Quantifying prediction uncertainty in field-scale soil maps generated by machine learning

García Seleme, F.A., Paccioretti, P., Balzarini, M., Córdoba, M.

Statistics and Biometry Linked Group. Phytopathology and Agricultural Modelling Unity (UFyMA). National Scientific and Technical Research Council (CONICET) – National University of Córdoba (UNC). Córdoba, Argentina.

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.

This study evaluated different approaches for quantifying uncertainty in machine-learning-based digital soil mapping using soil organic matter data from several agricultural fields. Two quantile-based algorithms, Quantile Regression Neural Networks (QRNN) and Quantile Regression Forests (QRF), were compared. Uncertainty was estimated through model-based quantile predictions, bootstrap resampling, and the Area of Applicability (AOA). Results showed that QRF achieved lower prediction errors in 57% of the fields, particularly when sample sizes were small, while QRNN produced wider prediction intervals and higher coverage probabilities. Spatial patterns of uncertainty were consistent across fields. Overall, integrating accuracy and uncertainty metrics provided a more comprehensive evaluation of soil maps, enabling more informed management decisions and helping identify areas where additional sampling or more cautious recommendations may be required.

Keywords.
spatial interpolation, neural networks, regression forest, digital soil mapping

Introduction

Field-scale maps of soil properties are a key component of precision agriculture, as they are routinely used as inputs for variable-rate fertilization, zone delineation, and site-specific management. Conventional geostatistical methods such as Kriging (Oliver, 2010) are commonly used to obtain these maps. This technique provides uncertainty maps, giving individual standard deviation values for each predicted position. Additionally, Kriging can be integrated with multiple regression models (regression Kriging; Odeh et al., 1995) in order to use covariates to augment the map's accuracy.

Machine learning (ML) and deep learning algorithms were recently applied to digital soil mapping, increasing the accuracy of spatial predictions (Pereira et al., 2022; Liu et al., 2024; García Seleme et al., 2026). However, these methods do not inherently yield uncertainty estimates as Kriging does, limiting the reliability of soil maps as decision-support tools. Quantifying prediction uncertainty is essential not only to assess map quality, but also to identify areas where predictions are less reliable and data-driven decisions should be taken with caution.

The literature currently identifies three main strategies to achieve uncertainty measures for ML spatial predictions (Paranavithana et al., 2024). First, the usage of model-embedded quantile predictions, using quantile regression algorithms such as Quantile Regression Forest (QRF; Meinshausen & Ridgeway, 2006). Hengl et al. (2018) suggested a formula (equation 1) that estimates individual standard deviation values for each predicted site using QRF.

$$\widehat{SD}_i = \frac{\widehat{Q_{0.841_i}} - \widehat{Q_{0.159_i}}}{2} \quad (1)$$

where $\widehat{SD}_i$ is the predicted standard deviation for each *i* site and $\widehat{Q_{0.841_i}}$ and $\widehat{Q_{0.159_i}}$ the predicted 0.841 and 0.159 quantiles for each *i* site, respectively.

This formula estimates a 68.27% prediction interval which is used to obtain the standard deviation value. There is no published evaluation of this formula on other quantile regression algorithms, such as quantile regression neural networks (QRNN; Cannon, 2011).

Another method, mainly useful for non-quantile algorithms, applies resampling techniques on the original training dataset. Paranavithana et al. (2024), evaluating tropical paddy soils of Sri Lanka, performed bootstrapping (Efron, 1979) a hundred times on the training dataset, and fitted Regression Forest for each one, obtaining a hundred predictions for each position. With these values, estimated 90% prediction intervals. Kakhani et al. (2024) performed bootstrapping on LUCAS land use and coverage dataset and fitted Random Forest algorithms. This paper didn't report the prediction intervals, but used them to estimate several uncertainty measures.

A third method, recently developed by Meyer & Pebesma (2021), called Area of Applicability (AOA), evaluates the covariates distribution at the training dataset and compares it to the prediction sites' covariates distribution. The standardized covariates are used to create a Dissimilarity Index (DI), which is used to obtain a critical DI, i.e., the maximum dissimilarity between a predicted point and the training data that allows a confident prediction. Following this calculus, DI values are estimated for each prediction site, and then are compared to the critical DI. Sites with a DI lower than the critical are considered inside the area of applicability of the model, and those with a higher DI are considered outside of this area. The final report often includes the percentage of the original prediction area that gets included in the AOA and a map of the included/not included areas. Paranavithana et al. (2024) used this method in addition to the mentioned bootstrapping technique. All of the published applications of this method took place on large-scale mapping topics, and the method wasn't evaluated yet at field-scale. Considering the real limitations of soil sampling on agricultural fields, the AOA index could prove useful as a map validation tool.

