Soil analysis with machine-learning-based processing of stepped-frequency GPR field measurements: Preliminary study

Chunlei Xu[a,*], Michael Pregesbauer[b], Naga Sravani Chilukuri[a], Daniel Windhager[a], Mahsa Yousefi[b], Pedro Julian[a,c] and Lothar Ratschbacher[a]

[a]Silicon Austria Labs GmbH, Intelligent Wireless Systems Division, Sciencepark 66c, Linz, 4040, Austria
[b]Geoprospectors GmbH, Wienersdorfer Straße 20-24, Traiskirchen, 2514, Austria
[c]Universidad Nacional del Sur IIIE-DIEC, San Andres 800, Bahia Blanca, Argentina

ARTICLE INFO

Keywords:
machine learning
soil analysis
stepped frequency radar
electromagnetic induction
GPR

ABSTRACT

Ground Penetrating Radar (GPR) has been widely studied as a tool for extracting soil parameters relevant to agriculture and horticulture. When combined with Machine-Learning-based (ML) methods, high-resolution Stepped Frequency Countinuous Wave Radar (SFCW) measurements hold the promise to give cost effective access to depth resolved soil parameters, including at root-level depth. In a first step in this direction, we perform an extensive field survey with a tractor mounted SFCW GPR instrument. Using ML data processing we test the GPR instrument's capabilities to predict the apparent electrical conductivity (ECaR) as measured by a simultaneously recording Electromagnetic Induction (EMI) instrument. The large-scale field measurement campaign with 3472 co-registered and geo-located GPR and EMI data samples distributed over ∼6600 square meters was performed on a golf course. The selected terrain benefits from a high surface homogeneity, but also features the challenge of only small, and hence hard to discern, variations in the measured soil parameter. Based on the quantitative results we suggest the use of nugget-to-sill ratio as a performance metric for the evaluation of end-to-end ML performance in the agricultural setting and discuss the limiting factors in the multi-sensor regression setting.
The code is released as open source and available at https://opensource.silicon-austria.com/xuc/soil-analysis-machine-learning-stepped-frequency-gpr.

1. Introduction

The availability of accurate soil information is an essential ingredient for many aspects of land management, ranging from resource-efficient agri- and horticulture to hydrological hazard mitigation (Zhuo et al., 2019). In recent years climate change has further increased the importance of land management based on timely, location-resolved data. On the one hand, widespread changes of weather patterns imply that historically established practices of land use need to be adapted in many parts of the world. On the other hand, land management techniques themselves are required to become more sustainable to reduce their contribution to environmental degradation (Zhou et al., 2019).
Depending on the spatio-temporal scale and accuracy requirements for soil information, different measurement methods are in use today. At the local level, (networks of) sensors deployed in the soil yield direct, quasi instantaneous measurements of soil parameters, including depth-resolving measurements (Francia et al., 2022). At large scales, satellite-based systems allow the estimation of soil properties based on remote-sensing data, with typical revisit times on the order of days (Liang et al., 2021; Dorigo et al., 2017). In-between these two extremes, and as the focus of this article, tractor-mounted instruments can provide soil data in a resource efficient way that allows to directly determine process parameters in time for subsequent (agri- and horticultural) equipment.
The two most established soil sensing techniques suitable to tractor mounting are Electromagnetic Induction (EMI) and Ground Penetrating Radar (GPR). For both geophysical methods, an extensive body of literature has built up over the last decades on how to relate instrument readouts to soil parameters (Pathirana et al., 2023).
EMI instruments measure the apparent electrical conductivity (ECaR) in the surrounding soil using low-frequency (VLF) electromagnetic waves. The probing frequency and the specific configuration of coils (spacing, orientation and height) determine the instrument sensitivity to subsurface locations at different depth (van't Veen et al., 2022; Schmäck et al., 2022). The deduction of soil properties of interest such as salinity water, volume fraction, bulk density, clay mass fraction and organic matter mass fraction among others, from a single measured quantity, namely the electrical conductivity, has been the subject of intense study, but typically relies on semi-empirical models (Visconti and de Paz, 2021).
GPR enables depth-resolving soil sensing by measuring the propagation of pulsed or frequency-stepped/modulated (VHF-UHF) radio frequency waves directed at the ground (Akinsunmade et al., 2019; Lombardi and Lualdi, 2019). The soil parameters determine the relative dielectric permittivity, the conductivity and potentially the magnetic susceptibility distribution in the ground, which in turn shape the radar signal propagation. The target application determines the choice of the TX and RX antenna configurations, including the options: air-coupled vs. ground-coupled; with or without

*Corresponding author
chunlei.xu@silicon-austria.com (C. Xu)
https://geoprospectors.com/ (M. Pregesbauer)
ORCID(s): 0000-0001-8704-4214 (C. Xu); 0000-0002-6308-4497 (P. Julian); 0000-0002-2631-0977 (L. Ratschbacher)

buried reflector; monostatic, fixed-offset or variable-offset. In GPR measurements, the radar center frequency constitutes a compromise, as higher frequencies generally lead to high spatial resolution, but implies lower signals from interfaces buried deeper below the surface, as well as, higher sensitivity to surface roughness. In the agricultural context GPR radar techniques have been studied for extracting a multitude of soil parameters, including soil layer thickness, soil density and Soil Water Content (SWC) (Klotzsche et al., 2018; Huisman et al., 2003), as well as, for root and seed localization (Mapoka et al., 2020; Sun et al., 2023), among many others.

