




The effect of local samples in the accuracy of mid-infrared (MIR) and X-ray fluorescence (XRF) -based spectral prediction models

V. Vona¹ · S. Sarjant³ · B. Tomczyk³ · M. Vona² · R. Kalocsai¹ · I. M. Kulmány¹ · G. Jakab⁴ · A. Ver¹ · G. Milics⁵  · Cs. Centeri⁵

Accepted: 20 July 2022 / Published online: 20 August 2022
© The Author(s) 2022

Abstract

Within the soil spectroscopy community, there is an ongoing discussion addressing the comparison of the performance of prediction models built on a global calibration database, versus a local calibration database. In this study, this issue is addressed by spiking of global databases with local samples. The soil samples were analysed with MIR and XRF sensors. The samples were further measured using traditional wet chemistry methods to build the prediction models for seventeen major parameters. The prediction models applied by AgroCares, the company that assisted in this study, combine spectral information from MIR and XRF into a single ‘fused-spectrum’. The local dataset of 640 samples was split into 90% train and 10% test samples. To illustrate the benefits of using local calibration samples, three separate prediction models were built per element. For each model, 0%, 50% (randomly selected) and 100% of the local training samples were added to the global dataset. The remaining 10% local samples were used for validation. Seventeen soil parameters were selected to illustrate the differences in performance across a range of soil qualities, using the validation set to measure performance. The results showed that many models already exhibit an excellent level of performance ($R^2 \geq 0.95$) even without local samples. However, there was a clear trend that, as more local calibration samples were added, both R^2 and ratio of performance to interquartile distance (RPIQ) increase.

Keywords Soil testing · Prediction model · Mid-infrared spectroscopy · X-ray fluorescence · Spiking

Introduction

Hungarian agricultural soil fertility has been decreasing due to the imbalance between nutrient input and output. Simultaneously, soils have become more vulnerable due to the extreme weather conditions (e.g. drought sensitivity, erosion). Management of the steadily decreasing soil organic matter and lack of organic matter (manures, crop residues, etc.)

Extended author information available on the last page of the article

therefore becomes crucial in maintaining soil fertility in the middle term. To harmonize the preservation of soil fertility with farming objectives and environmental requirements in the 2021–2027 Common Agricultural Policy (CAP) of the European Union, there is a need for proper soil nutrient management strategies. These strategies should be based on information about the current status of soil fertility. Since traditional soil testing is time-consuming and expensive, there is a need for techniques and instruments that allow rapid, affordable and precise routine soil testing. As a result, general interest in using diffuse reflectance infrared spectroscopy testing of soil physical, chemical and biological properties has increased (Soriano-Disla et al., 2014).

This technique has various advantages such as (a) it is non-destructive, (b) requires relatively little sample preparation and (c) does not involve any (hazardous) chemicals. Measurements only take a few seconds and many soil properties can be estimated from a single scan. Moreover, the technique allows for flexible measurement configurations, and in-situ as well as laboratory-based measurements (Viscarra Rossel et al., 2006).

Mid-infrared (MIR [3–50 μm]) spectroscopy is a widely used tool to estimate particle size distribution, organic carbon content and other chemical and physical soil properties (Terra et al., 2015) based on the detection in the 4000 — 400 cm^{-1} wave number range. It typically performs better than near-infrared spectroscopy, detecting the fundamental vibrations instead of overtones (Seybold et al., 2019). Some soil properties, such as organic matter (aliphatics, aromatics, carbonyls, etc.) and mineral components (CaCO_3 ; clay minerals, etc.) are believed to connect to distinct wavenumbers. In contrast, others with no standard chemical composition (pH, texture, cation exchange capacity, aggregate stability) are estimated based on the whole spectrum (Ma et al., 2018). Prediction accuracy, however, is affected by potential overlaps of the peaks in the former case, and noise and multicollinearity in the latter.

X-ray fluorescence (XRF [0.01 to 10 nm]) spectroscopy is based on detecting the secondary spectral lines at various wavelengths emitted by chemical elements (Nawar et al., 2019). This is a robust method for predicting heavy metals. However, it is less effective for analyzing low atomic number elements, i.e., K, P, Ca and Mg (Kaniu et al., 2012). The combined application of MIR and XRF spectroscopy was first reported by Towett et al., (2015) and found to be promising due to their synergistic effect on predicting soil properties (Naimi et al., 2022; O'Rourke et al. 2016).

