118. Potential of interpolation methods for modelling the spatial variability of soil parameters

L.D. Bejarano[1,*], D.A. Castañeda[2] and L.R. Amaral[1]
[1] *University of Campinas (UNICAMP), School of Agricultural Engineering, Av. Cândido Rondon, 501, 13083-875, Brazil; * laura.bejarano@feagri.unicamp.br*
[2] *National University of Colombia, School of Agricultural Sciences, St 65 #. 59A - 110, 050034, Colombia*

Abstract

Soil is highly variable, and an efficient survey requires consideration of the characteristics of the interpolation method and sampling density. The objective of this study was to analyse the impact of sampling density on the accuracy of different interpolation methods in a sugarcane area in Brazil, evaluating five interpolation techniques and three levels of sampling density. During the investigation, it was found that increasing sampling density does not always improve interpolation accuracy, while machine learning methods lost effectiveness due to the reduced number of samples. This highlights the importance of selecting an appropriate interpolation method to generate accurate maps based on the sampling characteristics and the study parameters.

Keywords: digital soil mapping, geostatistics, Sampling density, soil fertility

Introduction

The soil is a heterogeneous system whose properties vary spatially and temporally, depending on the processes governing soil formation and agricultural practices (Pusch *et al*., 2022). To understand this behaviour and improve agricultural decision-making, soil mapping is carried out based on grid sampling without consensus on the size of grid or the interpolation methods. Achieving adequate accuracy requires dense soil sampling, which demands significant effort, time, and resources (Abdel Rahman *et al*., 2021). Therefore, it is essential to use appropriate interpolation techniques and efficient sampling configurations to reduce errors and uncertainty.

Most studies rely on the importance of choosing the best interpolator but do not consider the sampling density (Pusch *et al*., 2022) even if this is a critical factor in determining costs and efficiency in the mapping process in precision agriculture. Moreover, some studies centre on the optimal sampling density but using just one interpolation method (Yang *et al*., 2020); however, the wide variety of interpolation methods available has advantages and disadvantages depending on the type and sample size, and context of the variable to interpolate.

Qu *et al*. (2024) mention that mapping performance depends on the method and sampling density. In their study, they found that for sand content prediction, ordinary kriging (OK) performed better with high sampling density, outperforming machine learning (ML) methods; however, with sparser densities, ML methods were better, suggesting that selection of the interpolation method depends on the nature of the variable, availability of samples and the covariates used in multivariate methods. This study aimed to evaluate whether increasing the sampling density improves accuracy of the interpolation methods and whether implementing more complex models improves the mapping accuracy compared to simpler methods.

Materials and methods

Study area and soil sampling

The research was conducted in a sugarcane production area in the municipality of Sales Oliveira, São Paulo, Brazil (20°51′26.17″S; 47°57′4.21″W). This area covers 70 ha, and the soil is predominantly classified as an Oxisol (USDA, 2022).

In 2023, soil sampling was conducted at a density of 6 sample/ha, collected at a depth of 0.25 m using an all-terrain vehicle equipped with an auger (Melo and Amaral, 2024). Each sample comprised six subsamples collected within a 5 m radius from the central point. In the laboratory, phosphorus content (P, mg/dm^3) was determined. Four data subsets were generated based on these samples, corresponding to three sampling densities for evaluation (1 sample/ha, 0.5 sample/ha, 0.2 sample/ha) and one external validation set comprised by the samples not included in the previous subsets (Figure 1).