For both instruments and their respective parameter estimation methods, it is commonly required to perform site-specific calibration since first principle approaches without free parameters tend to yield precise results only for well-defined laboratory settings. Taking the example of SWC, which represents the most studied soil parameter (Pathirana et al., 2023), analysis models have been developed based on simulation and/or controlled laboratory measurements with prepared soil compositions (Tran et al., 2012). In field measurements, model-based estimations of soil parameters from the aggregate EMI or GPR measurements encounter challenges due to the complexity of soil composition and structure, vegetation and surface morphology, as well as, the presence of machinery adjacent to the instrument.

In recent years ML approaches have started to supplement the traditional approaches in GPR analysis, exemplified by the use of deep neural networks for direct velocity inversion with GPRNet (Leong and Zhu, 2021) and its further development in the agricultural context for depth resolved SWC profiling (Li et al., 2023). *Terry et al.* (Terry et al., 2023) conducted field measurement with multiple geophyiscal instruments (including GPR) over a substantial field size with controlled moistening conditions and used machine learning to find optimized combinations of classical signal processing features for moisture estimation. The measurements were conducted using low frequency radar and profited from a clear layering in the soils structuring and probed soil over depths of ten meters and beyond. On the more challenging task of depth-resolved soil parameter estimation at the root zone level (from the surface to several tens of centimeters of depth) promising results were demonstrated by *Filardi et al.* (Filardi et al., 2023) using higher frequency SFCW radar and ML based processing based on data from a set of several tens of field positions. The employed field sampling clearly is the gold standard in soil parameter characterizations, but is typically incompatible with the large datasets required for thorough development and validation of data-driven methods.

In this paper we thus follow the idea of *Jonard et al.* (Jonard et al., 2013) to record a large field area with both EMI and GPR instruments simultaneously. Rather than using the GPR data to predict a specific soil property (e.g. such as soil moisture) that might suffer from aforementioned calibration effects or require substantial manual effort for measurement acquisition (e.g. from gravimetric and chemical analysis), we directly predict the co-measured EMI values that serve as a proxy for relevant soil parameters. The primary aim of this work is to investigate the capability of end-to-end machine learning methods to extract soil properties from high frequency GPR measurements in a large scale field campaign setting. Our work is based on SFCW GPR measurements in the comparatively high frequency range from 1.3 to 2.9 GHz, which suffer from a high sensitivity to surface morphology and strong attenuation, but carry the potential of further depth-resolved parameter extraction at the root level. A secondary aim of this publication is to make the obtained dataset available for further development of ML based methods in the field of precision agriculture.

2. Materials and Methods

2.1. Sensor Specification

For the measurement campaign a *Toro Reelmaster 5510* tractor was equipped with the EMI instrument *Topsoil Mapper* by *Geoprospectors* in the front and a newly developed SFCW GPR radar instrument in the back (see Fig. 1) and included into a data acquisition solution with GPS information (*Stonex s10a*) for geo-location and time stamping. The EMI instrument operating at a frequency of 9 kHz and horizontal coil alignment was mounted with a clearance of 20 cm above ground and recorded raw apparent electrical conductivity (ECaR) values with a sampling rate of 5 Hz. *Raw* in this case refers to the fact that the mounting of the otherwise self-calibrating EMI instrument on the mower introduced a baseline offset due to the presence of high conductive material. The GPS information was recorded at a rate of 1Hz.

2.1.1. SFCW GPR

The bistatic, air-coupled GPR radar setup features a single channel SFCM radar prototype by *Geoprospectors* with fixed-offset Vivaldi antennas for the transmitter and the receiver side. The detailed design parameter of the setup shown in Fig. 1 are listed in Table 1.

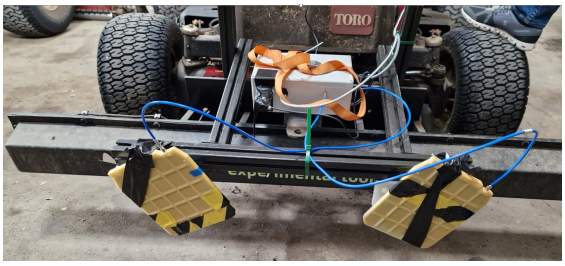

Figure 1: Experimental setup with the SFCW radar with air-coupled, fixed offset, Vivaldi antennas mounted directly behind the transversal EMI instrument bar on a *Toro Reelmaster 5510* tractor.