To make better use of spectroscopy, the establishment of soil spectral libraries covering different soil types has been strongly recommended (Viscarra Rossel, 2009). Most investigations have been focused on a single, quasi-homogeneous spectral library dataset. Nonetheless, these datasets are not necessarily comparable directly due to their various sample handling (e.g. drying, milling, sieving) and spectroscopic conditions. Extending the applicability of such libraries by spiking with local samples was introduced (Brown, 2007; Viscarra Rossel, 2009), which significantly improved prediction accuracy (Guero et al., 2016; Seidel et al., 2019) even at the field scale (Breure et al., 2022).

Eight years ago, AgroCares, a company utilising spectroscopic sensors to estimate soil and feed quality, started a research program focusing on the development and implementation of a soil testing method using mid-infrared (MIR) and X-ray fluorescence (XRF) sensor technology (Dimpka et al., 2017). One of the major challenges of such a concept is the derivation of reliable prediction models. For this, machine learning techniques were applied to link the wet chemical parameters of the soil calibration samples to spectra obtained from

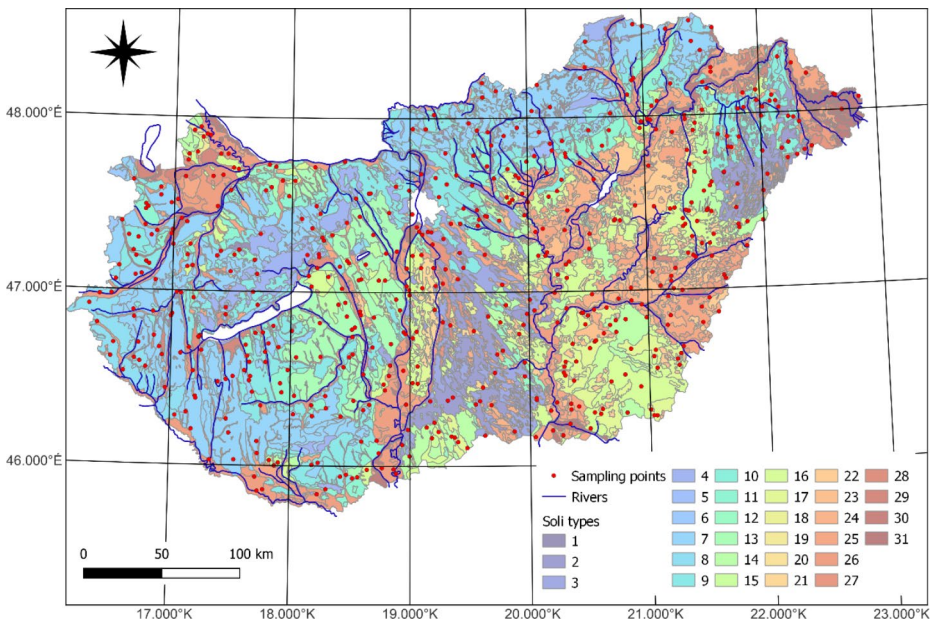


Fig. 1 Location of the local calibration and validation samples (Source of background soil map: Agrotopo (2008) database)

the MIR/XRF sensors. To make better use of spectroscopy, the establishment of soil spectral libraries covering different soil types has been strongly recommended (Viscarra Rosset, 2009). AgroCares utilises its global spectral calibration library of various soil types by building temporary local models to produce one-off predictions when prompted to produce a prediction. In order to develop prediction models for Hungary, a calibration and validation study started in 2017. In total 640 geo-referenced samples were collected to cover the different soil properties in Hungary. These samples are suitable for testing the effect of localised sample injection in a global model for Hungary. The aim of this study is to compare the effect of adding variable numbers of local calibration samples to a global soil calibration database on model performance within that local region.