Spatial dependence

The degree of spatial dependence of the variables was determined by fitting spherical, exponential, and Gaussian stochastic spatial models to experimental variograms. The best model was selected through leave-one-out cross-validation using the coefficient of determination (R^2) and root mean square error (RMSE) values. Spatial dependence (SD) was calculated using Eq. (1), applying the classification by Cambardella *et al.* (1994), where SD≤25% is considered a strong spatial dependence, 25<SD≤75% is moderate, and SD > 75% is weak. Moran's Index (MI) was also calculated to assess variable autocorrelation.

$$SD = (\text{nugget}/\text{sill}) * 100 \quad (1)$$

Interpolation methods

Five spatial interpolation methods, comprising geostatistical, deterministic, and machine learning (ML) approaches, were selected to predict the soil P content:

1. Ordinary kriging (OK): a geostatistical method that makes predictions based on stochastic spatial models, considering the data's semivariance (Oliver and Webster, 2015). The theoretical model with the best adjustment determined in the spatial dependence step was used.
2. Kriging with external drift (KED): a multivariate geostatistical interpolation method that estimates a target variable based on auxiliary variables (covariates) that show a linear correlation with the primary variable (Goovaerts and Kerry, 2010). Like OK, the best theoretical model was selected.

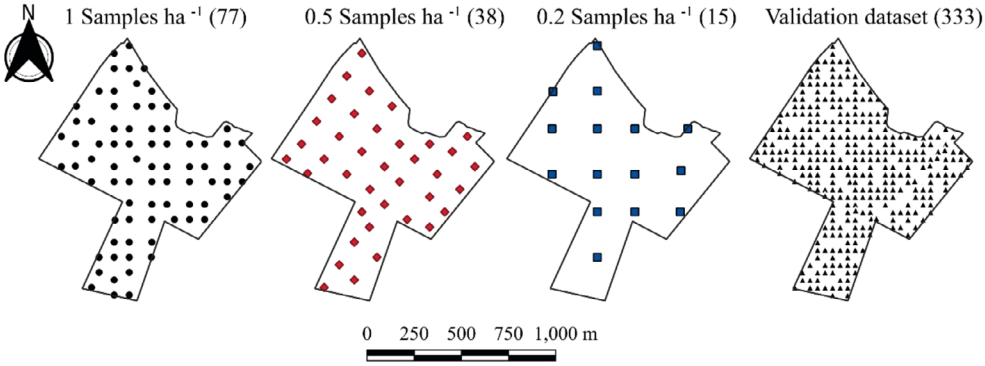

Figure 1. Spatial map of sampling dataset and validation dataset.

3. Inverse Distance Weighting (IDW): a univariate deterministic method that uses nearby samples to estimate the values of the target variable. It considers the power p, which controls the importance of the sampled values based on their distance from the points to be interpolated. In this case, the power was selected as 2, one of the most commonly used values (Wong, 2017).
4. Support vector machine (SVM): this was used in its regression version for value prediction. This machine learning method employs kernel functions to project the data into a hyperspace, where nonlinear patterns are utilised in simplified versions (Pereira *et al.*, 2022); this method is also multivariate and uses the covariates for training and prediction.
5. Spatial Random Forest (RF): a ML method based on decision trees in which segmentations are generated to minimise variance. This method uses covariates to make predictions. Since the conventional RF technique does not consider the spatial component, a covariate defined as the observations at the n nearest locations and the distances from these locations to the prediction location is used to make predictions (Sekulić *et al.*, 2020).

Covariates and selection

Soil apparent electrical conductivity (ECa) and soil apparent magnetic susceptibility (MSa) obtained with an EM38 sensor at a depth of 0.35 m were used. In addition, a digital elevation model (DEM) obtained from the PALSAR sensor was used to determine the elevation of the study area, which was re-sampled by the bilinear method to get a spatial resolution of 10 m.

The covariates were standardised using a *z*-score, and subsequently, the Pearson correlation of the covariates with the target variable was calculated with a 95% confidence, selecting those that presented $|r>0.2|$ and that did not present collinearity between them (Pearson correlation $|r>0.70|$).

Validation

Using the external validation dataset (n=333) (Figure 1), Lin's concordance correlation coefficient (LCC) and RMSE were determined for each interpolated map.