2.2. Study Site and Data Source

The study sites for the field campaign were fairway 14 and fairway 16 of the Fontana Golf Club south of Vienna, Austria, located at (47°58’29"N, 16°18’25"E,) with ∼220m

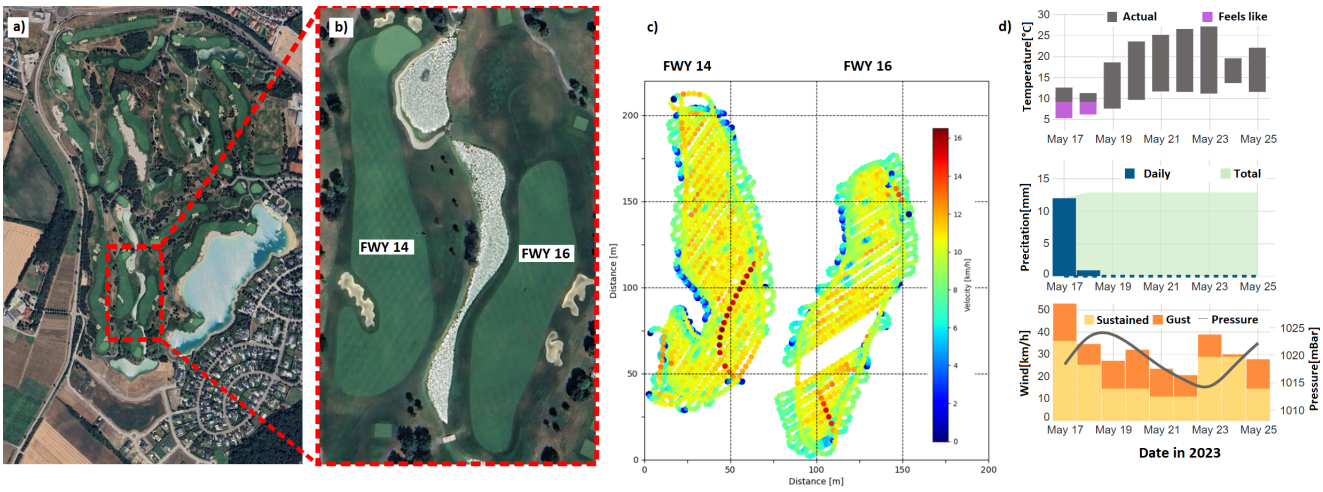

Figure 2: Site of the field campaign. (a) Overview satellite map of the Fontana Golf Club south of Vienna, Austria. Detailed satellite imagery (b) and velocity map (c) of fairways FWY14 (left) and FWY16 (right) used in study. (d) Weather conditions in the week leading up to the measurements taken on 25.5.2023.

Table 1
Stepped Frequency Radar Setup Parameters.

Frequency	1.3-2.9 GHz
Total sweep time	70 ms
Number steps	400
Max. radar location sampling rate	10 Hz
Antenna separation	60 cm at feed points
Ground clearance	15 cm
Angle to vertical	23 ° for both antennas
Antenna Gain	~7 dBI const. over bandwidth

of elevation. The data was recorded on 25.05.2023 following several days of dry, warm and windy weather. As shown in Fig. 2 the fairways were successively covered in parallel lanes with a separation on the order of 1.5 meters and at driving speeds up to 14km/h. The EMI, GPR and GPS recordings were co-registered and re-sampled at the temporal sampling rate of the GPR instrument.

2.3. Machine Learning

To quantitatively assess the capability of ML-based processing methods to estimate soil parameters from the high-resolution SFCW GPR, we formulate the problem as a regression task. In this supervised setting, the continuous and scalar value of the raw apparent electrical conductivity ECaR as measured by a simultaneously recording EMI instrument (dependent variable) is estimated from the SFCW measurement vector that represents the independent variables. The full code is released as open source and available at https://opensource.silicon-austria.com/xuc/soil-analysis-machine-learning-stepped-frequency-gpr.

2.3.1. Data Preprocessing

In the first step, the original data set (Fig. 2 (c)) is spatially and temporally filtered to reduce the potential impact on the instrument responses of tractor turns and velocity changes. Specifically, data points associated with turning paths and other non-parallel sampling paths at the beginning and the end of the measurement sequence, as well as those corresponding to extremely high or low velocities, are removed from the dataset. For the supervised regression task, a dataset with input and output pairs, $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_i, y_i)..., (\mathbf{x}_N, y_N)\}$ is created with the predictor variable vectors $\{\mathbf{x}_i\}$ being the single channel radar readings at the 400 frequency steps of the SFCW radar and scalar EMI ECaR as the scalar target variables $\{y_i\}$ for each location i. Further, the outliers corresponding to the upper and lower 0.5% of EMI ECaR values are removed. The final data set has $N = 3472$ entries with single channel radar values at each of the 400 frequency steps, the measured EMI ECaR values, as well as geographic coordinates associated with calculated tractor speed.