This study evaluates strategies to quantify prediction uncertainty in machine-learning-based digital soil mapping.

Materials and Methods

To illustrate results, soil organic matter (SOM) datasets from seven agricultural fields in the Pampean region of Argentina were used: two fields were located in Santiago del Estero province, while the other five fields were located in the province of Córdoba. Each field was sampled at a density of 0.3 samples per hectare, recording several soil features (e.g., SOM, cation exchange capacity, sulphur, pH). High-density covariates were obtained for these sampling sites: crop yield, electrical conductivity at surface and sub-surface levels, and elevation. These covariates were standardized, interpolated using Kriging and then the prediction for each sampling site was extracted. The final dataset contained 217 rows.

Two quantile-based algorithms were compared for mapping: Quantile Regression Neural Networks (QRNN) and Quantile Regression Forests (QRF). Each algorithm was fitted separately for each one of the seven fields. Crop yield, surface and sub-surface level electrical conductivity and elevation were used as covariates. Overfitting was evaluated in the different algorithms, using L1 loss function if required.

Three uncertainty estimation approaches were used. First, model-embedded quantile prediction was performed in order to derive prediction intervals and standard deviation (SD) of predictions using equation 1. In second place, following the Paranavithana et al. (2024) procedure, bootstrap resampling was performed on the training dataset, obtaining 100 bootstrapped datasets, fitting QRNN and QRF algorithms with each one of them, obtaining 100 predictions for each site. The SD of these predictions was estimated. Additionally, the uncertainty metrics reported by Kakhani et al. (2024) were calculated, using 90% prediction intervals for each site. This included Prediction Interval Coverage Probability (PICP), Prediction Interval Normalized Average Width (PINAW) and Interval Score (IS). Finally, for each field the Area of Applicability (AOA) was estimated. The average Dissimilarity Index (DI) was also calculated in order to evaluate its relation with established numerical uncertainty metrics.

The accuracy of the different algorithms was evaluated with the normalized Root Mean Squared Error (nRMSE). Model validation occurred with k-fold cross-validation (k = 5). Uncertainty maps were produced with the three different methods previously detailed. The analysis was performed using software R and the packages *caret* (Kuhn, 2008), *h2o* (Fryda et al., 2024), *gstat* (Gräler et al., 2016), *meteo* (Sekulic et al., 2020), and *sf* (Pebesma & Bivand, 2023).

Results and Discussion

Across fields, QRF achieved lower RMSE than QRNN in 4 of 7 fields, mainly in those with smaller sample size (Table 1). The nRMSE values obtained in this study are similar to those obtained by Liu et al. (2022, nRMSE = 8%) with similar sampling conditions, and those reported by Guo et al. (2013, nRMSE ~10%) with similar covariates.

Table 1. Observed mean (Avg., %) and standard deviation (SD_{Obs}) of soil organic matter; normalized root mean squared error (%) and standard deviation (SD_{Pred}) of predictions obtained via Quantile Regression Forest (QRF) or Quantile Regression Neural Networks (QRNN).

Field	Avg.	SD_{Obs}	nRMSE		SD_{Pred}	
			QRF	QRNN	QRF	QRNN
A	1.9	0.38	11.7	11.0	0.30	0.33
B	2.1	0.35	12.3	11.1	0.17	0.30
C	2.1	0.35	19.8	21.2	0.18	0.30
D	1.8	0.25	14.4	16.7	0.11	0.25
E	1.5	0.27	16.4	18.9	0.08	0.26
F	2.7	0.28	9.2	6.8	0.14	0.20
G	2.3	0.26	12.7	18.1	0.06	0.27

The SD of the predictions on each field were similar to the respectively observed SD (Table 1),

most commonly lower. QRF predictions were less variable between sampling sites, while QRNN showed similar variability inside each field. The high capability of the neural networks to detect small details and non-linear relations on the data may explain this difference. However, this similarity at the SD level was not reflected in improved accuracy metrics.