Results

The analysis of spatial dependence and autocorrelation of P content showed variations in spatial structure depending on the sampling density. As sampling distances increase, both MI and range increase and the nugget increase, indicating weaker autocorrelation (Figure 2). As seen in the figure 2, for 0.2 samples ha^{-1}, there is not enough samples to adjust the semivariogram model due to low number of samples (15). However, despite the semi-variogram adjustment requires at least 50 samples for the adjustment to be reliable, the analysis was conducted to reflect industry practice, where this method is applied despite not meeting the sample requirement.

The P content correlates significantly with ECa and Elev (r=0.49 and 0.28, respectively) at 1 sample/ha, having moderate and weak correlation. At lower densities, it correlates significantly with MSa, with r=0.54 at 0.5 sample/ha and r=0.60 at 0.2 sample/ha. Thus, for the density of 1 sample ha^{-1}, ECa and Elev were selected; for the other densities, MSa was used as a covariate.

While most interpolation methods exhibited a smoothing effect as sampling density decreased, KED and SVM retained finer spatial details due to the influence of covariates. This effect was not observed with RF, which displayed smoothed results across all densities. Even though IDW showed a smoothing effect, it introduced artefacts and unrealistic patterns in the higher densities (Figure 3). Except for SVM and IDW, all the interpolation methods evaluated showed an increase in LCCC and a decrease in RMSE as the sampling density increased. The SVM exhibited a drop in LCCC at intermediate density and increased again at higher density; on the other hand, the IDW showed a slight decrease in LCCC at higher density, a pattern also reflected by the increase in RMSE (Figure 4). These results emphasise the distinct responses of each method to varying sampling densities.

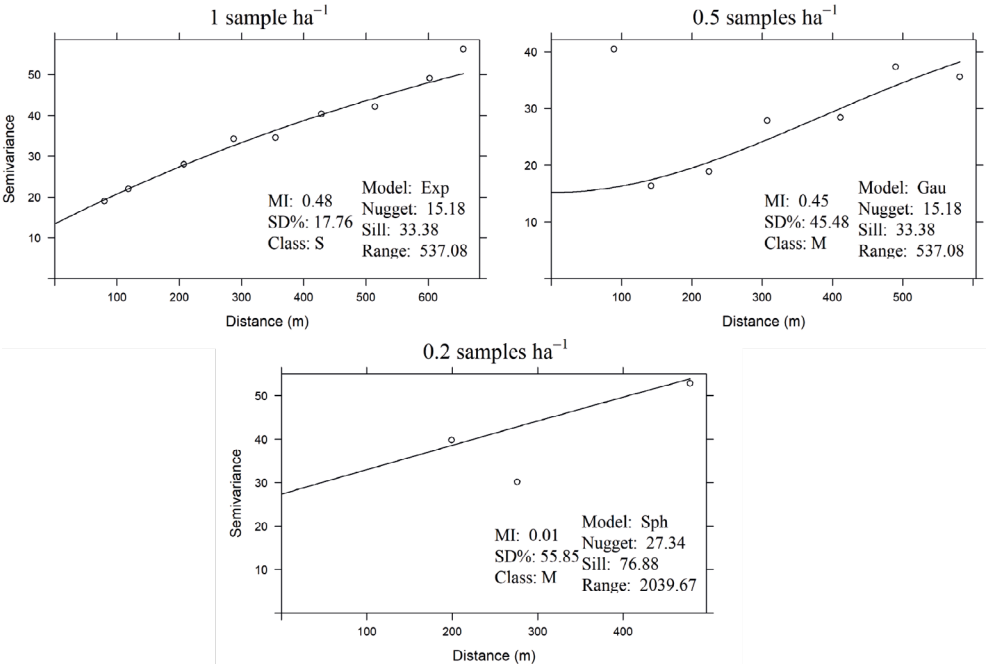

Figure 2. Semiovariograms and spatial dependency for P content. SD: Spatial dependence degree; MI, Moran's Index; Exp, Exponential model; Gau, Gaussian model; Shp, spherical model; M, moderate; S, strong.