In the second step, the observed predictor variables are normalized to obtain the independent features that serve as input for our ML regression models. This is done by subtracting for each frequency step of a measurement sample the mean value over all samples (slow time), following a similar approach as in (Filardi et al., 2023). The mean values of the training dataset are hereby used to also normalize the test dataset features. Furthermore, a transformation that yields zero mean and standard deviation of one (for the target variables on the training dataset) is applied to all target variables with the purpose to improve learning during the regression task.

2.3.2. Regression models and performance metrics

In this study, the classical ML based regression methods Linear Regression (Linear), Random Forest Regression (RFR) and k-nearest-neighbor-based regression (KNR) are employed to predict the ECaR value from the SFCW GPR. The Mean Squared Error (MSE) is used as the loss function for all models.

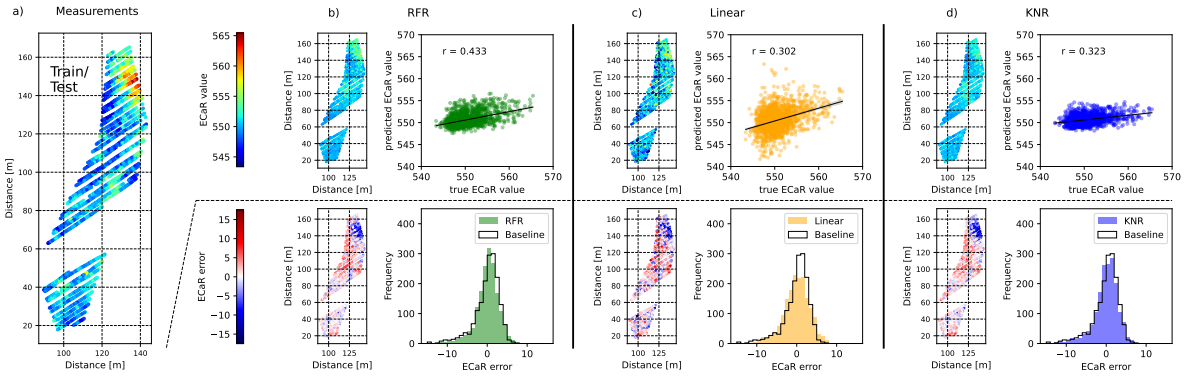

Figure 3: Results of the prediction of EMI values based on GPR data for various ML models. A geo-randomized five-fold cross-validation of the measurements of fairway 16 (a) is used for training and testing. Results of random forest regression, linear regression and k-nearest neighbor regression are shown in (b)-(d), respectively. Prediction maps and the scatter plots (with linear curve fit and sample Pearson correlation coefficient r) are shown in the top row; error maps and error histograms are shown in the bottom row. The baseline results outlined in the histogram plot represent the error with respect to a trivial uniform prediction with a value corresponding to the average of all training measurements. Note: The raw apparent electrical conductivity values (ECaR) contain a baseline shifted due to the proximity of high-conductivity materials present in the mower.

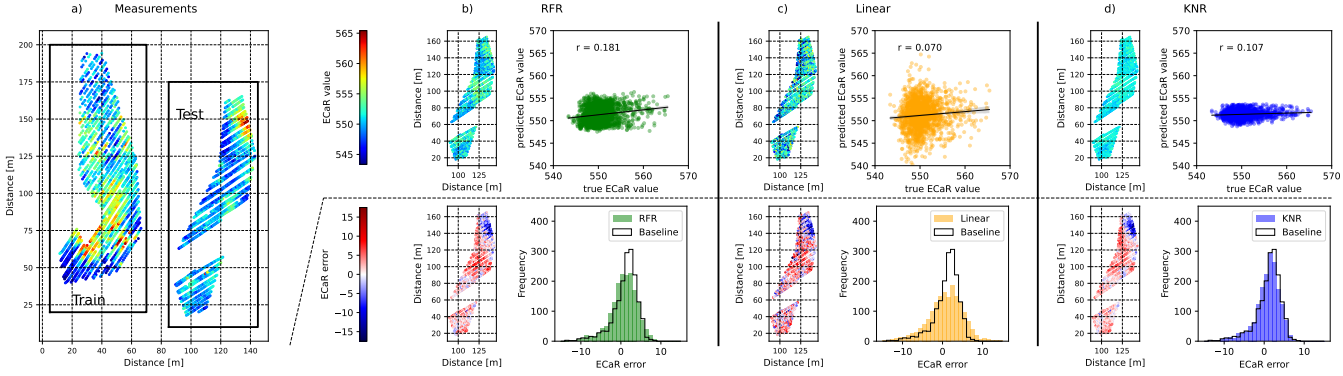

Figure 4: Results of the prediction of EMI values based on GPR data for various ML models. The models are trained on fairway 14 and tested on fairway 16 (a). The presentation of results for random forest regression (b), linear regression (c) and k-nearest neighbor regression (d) follows the description in Fig. 3.