Materials and methods

In Hungary, 640 geo-referenced soil calibration samples were taken. The locations were selected using the conditioned Latin Hypercube method (Minasny & McBratney, 2006; Roudier & Hedley, 2013). Variables used to stratify the samples were land use, soil type, climate data, accessibility and market value. In the selection of sample locations, it was considered if it is a high-priority agricultural area or not. The focus was on the most intensively used arable land (Fig. 1).

1 – Stony soils (solid rock on or near the surface), 2 – Blown sand, 3 – Humous sandy solis, 4 – Rendzinas, 5 – Erubase soils, 6 – Acidic, non-podzolic brown forest soils, 7 – Brown forest soils with clay illuviation (Sol brun lessivee), 8 – Pseudogleys, 9 – Brown earth (Ramann brown forest soils), 10 – „Kovárvány” brown forest soils (sandy brown

forest soils with thin interstratified layers of colloid and sesquioxide accumulation), 11 – Chernozem brown forest soils, 12 – Chernozem-type sandy soils, 13 – Pseudomycelial (calcareous) chernozems, 14 – Lowland chernozems, 15 – Lowland chernozems with salt accumulation in the deeper layers, 16 – Meadow chernozems with salt accumulation in the deeper layers, 17 – Meadow chernozems with salt accumulation in the deeper layers, 18 – Solonchak-solonetz, 19 – Terrace chernozems, 20 – Solonchaks, 21 – Solonchak-solonetz, 22 – Meadow solonetz, 23 – Meadow solonetz turning into steppe formation, 24 – Solonchak meadow soils, 25 – Meadow soils, 26 – Meadow alluvial soils and alluvial meadow soils, 27 – Peaty meadow soils, 28 – Peats, 29 – Ameliorated peats, 30 – Soils of swampy forests, 31 – Alluvial soils.

Soil samples were taken with an Edelman auger from the 0–200 mm top layer from one point. The top 20 mm of soil from the sample auger were removed, in order to remove any plant debris that might have fallen into the drill hole. 1 kg each of the soil samples were placed into 2 bags.

Soil samples were prepared according to ISO norm 11464:2006. The preparation included drying at 40 °C, and if necessary, the soil sample was crushed and sieved through a 2 mm sieve. The fraction less than 2 mm was divided into a portion mechanically using a divider. A subsample of 30 g was retrieved and milled with a ball mill to 0.2 mm particle size. Each milled sample was analysed with Alpha I Bruker (MIR) (Bruker Corporation, Billerica, USA) and Epsilon 1 (EDXRF) (Malvern Panalytical, Malvern, UK) sensors. The exact same instrumentation and procedures were used to create the AgroCares spectral library. This adherence to protocol helped eliminate potential discrepancies related to analytical procedure.

The samples were measured using traditional wet chemistry methods to define the reference dataset for the calibration and validation sets. Seventeen parameters were measured:

- pH in a 1:5 (volume fraction) suspension of soil in water (pH in H₂O), in 1 mol/l potassium chloride solution (pH in KCl) according to NEN-ISO 10,390, 2005,
- organic carbon (Organic C) by dry combustion, quantified with Elemental analyzer Rapid CS Cube (Elementar) according to EN 15,936, 2012,
- total nitrogen (total N) determined via Dumas method using elemental analyzer Vario MAX CN Cube (Elementar) according to NEN-EN 16168:2012,
- clay content (Clay) by using laser diffraction method by Mastersizer 3000 (Malvern Panalytical) according to ISO 13320:2009,
- total phosphorus (total P), calcium (total Ca), magnesium (total Mg) potassium (total K), measured with Epsilon 1, EDXRF (Malvern Panalytical) (according to ISO 18227:2014,
- exchangeable calcium (exchangeable Ca), magnesium (exchangeable Mg), potassium (exchangeable K) and cation exchange capacity (CEC) using a hexamminecobalt trichloride solution as extractant, quantified with ICP-MS 7700 (Agilent) according to ISO 23470:2007,
- Plant available phosphorus (plant available P), calcium (plant available Ca), magnesium (plant available Mg), potassium (plant available K), determined using Mehlich-3 extraction, quantified with ICP-MS 7700 (Agilent) according to Wolf and Beegle (2009).

