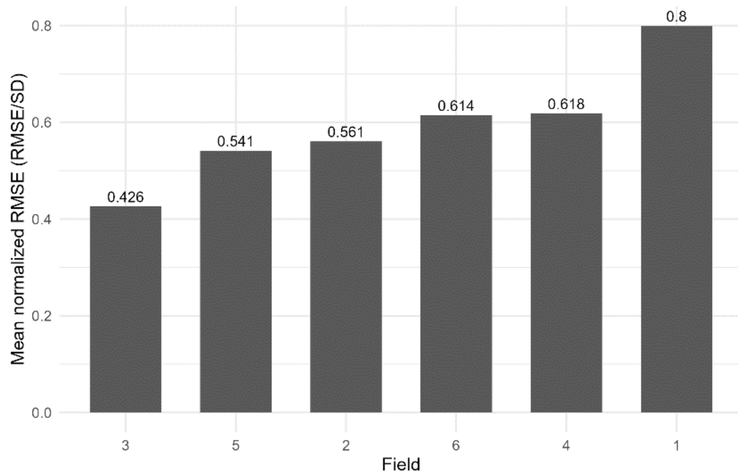

Fig. 6. Field-level interpolation performance based on the balanced comparison among fields. Values represent mean normalized RMSE (RMSE/SD) across common attributes, sampling density, and interpolation methods.

The analysis of factor importance reinforced this interpretation (Figure 7). In the complete dataset, soil attribute was the main source of variation in interpolation error, followed by sampling density, field, and interpolation method. This indicates that the intrinsic spatial behavior of each soil attribute strongly affected prediction quality, and that reducing sampling density was also an important source of error. In the field-balanced dataset, however, field became the dominant factor, followed by soil attribute and method. This result suggests that local field conditions, including management history and spatial heterogeneity, had a strong influence on interpolation performance when fields were compared under equivalent conditions.

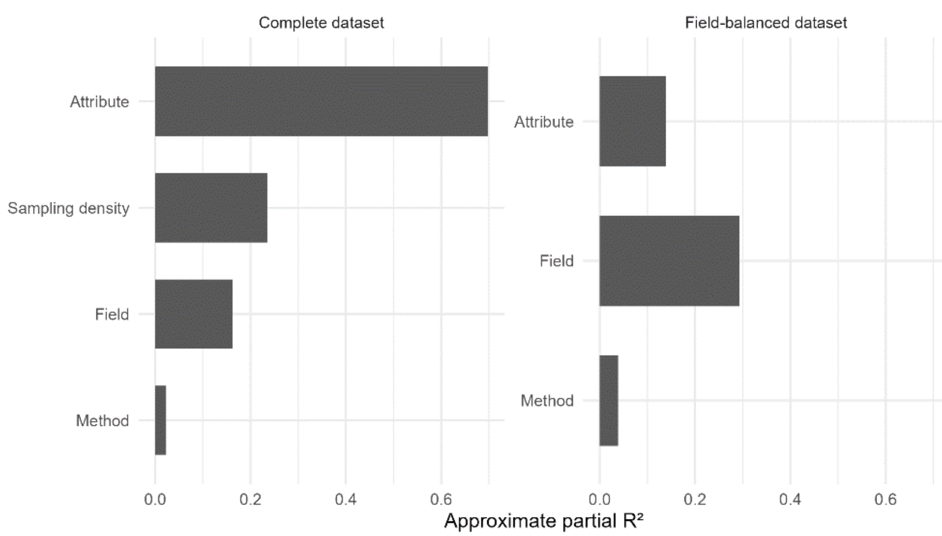

Fig. 7. Relative importance of factors affecting interpolation error. Approximate partial R^2 values were obtained from linear models using RMSE/SD as the response variable for the complete dataset and for the field-balanced dataset.

CONCLUSIONS

Reducing soil sampling density increased interpolation error relative to the 1 ha point^{-1} reference condition. The 2 ha point^{-1} density was the most reliable reduced sampling strategy, showing the lowest normalized error across interpolation methods, fields, and soil attributes. Densities of 3 and 4 ha point^{-1} may be considered secondary alternatives when operational constraints require further sampling reduction, but they already involve measurable loss of spatial agreement. Densities of 5–6 ha point^{-1} should be used with caution, while densities of 7 ha point^{-1} or higher were less stable and are not recommended as general sampling strategies under the evaluated conditions.

The comparison among interpolation methods showed that RF had the lowest mean normalized error, whereas OK was the most frequently best-performing method across individual field–density–attribute combinations. Thus, RF showed better average global efficiency, but OK remained the most consistent method. RFRK did not provide a consistent gain over RF or OK.

Interpolation performance also varied strongly among fields, indicating that local field conditions affected the reliability of reduced-density sampling. Therefore, practical recommendations for soil sampling density should consider not only the interpolation method, but also the target soil attribute and the spatial complexity of each field.

Overall, the results support 2 ha point^{-1} as the general recommended reduced sampling density for maintaining spatial prediction quality, with 3–4 ha point^{-1} as possible alternatives depending on operational constraints and acceptable levels of uncertainty.

Future work may expand this analysis by testing additional artificial intelligence approaches and by incorporating an economic assessment of sampling-density reduction. This would allow the trade-off between reduced sampling costs and loss of spatial resolution to be evaluated more directly in the context of site-specific management decisions.

Acknowledgments

The authors would like to thank CAPES for the first author scholarship (Proc. 88887.084394/2024-00) and APagri Consultoria Agronômica for supporting this study.

References

BREIMAN, L. Random forests. Machine Learning, v. 45, p. 5–32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324.

CAMBARDELLA, C. A. et al. Field-scale variability of soil properties in central Iowa soils. Soil Science Society of America Journal, Madison, v. 58, n. 5, p. 1501–1511, 1994.

MATTOSO, M. J.; GARCIA, J. C. Análise econômica da agricultura de precisão. In: CONGRESSO BRASILEIRO DE AGRICULTURA DE PRECISÃO, 2., 2006, São Pedro, SP. Anais... Piracicaba: USP/ESALQ, 2006. Disponível em: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/485410/analise-economica-da-agricultura-de-precisao.

MOLIN, J. P. Agricultura de precisão: números do mercado brasileiro. Boletim Técnico 03. Piracicaba: Laboratório de Agricultura de Precisão, USP/ESALQ, 2017.

VALENTE, D. S. M.; PEREIRA, G. W.; QUEIROZ, D. M. de; ZANDONADI, R. S.; AMARAL, L. R. do; BOTTEGA, E. L.; COSTA, M. M.; COELHO, A. L. de F.; GRIFT, T. Accuracy of various sampling techniques for precision agriculture: a case study in Brazil. Agriculture, v. 14, n. 12, art. 2198, 2024. Disponível em: https://doi.org/10.3390/agriculture14122198.

WEBSTER, R.; OLIVER, M. A. Geostatistics for environmental scientists. 2. ed. Chichester: John Wiley & Sons, 2007.

ZHANG, L. et al. Mapping topsoil total nitrogen using Random Forest and modified regression kriging in agricultural areas of Central China. Plants, v. 12, n. 7, art. 1464, 2023. Disponível em: https://doi.org/10.3390/plants12071464.