For the assessment of the prediction performance and the relative comparison, the Mean Squared Error (MSE), the Mean Absolute Error (MAE) and the sample Pearson correlation coefficient (r) are computed for all models and data evaluation scenarios. In each data evaluation scenario, the RFR and KNR models were first optimized through hyperparameter grid searches in (repeated) nested cross-validation (CV) settings on the datasets (Bates et al., 2023; Bradshaw et al., 2023; Varma and Simon, 2006; Krstajic et al., 2014). The best model architectures for KNR and RFR were selected according to the MSE metric. The Linear model and the respective best-performing KNR and RFR model architectures were then trained and evaluated in terms of performance and error estimation by employing repeated cross-validation (Jiang and Wang, 2017). In this way the average values of performance metrics and their (one-sigma) confidence intervals as shown in Table 2 are calculated based on 50 evaluations that arise from the ten repetitions of the 5-fold cross validation.

3. Results

Based on the data of the two fairways described in Sect. 2.2, two different scenarios are studied.

In the first scenario, only data from fairway 16 is used. The RFR and KNR models are optimized for hyperparameters via repeated nested CV. The best architectures for each ML model are then trained and tested using repeated 5-fold CV, where the data splits of the fairway 16 are performed randomly by completely disregarding the spatial location information. In this way, train and test data points are expected to be in close spatial proximity to each other (cf. Fig. 3 a)).

In the second scenario, fairway 14 is used for traing and fairway 16 is used for testing. The RFR and KNR model are optimized for hyperparameters via repeated nested CV on data from fairway 14. The ML models with best architectures are selected for repeated 5-fold CV with 4-fold of data taken from fairway 14 for training and the entire data of fairway

	5-fold cross-val. fwy 16				Train on fwy14, Test on fwy16			
	RFR	Linear	KNR	Baseline	RFR	Linear	KNR	Baseline
MAE	**2.11±0.10**	2.61±0.12	2.22±0.10	2.31±0.11	2.72±0.06	3.57±0.11	**2.60±0.03**	2.70±0.02
MSE	**8.10±0.86**	12.15±1.92	8.93±0.88	9.94±1.07	11.99±0.33	21.66±1.38	**10.85±0.18**	11.23±0.09
r	**0.43±0.06**	0.33±0.04	0.32±0.05	n.a.	**0.18±0.02**	0.08 ± 0.02	0.11±0.02	n.a.

Table 2
Summary of prediction performance data based on random forest regression (RFR), linear regression (Linear) and k-nearest neighbor regression (KNR) models for the two studied train/test scenarios.

	Train and Test on Fwy 16			Train on Fwy14, Test on Fwy 16			Ground Truth Fwy 16
	RFR	Linear	KNR	RFR	Linear	KNR	
Range	17.62±0.20	18.18±1.09	19.09±0.31	26.10±1.29	17.41±2.30	21.03± 1.42	15.82
Nugget	0.67±0.04	5.93±0.57	0.50±0.02	1.83±0.26	11.35±1.06	0.54±0.05	1.51
Sill	1.78±0.04	7.96±0.92	1.13±0.02	3.88±0.47	13.00±1.40	0.81±0.07	9.71
NSR	**0.38±0.02**	0.75±0.02	0.44±0.01	**0.47±0.03**	0.87±0.04	0.66±0.02	0.16

Table 3
Variogram parameters as extracted from spherical fits to the prediction and ground truth variograms in Fig. 5. The correlation of the nugget-to-sill ratio (NSR) with the performance metrics in Table 2 shows that the NSR could be used as an indicator for prediction performance, which does not require the availability of ground truth measurements for its evaluation.

16 evaluated for testing. As a consequence, there is a clear spatial separation between the train and test data points (cf. Fig. 4 a)).

The results of the first scenario are graphically summarized in Fig. 3 b), c) and d) for the RFR, the Linear and the KNR models respectively. By comparing the heatmap plots of predicted EMI values for all three models (upper plots in Fig. 3 b), c) and d)), with the measured EMI value map (Fig. 3 a)), it is evident that predictions contain features with similar geographic sizes and some correlation in terms of locations, but with significant lower dynamic range. The scatter plots of predicted vs. true EMI values also highlight the fact that high and low values are challenging to predict, but also show the significant variances in the prediction error. Whereas the fitted linear slope in the scatter plot data is highest for the Linear model, the RFR exhibits a significantly lower scattering of the prediction error. Overall the Pearson correlation coefficient is highest for the RFR with a value of $r = 0.425$, which we thus consider the best performing model. This is also reflected in the error histogram plot, where the RFR leads to a slight narrowing with respect to the baseline, which naively assumes the mean of the target values of the training data as the prediction value for the test data. The summarized error metrics in Table 2, show that, while the RFR clearly performs best, the ranking of models depends on the chosen error metric.

The results of the second scenario with geo-location separated training and test data are presented in Fig. 4. The RFR regressor model features the highest sample Pearson correlation coefficient for EMI value prediction based on the GPR radar data, but ranks second in terms of MAE and MSE performance to the KNR, which is the only model beating the baseline (see Table 2). In the second scenario the overall performance is degraded by the significantly reduced slope of the predicted vs. the true EMI data for all models when compared to the first performance.