The AgroCares workflow creates prediction models that combine the spectral information from MIR and XRF into a single ‘fused-spectrum’. This means that spectral information

present in either sensor are simultaneously utilised for a better overall prediction than either sensor individually (Elmenreich, 2002). This can be seen in the [results](#) section in Table 1.

To build a prediction model, spectral data from both sensors is required, as well as the wet-chemical reference data to be predicted by the model. In 2021, the models of AgroCares were built on approximately 17,000 soil samples, collected from 35 countries (primarily in Africa, Asia and Europe). The prediction models were built using the WEKA software (Hall et al., 2009) and used the ADAMS knowledge flow (Reutemann & Vanschoren, 2012) to prepare the data. Before training the primary model, the data was first cleaned of outliers by training a simple linear regression algorithm with 10-fold cross validation. All samples that were misclassified by this cleaning algorithm by a factor of m ($m=6$ in experiments) were removed. The cleaned samples were then used to build each prediction model, one model per soil property. Each model was built by first applying a Segmented Savitzky-Golay (SSG) filter (Geise and French, 1955) with $n=4$ segments for both MIR and XRF spectra, then partial least squares (PLS) regression (Vinzi et al., 2010) to decompose the spectral information into approximately 10–20 components per sensor input (this varies per model). The PLS data from both MIR and XRF was then concatenated and used as input for a locally weighted learning (LWL) algorithm (number of neighbours, $k=300$) (Atkeson et al., 1997) wrapped around a Gaussian Processes (GP) regressor (Rasmussen & Williams, 2006). The parameters for each of these processes (e.g. number of PLS components, window size of Savitzky-Golay, etc.) were defined through an optimisation process for each soil property using the full global dataset. The LWL component is the crucial component here – despite using a global dataset, the LWL builds a temporary GP regressor on k of the most similar training samples (based on the PLS components) at prediction time, and uses this model to predict the soil property value.

Other studies have also utilised a fusion of MIR and XRF sensors to show improved combined performance. O'Rourke et al. (2016) used Cubist regression models (Holmes et al., 1999; Quinlan, 1993) to predict soil parameters for the NSDB Republic of Ireland Database (Fay et al., 2007). This technique also showed improvements when using fused spectral inputs, but the scope of data is much smaller than this study. Kandpal et al., (2022) also presented favourable results for fused sensor approaches, using different PLS variants as the primary prediction method (and various preprocessing techniques). But this method also used a relatively small number of samples ($n=196$). The key difference in the work in this study is the use of the LWL algorithm to dynamically create localised prediction models for every new sample. This technique takes advantage of a large soil library to build focused models on the soil sample being predicted.

The local dataset of 640 Hungarian soil samples was split into 90% training and 10% test samples. To illustrate the benefits of using local calibration samples, three separate prediction models were built per soil property. For each model, 0%, 50% (randomly selected) or 100% of the local Hungarian training samples (90%) were added to the global calibration dataset of approximately 17,000 samples (where only 300 are selected per sample, as defined above). The remaining 10% local samples were used for validation. Seventeen soil parameters were selected to illustrate the differences in performance across a range of soil qualities, using the validation set to measure performance. Two evaluation metrics are presented: R^2 , the coefficient of determination – a measure of how closely the predictions match the actual values; and ratio of performance to interquartile distance (RPIQ), a metric for representing the root mean squared error as a factor of the range (Bellon-Maurel

et al., 2010). Because the validation samples are constant, an increase in RPIQ represents a decrease in error. To interpret and to compare the results of the model performance in Hungary, the predictions based on R^2 values were classified according to Malley et al., (2004), as follows: excellent ($R^2 > 0.95$), successful ($0.90 \leq R^2 \leq 0.95$), moderately successful ($0.80 \leq R^2 < 0.90$) and moderately useful ($0.70 \leq R^2 < 0.80$).

Results

The experiment was conducted for: pH(KCl), pH(H₂O), organic C, total N, clay content, total P, total Ca, total Mg, total K, exchangeable Ca, exchangeable Mg, exchangeable K, CEC, plant available P, plant available Ca, plant available Mg and plant available K.