Focusing on uncertainty metrics, both the quantile difference and the bootstrapping methods report generally lower prediction uncertainty on QRF algorithms (Table 2). As an exception, both SD estimations were lower for QRNN than for QRF at the field D. The quantile-difference SD is lower for QRNN at the field E by a minimum margin (0.009). Although these two metrics are considered SD estimations, their values show substantial differences for all fields, showing inconsistency. Resampling method delivered lower estimates for all fields and algorithms, excepting QRNNs on field E.

Table 2. Average on-site standard deviation of predictions, estimated by quantile difference (Q-SD) and bootstrapping (Bs-SD) for Quantile Regression Forest (QRF) and Quantile Regression Neural Networks (QRNN) methods.

Field	Q-SD		Bs-SD	
	QRF	QRNN	QRF	QRNN
A	0.167	0.514	0.144	0.299
B	0.223	0.308	0.119	0.149
C	0.274	0.627	0.082	0.322
D	0.186	0.127	0.110	0.097
E	0.179	0.170	0.109	0.232
F	0.189	0.537	0.105	0.254
G	0.179	0.372	0.092	0.278

The different uncertainty metrics used in Kakhani et al. (2024) study show higher PINAW and higher PICP values for QRNN models in five of the seven fields (Table 3). The average PINAW for QRNN models is 80% higher than the one obtained with QRF models. The average PICP on the other hand is 37.5% higher on QRNN models. In the fields B and D QRF prediction intervals got higher PICP and PINAW, but the difference between QRF and QRNN is minimal (less than 10%). The Interval Score (IS) is consistently lower (i.e., better) in QRNN models. QRNN models produce wider intervals which naturally get a better coverage of the observed values. QRF on the other hand produces more precise prediction intervals that often fail to include the observed values. These metrics must be evaluated as a whole, since they analyse different issues on the prediction intervals. The dissimilarity index was positively correlated with other uncertainty metrics, supporting its use as a practical indicator of spatial prediction reliability.

Table 3. Uncertainty metrics per field obtained from prediction intervals generated with Quantile Regression Forest (QRF) or Quantile Regression Neural Networks (QRNN). IS: interval score; PICP: prediction interval coverage probability; PINAW: prediction interval normalized average width; DI: dissimilarity index; AOA: area of applicability.

Field	QRF					QRNN				
	IS	PICP	PINAW	DI	AOA	IS	PICP	PINAW	DI	AOA
A	1.38	60.0	0.184	0.911	96.0	1.03	94.5	0.503	0.911	96.7
B	1.88	59.7	0.164	0.865	99.0	1.61	56.1	0.150	0.873	99.1
C	3.46	41.2	0.333	1.101	94.6	2.21	70.6	0.719	1.120	94.0
D	1.68	62.5	0.416	0.917	97.8	1.22	50.0	0.395	0.878	97.3
E	1.88	58.8	0.307	1.030	95.4	1.22	82.4	0.467	1.023	95.4
F	1.64	50.0	0.310	1.422	83.2	0.70	96.4	0.625	1.430	84.2
G	2.63	40.7	0.163	1.076	96.8	1.98	63.0	0.518	1.160	94.6

Although field-level sample size affected accuracy metrics—yielding RMSE values 35% lower in fields with more than 25 samples—it did not substantially influence uncertainty metrics. QRNN-derived uncertainty showed no differences across sample sizes, whereas QRF models exhibited a 25% reduction with the biggest sample sizes.

The SD maps show scale differences between QRF and QRNN (Figs. 1 and 2), particularly at fields A and F. Maps of quantile-difference and resampling SD estimations exhibited strong spatial coherence for each method, showing some differences between QRF and QRNN. Most of the maps didn't show lines or other undesired artifacts.