As additional metric for characterizing the performance of the regression models, variograms have been calculated from the prediction values and fitted by spherical covariance models for parameter extraction (Müller et al., 2022). Variograms, which quantify the spatial variability of a soil property under study, play an important role in geostatistical estimation, in particular, for the optimal interpolation of measured data points (Mzuku et al., 2005; Cressie, 1993). In Fig. 5 and Table 3 the predicted nugget (describing the (extrapolate) variance between samples at vanishing distances), sill (corresponding to the variance of samples at (infinitely) large distances), range and Nugget-to-Sill Ratio (NSR) are compared to those of the measured EMI data. Motivated by the fact that the nugget value of a variogram, which describes the (extrapolate) variance between samples at vanishing distances, has a lower bound by the intrinsic variance of the measurement/estimation method (Rossel and McBratney, 1998), we test the hypothesis that the nugget-to-sill (NSR) ratio could serve as a meaningful metric for model performance evaluation (Karl and Maurer, 2010). A low ratio of the nugget to the sill indicates a strong spatial dependence of the predicted values and hence could indicate better model performance. Importantly, and in contrast to the other error metrics, the calculation of the NSR does not require a ground truth. Indeed, the measured ground truth data features the lowest NSRs, followed by the RFR models, followed by the KNR and the Linear regressor. The ranking according to NSR hence agrees with the one according to the sample Pearson coefficient and correlates well with the ones according to MAE and MSE (cf. Table 3 and Table 2).

4. Discussion

From a practical perspective, the first scenario with spatially inter-dispersed training and test data, is relevant for

soil parameter interpolation, where information from more frequent readings of one sensor is used to improve the interpolation of less frequently available soil data from another sensor (e.g. manual sampling). The second scenario is representative of the requirements for standalone instrument for soil analysis, which gets initially calibrated on specific data and subsequently performs soil analysis measurements in the field. For the study case of predicting EMI values from GPR data, significantly better results were achieved in the first scenario than in second scenario, as quantified by the metrics in Table 2 and the NSR error metrics in Table 3. We attribute the observed performance behaviour mainly to two aspects of the multi-sensor regression setting of our study. The major contribution limiting the regression performance is the reduced dynamic range of the predicted values as compared to the measured values. In the regression setting with least square loss, truly absent (or severely limited) correlation between variables will lead to an estimate at (or close to) the mean of the distribution of the target variable. For the first scenario, spatially inter-dispersed train/test scenario, a fundamental lack of correlation is expected due to the different sensing depth of the two sensors. Whereas the EMI sensing volume is estimated to extend to a depth of 90 cm, the high frequency GPR signals are strongly attenuated at this depth. The sensitivity analysis based on Lasso feature selection in (Filardi et al., 2023), for instance, shows that for the soil moisture parameter of that study, radar frequencies from 1.4 GHz to 2.0 GHz provide little to no information on soil water content at 0.4 m depth compared to frequencies from 0.4 GHz 1.4 GHz. For the second scenario - the spatially segregated case - we expect a further aspect of the difference in sensor modalities to contribute in a significant way. As outlined in the introduction, the EMI and GPR instruments can be understood to effectively respond to different physical soil properties, apparent electrical conductivity and the dielectric permitivity (profile), which in turn depend to a different degree on soil parameters, including, among others, SWC, salinity, and clay content. With insufficient training data to cover relevant soil parameter distributions, the machine learning model will lack the generalisation capabilities to overcome distribution shifts in training and test data. For the substantial but still limited data set underlying this study, the effect is expected to contribute to the degraded performance for the second scenario with respect to the first scenario. In addition, several potential technical contributions are worth mentioning. While certain considerations have been taken to reduce the dependencies of our GPR and EMI measurements on varying experimental conditions (cf. Section 2.3.1), residual or unaccounted effect, in particular due to the vehicle mounting, could still contribute to the observed regression performance limitations. Furthermore, imperfections in the instruments, including potential effects that are due to saturation of the radar measurements in our data set, could also play a role and will be the subject of future work.

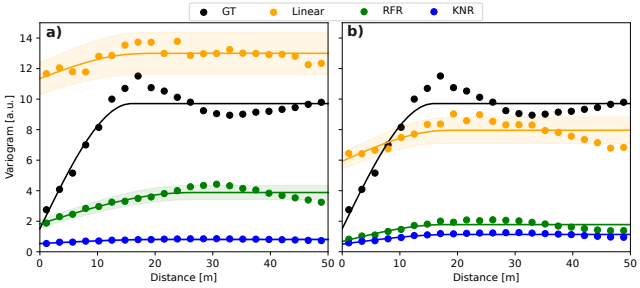

Figure 5: Variograms of the first scenario with training and testing on data from fairway 16 (LEFT), and the geographically-split second scenario with training on fairway 14 and testing on fairway 16 (RIGHT).