It is also noted that the two ML methods did not outperform the other interpolation methods at any of the sampling densities evaluated (Figure 4). The most accurate results for P interpolation in the evaluation area were achieved at the density of 1 sample ha^{-1}, with OK performing as the most effective method, closely followed by KED.

Discussion

In agricultural areas, smaller farms often have small sample numbers even with dense sampling, such as the 1 sample/ha proposed in this study. For sparser sampling densities, applying certain interpolation methods becomes problematic. Oliver & Webster (2015) indicated that at least 100 sampling points are required to estimate a variogram reliably using geostatistical methods, highlighting the limitations of having too few data points. Alternative approaches to adjust the variogram, such as residual maximum likelihood (REML), can offer better accuracy with fewer data (between 50 and 100) (Kerry and Oliver, 2007). However, in some cases, like this study with 38 samples at a density of 0.5 sample ha^{-1}, sparse densities may not reach the 50-sample threshold, making geostatistical methods impractical in these scenarios.

Thus, the smoothing effect can be observed as the density decreases, represented by the increase in nugget variance (Figure 2), showing how P loses spatial autocorrelation with distance (Webster and Oliver, 2007), which leads these geostatistical methods to smooth further to reduce the uncertainty in the prediction (Oliver and Webster, 2015).
Increasing the sampling density by reducing the distance between samples is expected to resolve the variability and spatial structure better of soil variables. This was seen by the change in spatial

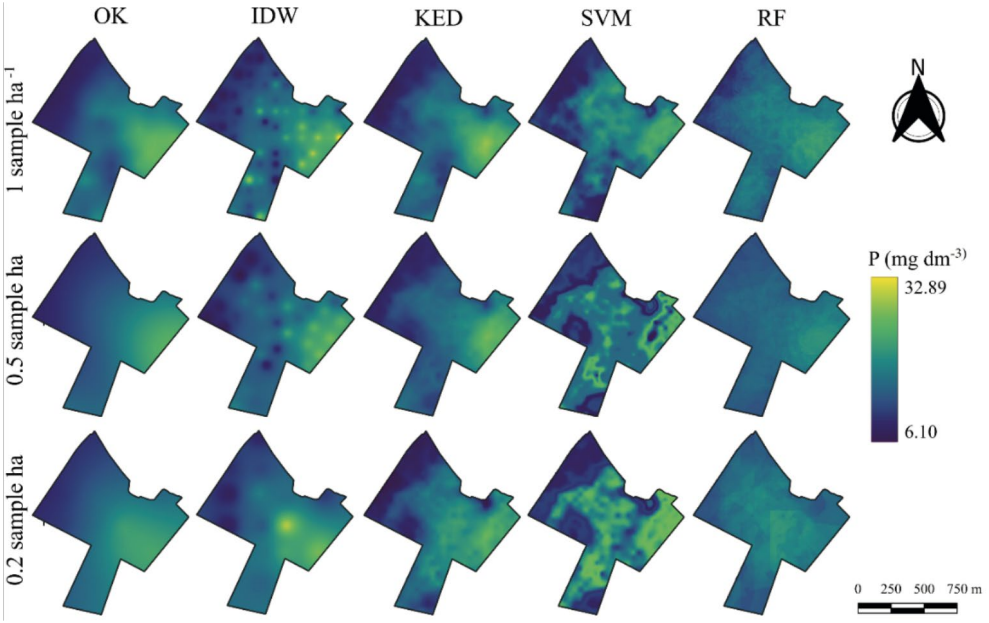

Figure 3. Spatial behaviour of phosphorus (P) content with different methods of interpolation and soil sampling densities.