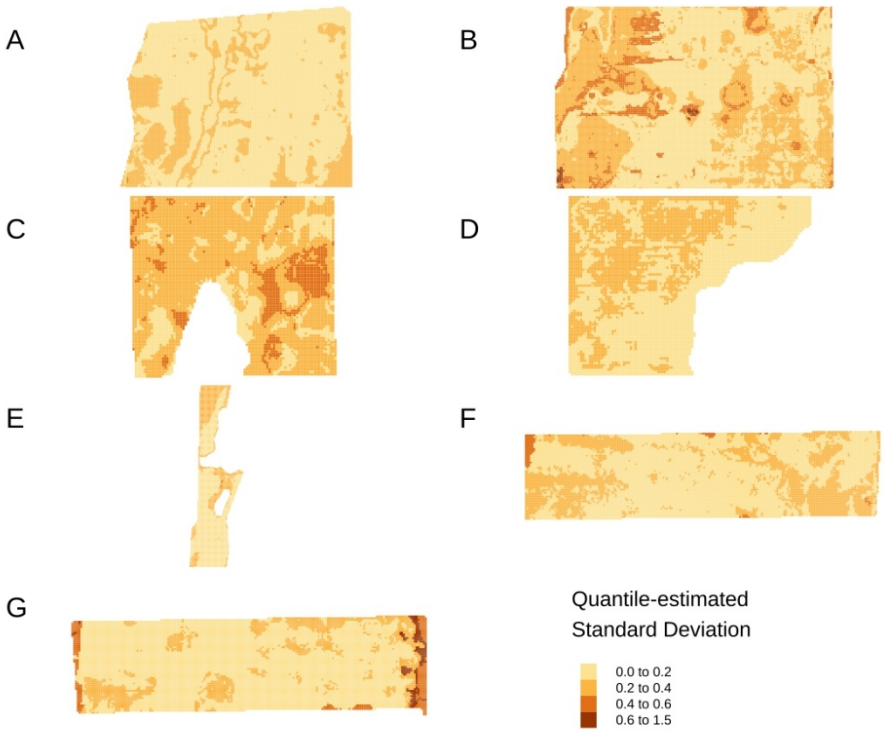

Fig 1. Standard Deviation of Quantile Regression Forest predictions estimation maps obtained using the quantile difference formula.

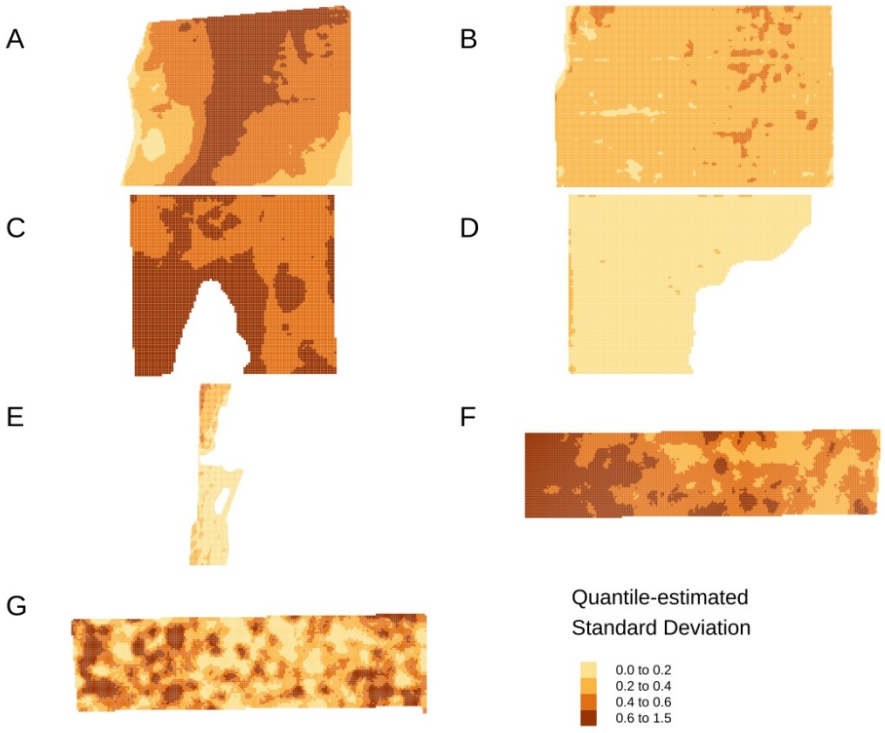

Fig 2. Standard Deviation of Quantile Regression Neural Networks predictions estimation maps obtained using the quantile difference formula.

AOA values exceeded 94% in six of the seven fields and both methods, indicating good representativeness of the training data. It was expected that the AOA didn't show differences between QRF and QRNN due to its calculus, based mainly on the covariates. The only factor that varies between methods within the AOA estimation is the different variable importance that each

method assigns to each covariate, and it was expected that this single difference did not affect the final output. However, slight differences appear on AOA values between fields despite using the same sampling method and density. Most of the area of “non-applicability” is found at the borders of the fields, with a notable exception being the field F, where a substantial portion at the SE was excluded from the AOA. The area of “non-applicability” didn’t correlate to areas of higher uncertainty as reported by the other metrics (Fig. 3).