5. Conclusion

This study has investigated the capabilities and the limitations of the end-to-end machine learning techniques to perform soil analysis from FMCW GPR measurements. In contrast to previous studies, e.g. by *Castrignanò et al.* (Castrignanò et al., 2017), which employed sensor fusion and geostatistical methods to compare EMI and GPR radar data and to delineate homogeneous zones in the field, our data-driven ML approach does not rely on domain-specific expert knowledge for calibration and enables quantitative performance assessments. In this context, the nugget-to-sill ratio, which strongly correlates with several standard ML figures of merit, has been identified as a promising indicator of performance that can be computed without ground truth measurements.

To establish high-frequency FMCW GPR radar as a geophysical instrument for precision farming, e.g. enabling depth-resolved soil analysis at root level, more application-specific data sets for supervised machine learning are needed. Multi-sensor field campaigns, such as the one reported here, are important to collect the large-scale labelled measurements of new instruments with manageable effort. The integration of additional sensors such as height sensors and optical cameras could enable the identification of undesired instrument sensitivities, including to driving conditions, surface morphology and vegetation. Further developments towards small form factor devices with data up-link, will enable future data collections to benefit from a roll-out into routine land management activities. Instrument data under a large variety of soil types and conditions can then, for example, be assessed and labelled in an automatized fashion by correlating with data from other sensor modalities. Regarding soil water content (SWC) estimation at low soil depth, low-resolution satellite surveying data can effectively label uniform fields over extensive spatial areas (Gardin et al., 2021), wheras UAV-based measurements have demonstrated their capability to provide data for correlating SWC measurements at smaller spatial scales (Bertalan et al., 2022).

Acknowledgement

This work has been supported by Silicon Austria Labs (SAL), owned by the Republic of Austria, the Styrian Business Promotion Agency (SFG), the federal state of Carinthia, the Upper Austrian Research (UAR), and the Austrian Association for the Electric and Electronics Industry (FEEI).

CRediT authorship contribution statement

Chunlei Xu: Writing - review & editing, Visualization, Data curation, Machine learning analysis, Methodology, Conceptualization. **Michael Pregesbauer:** Writing - review & editing, Validation, Hardware Setup, Field Campaign Measurements, Methodology, Conceptualization. **Naga Sravani Chilukuri:** Writing - review & editing, Data curation, Machine learning analysis. **Daniel Windhager:** Writing - review & editing, Data curation, Software, Hardware Setup. **Mahsa Yousefi:** Writing - review & editing, Hardware Setup. **Pedro Julian:** Writing - review & editing, Supervision, Conceptualization. **Lothar Ratschbacher:** Writing - original draft, Writing - review & editing, Visualization, Data analysis, Methodology, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The full code is released as open source and available at https://opensource.silicon-austria.com/xuc/soil-analysis-machine-learning-stepped-frequency-gpr.

References

Akinsunmade, A., Tomecka-Suchoń, S., and Pysz, P. (2019). Correlation between agrotechnical properties of selected soil types and corresponding gpr response. *Acta Geophysica*, 67.

Bates, S., Hastie, T., and Tibshirani, R. (2023). Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*.

Bertalan, L., Holb, I., Pataki, A., Szabó, G., Szalóki, A. K., and Szabó, S. (2022). Uav-based multispectral and thermal cameras to predict soil water content – a machine learning approach. *Computers and Electronics in Agriculture*, 200:107262.

Bradshaw, T., J., Huemann, Z., Hu, J., and Rahmim, A. (2023). A guide to cross-validation for artificial intelligence in medical imaging. *Radiology: Artificial Intelligence*, 5(4):e220232.

Castrignanò, A., Buttafuoco, G., Quarto, R., Vitti, C., Langella, G., Terribile, F., and Venezia, A. (2017). A combined approach of sensor data fusion and multivariate geostatistics for delineation of homogeneous zones in an agricultural field. *Sensors 2017, Vol. 17, Page 2794*, 17:2794.

Cressie, N. A. C. (1993). Geostatistics. In *Statistics for Spatial Data*, chapter 2, pages 27–104. John Wiley and Sons, Ltd.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S. I., Smolander, T., and Lecomte, P. (2017). Esa cci soil moisture for improved earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment*, 203:185–215.

Filardi, V., Cheung, A., Khan, R., Mangoubi, O., Moradikia, M., Zekavat, S. R., Wilson, B., Askari, R., and Petkie, D. (2023). Data-driven soil water content estimation at multiple depths using sfcw gpr. *2023 IEEE International Opportunity Research Scholars Symposium, ORSS 2023*, pages 86–91.

Francia, M., Giovanelli, J., and Golfarelli, M. (2022). Multi-sensor profiling for precision soil-moisture monitoring. *Computers and Electronics in Agriculture*, 197:106924.

Gardin, L., Chiesi, M., Fibbi, L., Angeli, L., Rapi, B., Battista, P., and Maselli, F. (2021). Simulation of soil water content through the combination of meteorological and satellite data. *Geoderma*, 393:115003.

Huisman, J. A., Hubbard, S. S., Redman, J. D., and Annan, A. P. (2003). Measuring soil water content with ground penetrating radar: A review. *Vadose Zone Journal*, 2:476–491.