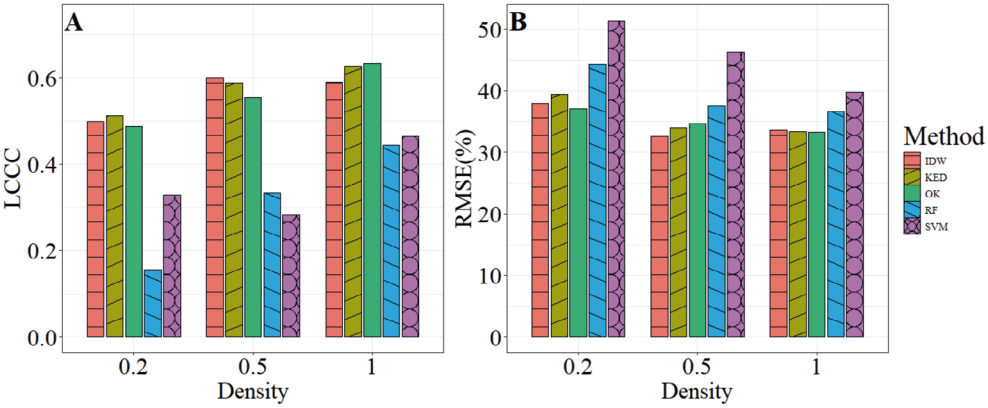

Figure 4. Validation metrics for phosphorus content (P) according to the sampling density and the interpolation method. (A) Lin's concordance coefficient correlation (LCCC); (B) root mean square error (RMSE).

dependence (SD) from moderate to strong as density increased, along with an increase in Moran's Index (MI) (Figure 2). This enhancement allows methods like OK, KED, and RF to deliver better results (Figure 4). These findings align with those of Qu *et al.* (2024), who found that increasing sampling densities improved the accuracy of geostatistical and machine learning models, with OK outperforming methods such as RF, SVM, cokriging, and Cubist. Similarly, our results confirmed OK as the most effective method at higher densities (Figure 4), with KED presenting comparable performance, emphasising the importance of both sampling density and method selection.

However, increasing density does not always improve the accuracy of interpolation models, as observed in the Inverse Distance Weighting (IDW) method (Figure 4), where a slight decrease in accuracy was noted at the highest density. This phenomenon can occur due to the creation of unnatural patterns known as bull's eyes, which may not adequately represent the soil spatial variability (Karp *et al.*, 2024) and which may be associated with the p value (set at 2) used during the evaluation, and can be improved by optimising this value to present a smoother result (Sobjak *et al.*, 2023). This aligns with the findings of Li (2010), who evaluated the prediction of organic matter content using different interpolation methods, such as OK, Universal Kriging (UK), and Regression Kriging (RK), at various sampling densities. This author found that, beyond a certain point, increasing density began to increase the RMSE of predictions, indicating that higher density does not always equate to greater interpolation accuracy.

On the other hand, ML methods did not perform well in any of the scenarios but were poor even at the sparsest density. This can be mainly attributed to the fact that these more advanced models have a limitation in the number of samples they require to be trained and outperform traditional methods (Sekulić *et al.*, 2020). However, this issue might also be related to the amount of data and correlation with covariates; Kerry *et al.* 2024 emphasised that using covariates improves model prediction when the Pearson correlation exceeds 0.40 with the soil property. In this study, ECa and MSa exhibited correlations above this threshold with P content, but Elev, which showed a weaker correlation, may have contributed less to the predictive performance. These findings underline the necessity of ensuring adequate sampling density and carefully selecting covariates with strong correlations to the target soil property to enhance the performance of multivariate models.

Conclusions

The sampling density and interpolation method are critical for generating accurate maps that reflect the spatial variability of the soil parameters. While increasing sampling density generally enhances map accuracy, its effect varies by interpolation methods. At higher densities OK, KED consistently performed well, whereas IDW and SVM showed reduced accuracy, particularly at denser configurations. These findings emphasise the need for studies on a selection of sampling configuration and interpolation methods in an agricultural context where areas tend to be smaller, and some methods may become less effective.