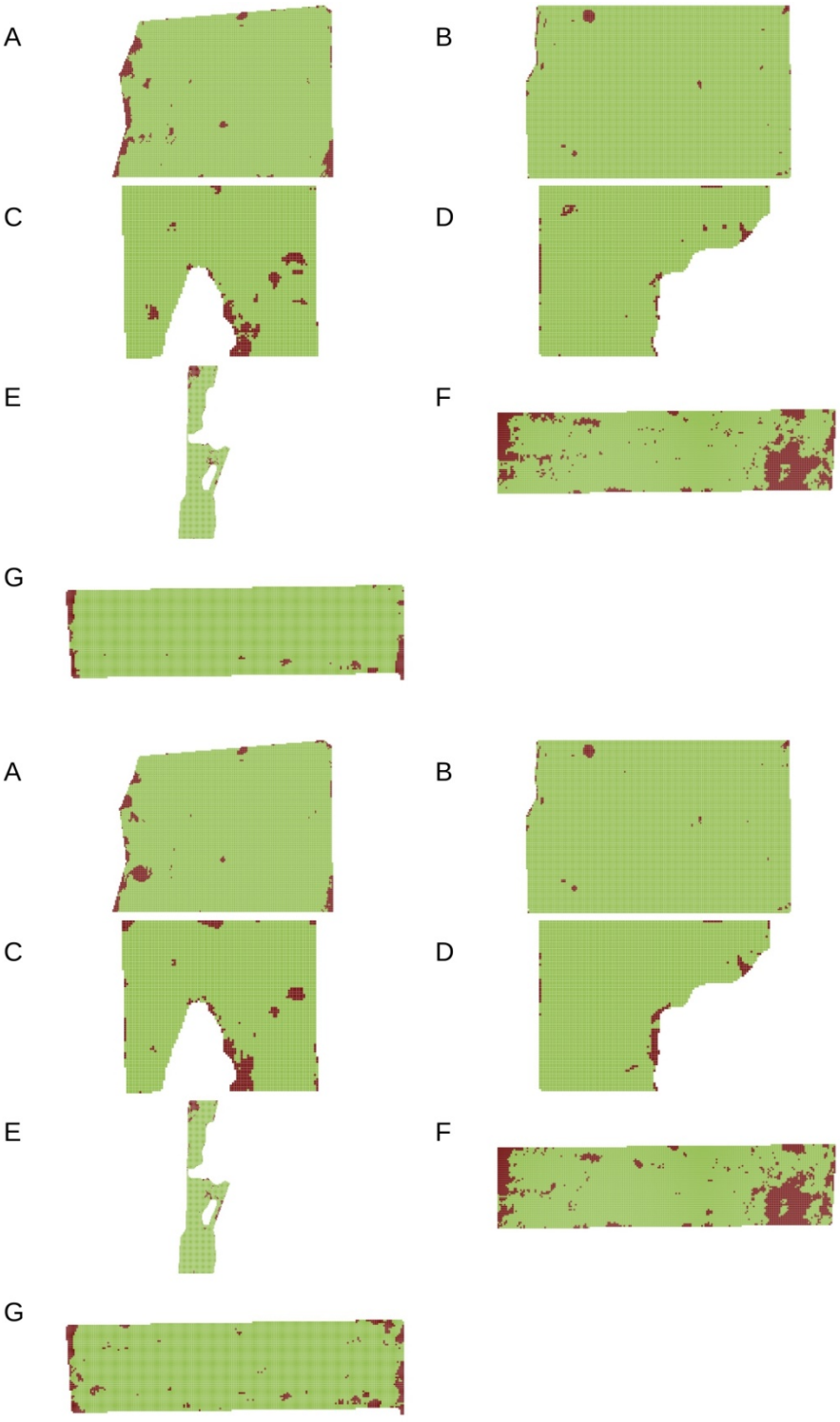

Fig 3. Area of Applicability of Quantile Regression Forest (up) and Quantile Regression Neural Networks (down) algorithms. Green: area of applicability; red: area of non-applicability.