Jiang, G. and Wang, W. (2017). Error estimation based on variance analysis of k-fold cross-validation. *Pattern Recognition*, 69:94–106.

Jonard, F., Mahmoudzadeh, M., Roisin, C., Weihermüller, L., André, F., Minet, J., Vereecken, H., and Lambot, S. (2013). Characterization of tillage effects on the spatial variation of soil properties using ground-penetrating radar and electromagnetic induction. *Geoderma*, 207-208:310–322.

Karl, J. W. and Maurer, B. A. (2010). Spatial dependence of predictions from image segmentation: A variogram-based method to determine appropriate scales for producing land-management information. *Ecological Informatics*, 5:194–202.

Klotzsche, A., Jonard, F., Looms, M., van der Kruk, J., and Huisman, J. (2018). Measuring soil water content with ground penetrating radar: A decade of progress. *Vadose Zone Journal*, 17:1–9.

Krstajic, D., Buturovic, L., Leahy, D., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6:10.

Leong, Z. X. and Zhu, T. (2021). Direct velocity inversion of ground penetrating radar data using gprnet. *Journal of Geophysical Research: Solid Earth*, 126:e2020JB021047.

Li, Z., Zeng, Z., Xiong, H., Lu, Q., An, B., Yan, J., Li, R., Xia, L., Wang, H., and Liu, K. (2023). Study on rapid inversion of soil water content from ground-penetrating radar data based on deep learning. *Remote Sensing*, 15:1906.

Liang, J., Liang, G., Zhao, Y., and Zhang, Y. (2021). A synergic method of sentinel-1 and sentinel-2 images for retrieving soil moisture content in agricultural regions. *Computers and Electronics in Agriculture*, 190:106485.

Lombardi, F. and Lualdi, M. (2019). Step-frequency ground penetrating radar for agricultural soil morphology characterisation. *Remote Sensing*, 11(9).

Mapoka, K. O., Birrell, S. J., and Eisenmann, D. J. (2020). Manual and ground penetrating radar field measurements of field-corn spacing, planting depth, and furrow feature identification. *Journal of Applied Geophysics*, 180:104125.

Mzuku, M., Khosla, R., Reich, R., Inman, D., Smith, F., and MacDonald, L. (2005). Spatial variability of measured soil properties across site-specific management zones. *Soil Science Society of America Journal*, 69(5):1572–1579.

Müller, S., Schüler, L., Zech, A., and Heße, F. (2022). Gstools v1.3: A toolbox for geostatistical modelling in python. *Geoscientific Model Development*, 15:3161–3182.

Pathirana, S., Lambot, S., Krishnapillai, M., Cheema, M., Smeaton, C., and Galagedara, L. (2023). Ground-penetrating radar and electromagnetic induction: Challenges and opportunities in agriculture. *Remote Sensing 2023, Vol. 15, Page 2932*, 15:2932.

Rossel, R. A. V. and McBratney, A. B. (1998). Soil chemical analytical accuracy and costs: implications from precision agriculture. *Australian Journal of Experimental Agriculture*, 38:765–775.

Schmäck, J., Weihermüller, L., Klotzsche, A., von Hebel, C., Pätzold, S., Welp, G., and Vereecken, H. (2022). Large-scale detection and quantification of harmful soil compaction in a post-mining landscape using multi-configuration electromagnetic induction. *Soil Use and Management*, 38(1):212–228.

Sun, D., Jiang, F., Wu, H., Liu, S., Luo, P., and Zhao, Z. (2023). Root location and root diameter estimation of trees based on deep learning and ground-penetrating radar. *Agronomy 2023, Vol. 13, Page 344*, 13:344.

Terry, N., Day-Lewis, F. D., Lane, J. W., Johnson, C. D., and Werkema, D. (2023). Field evaluation of semi-automated moisture estimation from geophysics using machine learning. *Vadose Zone Journal*, 22:e20246.

Tran, A. P., Ardekani, M. R. M., and Lambot, S. (2012). Coupling of dielectric mixing models with full-wave ground-penetrating radar signal inversion for sandy-soil-moisture estimation. *Geophysics*, 77.

van't Veen, K. M., Ferré, T. P. A., Iversen, B. V., and Børgesen, C. D. (2022). Using machine learning to predict optimal electromagnetic induction instrument configurations for characterizing the shallow subsurface. *Hydrology and Earth System Sciences*, 26(1):55–70.

Varma, S. and Simon, R. M. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91 – 91.

Visconti, F. and de Paz, J. M. (2021). Sensitivity of soil electromagnetic induction measurements to salinity, water content, clay, organic matter and bulk density. *Precision Agriculture*, 22:1559–1577.

Zhou, S., Zhang, Y., Williams, A. P., and Gentine, P. (2019). Projected increases in intensity, frequency, and terrestrial carbon costs of compound drought and aridity events. *Science Advances*, 5.

Zhuo, L., Dai, Q., Han, D., Chen, N., Zhao, B., and Berti, M. (2019). Evaluation of remotely sensed soil moisture for landslide hazard assessment. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12:162–173.