Table 1 shows the results of the experiments incorporating varying levels of local samples in the calibration set. Many models already exhibit an excellent level of performance ($R^2 \geq 0.95$), even without local samples. Overall, there is a clear trend that as more local calibration samples were added, both R^2 and RPIQ increase. For elements with already excellent performance, additional samples had little effect, but the effect is more pronounced for lower performing models. Notable examples include: pH(KCl), clay, exchangeable Ca, available Ca and Mg and total K – each showing a substantial reduction in RMSE and an increase in R^2 . Some models show no increase in performance: CEC, total Mg and available P; local samples have no effect on performance. The worst prediction was for plant available K, but adding calibration samples resulted in an R^2 improvement from 0.13 to 0.30 and a 10% increase in RPIQ.

Figure 2 illustrates the differences in performance of the prediction models on the Hungarian validation set when different amounts of Hungarian calibration data is used to train the prediction models. Those parameters were selected for visualization where the addition of local samples has a larger effect on both R^2 and RPIQ.

In models where R^2 performance is already high (> 0.95), additional Hungarian samples increase the RPIQ. In weaker models, such as available Mg and available Ca, the addition of local samples has a larger effect on both R^2 and RPIQ except for CEC and available P.

Clay showed an overall reduction in error across the entire range. Performance for plant available Ca is excellent from 0 to 10,000 in both models, but when local calibration samples are provided, the error from 10,000+ is notably reduced (by 11.08%). An even bigger effect is seen with plant available Mg. While the performance of exchangeable K is poor, the addition of local samples aids in centralising the predictions. The model for plant available K is very poor, but the trend clearly increases with the addition of local samples. The effect of the local sample was most visible on available Mg where the R^2 improved by 20% and the RPIQ increased by 78%.

Figure 3 illustrates the performance improvement of using ‘fused’ models for prediction. In all but two of the models: total Ca and plant available K, fused performance matches or exceeds that of the individual sensor models. In the total Ca model, XRF alone far exceeds the performance of the fused model, but performance was already at a very high level for all two sensors. For plant available K, performance was quite low for all three models and the increase seen for XRF alone is likely due to noise in the training process.

Table 1 A summary of the performance of the prediction models over the Hungarian validation set of 10% samples

Soil parameter	Evaluation metric	Hungarian calibration samples				
		MIR + XRF			MIR	XRF
		0%	50%	100%	100%	100%
pH(KCl)	RMSE	0.26	0.27	0.23	0.31	0.47
	R²	0.96	0.96	0.97	0.94	0.87
	RPIQ	7.47	7.24	8.54	6.24	4.08
pH(H₂O)	RMSE	0.30	0.29	0.29	0.32	0.45
	R²	0.90	0.91	0.91	0.88	0.78
	RPIQ	4.64	4.83	4.80	4.31	3.10
Organic C	RMSE	5.37	5.24	5.13	7.82	5.33
	R²	0.98	0.98	0.98	0.95	0.98
	RPIQ	3.19	3.27	3.34	2.19	3.22
Clay	RMSE	4.25	3.76	3.72	4.23	6.26
	R²	0.86	0.89	0.90	0.87	0.71
	RPIQ	3.91	4.41	4.47	3.92	2.65
CEC	RMSE	92.06	92.24	91.74	90.06	108.17
	R²	0.72	0.72	0.73	0.74	0.62
	RPIQ	2.10	2.09	2.11	2.15	1.79
Total N	RMSE	0.73	0.72	0.71	0.88	0.73
	R²	0.95	0.95	0.95	0.92	0.95
	RPIQ	2.35	2.36	2.39	1.93	2.34
Total P	RMSE	0.12	0.11	0.11	0.58	0.11
	R²	0.96	0.97	0.97	0.04	0.96
	RPIQ	3.72	4.06	4.19	0.76	3.91
Plant available P	RMSE	119.49	117.19	120.19	271.18	117.97
	R²	0.82	0.83	0.82	0.08	0.83
	RPIQ	0.91	0.93	0.91	0.40	0.93
Total Ca	RMSE	7.18	7.00	6.49	7.31	3.62
	R²	0.98	0.98	0.98	0.98	0.99
	RPIQ	6.02	6.17	6.66	5.91	11.95
Exchangeable Ca	RMSE	58.13	54.10	51.69	52.96	79.31
	R²	0.90	0.92	0.92	0.92	0.82
	RPIQ	3.31	3.56	3.72	3.63	2.43
Plant available Ca	RMSE	3653.35	3251.64	2764.41	3447.19	5077.21
	R²	0.88	0.90	0.93	0.89	0.77
	RPIQ	2.63	2.96	3.48	2.79	1.89
Total Mg	RMSE	2.17	2.09	2.23	2.23	2.30
	R²	0.91	0.92	0.90	0.90	0.90
	RPIQ	1.53	1.59	1.49	1.49	1.44
Exchangeable Mg	RMSE	18.72	17.29	17.13	24.30	34.77
	R²	0.85	0.87	0.87	0.74	0.47
	RPIQ	3.19	3.46	3.49	2.46	1.72
Plant available Mg	RMSE	414.04	322.03	232.37	341.63	503.75
	R²	0.77	0.86	0.93	0.84	0.66
	RPIQ	1.50	1.93	2.68	1.82	1.23
Total K	RMSE	1.26	1.15	1.03	1.76	1.20
	R²	0.94	0.95	0.96	0.89	0.95
	RPIQ	4.77	5.22	5.83	3.43	5.01