Conclusions

These results demonstrate that integrating accuracy and uncertainty metrics provides a more comprehensive assessment of spatial prediction quality. Uncertainty-aware soil property maps enable more robust interpretation of management zones and variable-rate recommendations, helping producers and agronomists decide where predictions can be trusted and where additional sampling or conservative management is required.

Acknowledgments

We acknowledge the financial support provided by the National Agency for the Promotion of Research, Technological Development and Innovation (ANPCyT–PICT 2021-0682 and 2021-00319), the National Scientific and Technical Research Council (CONICET), and the Secretariat of Science and Technology of the National University of Córdoba (UNC) for carrying out this research. We also thank the farmers and the technical staff of Experta AGD for their valuable collaboration in collecting the experimental data.

References

Cannon, A. (2010). Quantile Regression Neural Networks: Implementation in R and application to precipitation downscaling. *Computers & Geoscience*, *37*(9), 1277–1284. https://doi.org/10.1016/j.cageo.2010.07.005

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Fryda, T., LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Ahoyoum, P., Kurka, M., Malohlava, M., Poirier, S., Wong, W., Rehak, L., Eckstrand, E., Hill, B., Vidrio, S., Jadhawani, S., … Novotny, M. (2024). *H2o: R Interface for the 'H2O' Scalable Machine Learning Platform*. https://doi.org/10.32614/CRAN.package.h2o

García Seleme, F.A., Paccioretti, P., Balzarini, M., Córdoba, M. (2026). Machine Learning Algorithms Performance for Mapping Soil Organic Matter Variability at Field Scale. *SADIO Electronic Journal of Informatics and Operations Research*, *25*(2). https://doi.org/10.24215/15146774e105

Gräler, B., Pebesma, E., Heuvelink, G. (2016). Spatio-Temporal Interpolation using gstat. *The R Journal*, *8*, 204-218. https://doi.org/10.32614/RJ-2016-014

Guo, P., Wu, W., Sheng, Q., Li, M., Liu, H., Wang, Z. (2013). Prediction of soil organic matter using artificial neural network and topographic indicators in hilly areas. *Nutrient Cycling in Agroecosystems*, *95*, 333–344. https://doi.org/10.1007/s10705-013-9566-9

Hengl, T., Nussbaum, M., Wright, M., Heuvelink, G., Gräler, B. (2018). Random forest as a generic framework for predictive model for spatial and spatio-temporal variables. *PeerJ*, *6*, e5518. https://doi.org/10.7717/peerj.5518

Jurafsky, D., Martin, J.H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. (3rd ed.). Stanford University. https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf

Kakhani, N., Alamdar, S., Kebonye, N.M., Amani, M., Scholten, T. (2024). Uncertainty Quantification of Soil Organic Carbon Estimation from Remote Sensing Data with Conformal Prediction. *Remote Sensing*, *16*(3), 438. https://doi.org/10.3390/rs16030438

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical*

Software, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Liu, Q., He, L., Guo, L., Wang, M., Deng, D., Lv, P., Wang, R., Jia, Z., Hu, Z., Wu, G., Shi, T. (2022). Digital mapping of soil organic carbon density using newly developed bare soil spectral indices and deep neural network. *CATENA*, *219*, 106603. https://doi.org/10.1016/j.catena.2022.106603

Meinshausen, N., Ridgeway, G. (2006). Quantile Regression Forests. *Journal of Machine Learning Research*, *7*(6), 983–999. https://www.jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf

Meyer, H., Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, *12*(9), 1620–1633. https://doi.org/10.1111/2041-210X.13650

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J. (1995). Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, *67*(3-4), 215–226. https://doi.org/10.1016/0016-7061(95)00007-B

Oliver, M.A. (2010). *Geostatistical Applications for Precision Agriculture.* (1st ed.). Springer. https://link.springer.com/book/10.1007/978-90-481-9133-8

Paranavithana, T.M., Karunaratne, S.B., Wimalathunge, N., Malone, B.P., Macdonald, B., Bishop, T.F.A., Ratnayake, R.R. (2024). Unravelling spatial drivers of topsoil total carbon variability in tropical paddy soils of Sri Lanka. *Geoderma Regional*, *26*, e00745. https://doi.org/10.1016/j.geodrs.2023.e00745

Pebesma, E., Bivand, R. (2023). *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

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

R Core Team. (2025). *R: A Language and Environment for Statistical Computing* (Version 4.5.1) [Software]. https://www.R-project.org/

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