Acknowledgement

This project was funded by FAPESP (São Paulo Research Foundation) processes No. 2024/01557-3 and 2022/03160-8.

References

Cambardella, C.A., Moorman, T.B., Novak, J.M., Parkin, T.B., Karlen, D.L., Turco, R.F., Konopka, A.E. (1994). Field-scale variability of soil properties in central Iowa soils. Soil Science Society of America Journal, 58(5), 1501–1511. https://doi.org/10.2136/sssaj1994.03615995005800050033x

Goovaerts, P., & Kerry, R. (2010). Using ancillary data to improve prediction of soil and crop attributes in precision agriculture. In M.A. Oliver (ed.), Geostatistical Applications for Precision Agriculture. Springer, Dordrecht, pp. 167–194. https://doi.org/10.1007/978-90-481-9133-8_7

Karp, F.H.S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2024). Comparative study of interpolation methods for low-density sampling. Precision Agriculture, 25 2776–2800. https://doi.org/10.1007/s11119-024-10141-0

Kerry, R., & Oliver, M.A. (2007). Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. Geoderma, 140(4), 383–396. https://doi.org/10.1016/j.geoderma.2007.04.019

Kerry, R., Ingram, B., Oliver, M., & Frogbrook, Z. (2024). Soil sampling and sensed ancillary data requirements for soil mapping in precision agriculture II: contour mapping of soil properties with sensed z-score data for comparison with management zone averages. Precision Agriculture, 25 1212–1234. https://doi.org/10.1007/s11119-023-10108-7

Li, Y. (2010). Can the spatial prediction of soil organic matter contents at various sampling scales be improved by using regression kriging with auxiliary information? Geoderma, 159(1–2), 63–75. https://doi.org/10.1016/j.geoderma.2010.06.017

Melo, D.D., & Amaral, L.R. (2024). Replication data for: hierarchical stratification for spatial sampling and digital mapping of soil attributes. Research Data Repository of Unicamp. https://doi.org/10.25824/redu/8QITE4

Oliver, M.A., & Webster, R. (2015). Basic steps in geostatistics: the variogram and kriging. Springer International, Cham. https://doi.org/10.1007/978-3-319-15865-5

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

Pusch, M., Samuel-Rosa, A., Oliveira, A.L.G., Magalhães, P.S.G., & do Amaral, L.R. (2022). Improving soil property maps for precision agriculture in the presence of outliers using covariates. Precision Agriculture, 23(5), 1575–1603. https://doi.org/10.1007/s11119-022-09898-z

Qu, L., Lu, H., Tian, Z., Schoorl, J. M., Huang, B., Liang, Y., Qiu, D., & Liang, Y. (2024). Spatial prediction of soil sand content at various sampling density based on geostatistical and machine learning algorithms in plain areas. Catena, 234, 107572. https://doi.org/10.1016/j.catena.2023.107572

Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. Remote Sensing, 12(10), 1–29. 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

USDA. (2022). Keys to soil taxonomy, 13th edn. Available online at https://www.nrcs.usda.gov/sites/default/files/2022-09/Keys-to-Soil-Taxonomy.pdf

Webster, R., & Oliver, M.A. (2007). Geostatistics for environmental scientists, 2nd edn. Wiley, Chichester. https://doi.org/10.1002/9780470517277

Wong, D.W.S. (2017). Interpolation: inverse-distance weighting. In International Encyclopedia of Geography, 1–7. Wiley, Chichester. https://doi.org/10.1002/9781118786352.wbieg0066

Yang, L., Li, X., Shi, J., Shen, F., Qi, F., Gao, B., Chen, Z., Zhu, A.-X., & Zhou, C. (2020). Evaluation of conditioned Latin hypercube sampling for soil mapping based on a machine learning method. Geoderma, 369, 114337. https://doi.org/10.1016/j.geoderma.2020.114337