Table 1 (continued)

Exchangeable K	RMSE	7.08	6.79	6.83	9.65	8.86
	R²	0.60	0.63	0.63	0.25	0.37
	RPIQ	1.49	1.56	1.55	1.10	1.19
Plant available K	RMSE	405.69	381.51	364.04	348.37	250.32
	R²	0.13	0.23	0.30	0.36	0.67
	RPIQ	0.76	0.81	0.84	0.88	1.23

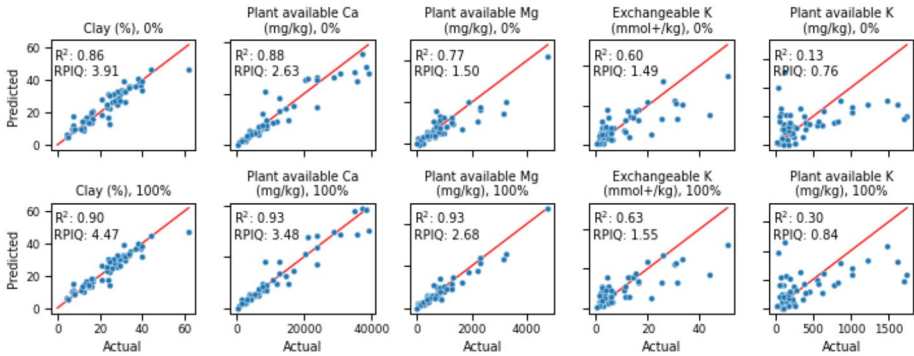
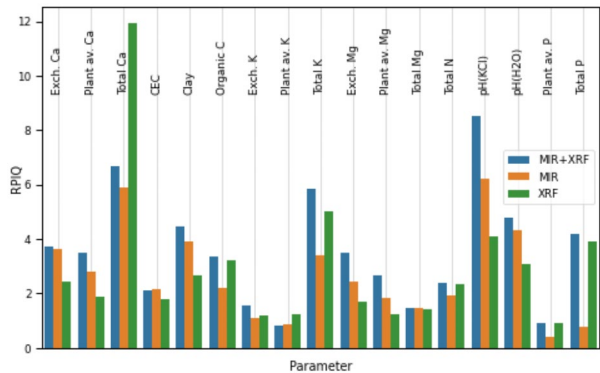


Fig. 2 Plots illustrating the performance of the Hungarian validation samples for clay, plant available Ca, plant available Mg, exchangeable K and plant available K. 0% added Hungarian calibration samples on the top row, 100% added Hungarian calibration samples on the bottom row

Fig. 3 A comparative RPIQ plot of MIR+XRF, MIR and XRF performance when utilising 100% of the local samples in the training set



Discussion

The paper aimed to compare the performance of a global dataset to that of a global dataset including local samples. Large, country, continental and global soil spectral libraries have already been developed and found to be useful for larger scale analyses and interpretation (Viscarra Rossel & Webster, 2012; Stevens et al., 2013; Shi et al., 2014; Viscarra Rossel et al., 2016). Results have shown that large libraries are not always accurate to predict soil

parameters of local samples (Stevens et al., 2013; Gogé et al., 2014; Guerrero et al., 2014; Clairotte et al., 2016; Brown, 2007). It was found that the addition of local samples from Hungary notably improved the performance for pH(KCl), clay, exchangeable Ca, available Ca and Mg and total K. The improvement was mainly expressed as a reduction in RMSE values and only moderately improved R^2 values, though some models showed no increase in performance. In the current study, local samples did not affect the performance of CEC, total Mg and available P. It seems that the global dataset of AgroCares already contains similar information on comparable soils from other countries in the prediction of these parameters and so no further improvements could be gained from local samples. The prediction of clay fraction was moderately successful without the local samples and became successful with local samples, but 100% of the local training dataset was needed for this improvement.

The predictions for plant available K was very poor, but could be seen to improve with the successive addition of local samples, which may indicate that the model performance would be further increased with extremely dense sampling. The positive effect of local samples on model performance could be explained with the statement of Viscarra Rossel et al., (2008): NIR spectra and soil properties can vary under different soil mineralogy and their content in soil organic matter. According to Fabien et al., (2012), when strong spectral features are related to the characteristic under study (as for CaCO_3 content), a wide national database can be used alone to calibrate accurate prediction models. In the other cases, for properties involving more diverse spectral regions, the usefulness of a large database spiked with local samples should be established. It confirms the study results of Soriano-Disla et al., (2014): ‘variables that are predicted by virtue of their correlations with infrared-active soil properties (indirect calibrations) frequently require the development of models for specific soil types, locations and particular environments’.

In the case of CEC, the local dataset involvement did not increase RPIQ, possibly because the non-Hungarian soil samples in the database are already good enough. Most studies aim to predict soil properties (including CEC) on a catchment or a regional scale. For example, Viscarra Rossel et al., (2006) reported a good R^2 (0.73) for CEC prediction in Australian soils. Terra et al., (2015) and Pinheiro et al., (2017) found $R^2=0.72$ and 0.68 and $\text{RMSE}=0.14$ and 5.86 for CEC prediction in south American tropical soils. Also, Ulusoy et al., (2016) published R^2 of 0.83 and RMSE of 1.45 at field scale investigations in Turkey. However, the prediction goodness might drop on a global scale, including several soil types with various environmental circumstances.

A notable difference between the correlation of R^2 and RPIQ can be seen with the plant available P content. The drastic difference between the two measures is likely because the validation set contains a few large values which artificially increase the R^2 . These large values would fall outside the inter-quantile range, but their errors are still factored into the RPIQ measurement, thereby resulting in a relatively small RPIQ. This is a weakness of the RPIQ metric.

Conclusions

In 2021, AgroCares had more than 17,000 soil samples in their global calibration dataset collected from 32 countries in Africa, Asia and Europe. The results of this study show that the global dataset of AgroCares already contains enough information on comparable

soils from other countries to predict the total calcium, total nitrogen and total phosphorus successfully, regardless of the number of Hungarian samples present. However, in weaker models, such as clay content, exchangeable potassium and all magnesium forms with R^2 lower than 0.95, the addition of local Hungarian samples has improved the quality of local predictions (except for total Mg where it decreased).

In the case of special soil parameters, it can be important to identify the local effects that are unique to a country or a region. Based on the experiment, it can be concluded that the model performance in Hungary benefits from the inclusion of local samples. Based on the cases where model performance was improved with the increasing number of local samples, it can be concluded that more data is useful. However, it is important to state that sample numbers did not always improve model performance and/or even decrease it. According to the results, the AgroCares concept for routine soil testing using MIR + XRF sensor technology is viable for Hungarian agriculture.

Funding Open access funding provided by Hungarian University of Agriculture and Life Sciences.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Agrotopo database (2008). : <https://www.elkh-taki.hu/hu/keptar/agrotopo> [last accessed: 07.07.2022]
- Atkeson, C. G., Moore, A. W., & Schaal, S. (1997). Locally weighted learning. *Lazy learning*. Dordrecht, The Netherlands: Springer. <https://doi.org/10.1007/978-94-017-2053-3-2>
- Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends in Analytical Chemistry*, 29(9), 1073–1081. <https://doi.org/10.1016/j.trac.2010.05.006>
- Breure, T. S., Prout, J. M., Haefele, S. M., Milne, A. E., Hannam, J. A., Moreno-Rojas, S., et al. (2022). Comparing the effect of different sample conditions and spectral libraries on the prediction accuracy of soil properties from near- and mid-infrared spectra at the field-scale. *Soil and Tillage Research*, 215, 105196. <https://doi.org/10.1016/j.still.2021.105196>
- Brown, D. J. (2007). Using a global VNIR soil-spectral library for local soil characterization and landscape modelling in a 2nd-order Uganda watershed. *Geoderma*, 140, 444–453. <https://doi.org/10.1016/j.geoderma.2007.04.021>
- Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N. P. A., Bernoux, M., et al. (2016). National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. *Geoderma*, 276, 41–52. <https://doi.org/10.1016/j.geoderma.2016.04.021>
- Elmenreich, W. (2002). Sensor Fusion in Time-Triggered Systems. PhD thesis, TU Wien, Inst. für Technische Informatik, Vienna, Austria
- Fabien, A. C., Wilkins, D. R., Miller, J. M., Reis, R. C., Reynolds, C. S., Cackett, E. M., et al. (2012). On the determination of the spin of the black hole in Cyg X-1 from X-ray reflection spectra. *Monthly Notices of the Royal Astronomical Society*, 424 (1) 217–223. <https://doi.org/10.1111/j.1365-2966.2012.21185.X>
- Fay, D., Kramers, G., Zhang, C., McGrath, D., & Grennan, E. (2007). *Soil geochemical atlas of Ireland*. Ireland: Teagasc and the Environmental Protection Agency
- Gogé, F., Gomez, C., Jolivet, C., & Joffre, R. (2014). Which strategy is best to predict soil properties of a local site from a national vis-NIR database? *Geoderma*, 213, 1–9. <https://doi.org/10.1016/j.geoderma.2013.07.016>

- Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R. A., Maestre, F. T., Mouazen, A. M., et al. (2014). Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset. *European Journal of Soil Science*, 65, 248–263. <https://doi.org/10.1111/ejss.12129>
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 10–18. <https://doi.org/10.1145/1656274.1656278>
- Holmes, G., Hall, M., & Frank, E. (1999). Generating rule sets from model trees. *Advanced Topics in XXXrtificial Intelligence. Lecture Notes in Computer Science*, 1747, 1–12. https://doi.org/10.1007/3-540-46695-9_1
- ISO. ISO 13320-1:2009 (2009). *Particle size analysis – laser diffraction methods – Part 1: General principles*.
- ISO 18227:2014 (2014). *Soil Quality Determination of elemental composition by X-ray fluorescence*.
- ISO 23470:2007 (2007). *Soil quality Determination of effective cation exchange capacity22 (CEC) and exchangeable cations using a hexamminecobalt trichloride solution*.
- Kandpal, L., Munnaf, M., Cruz, C., & Mouazen, A. (2022). Spectra Fusion of Mid-Infrared (MIR) and X-ray Fluorescence (XRF) Spectroscopy for Estimation of Selected Soil Fertility Attributes. *Sensors (Basel)*, 22(9), 3459. <https://doi.org/10.3390/s22093459>
- Kaniu, M. I., Angayo, K. H., Mwala, A. K., & Mangala, M. J. (2012). Direct rapid analysis of trace bioavailable soil macronutrients by chemometrics-assisted energy dispersive X-ray fluorescence and scattering spectrometry. *Analytica Chimica Acta*, 729, 21–25. <https://doi.org/10.1016/j.aca.2012.04.007>
- Ma, F., Du, C. W., Zhou, J. M., & Shen, Y. Z. (2018). Investigation of soil properties using different techniques of mid-infrared spectroscopy. *European Journal of Soil Science*, 70, 96–106. <https://doi.org/10.1111/ejss.12741>
- Malley, D. F., Martin, P. D., & Ben-Dor, E. (2004). In C. A. Roberts, J. Workman, Jr. and, & J. B. Reeves III (Eds.), *Application in analysis of soils. In Near Infrared Spectroscopy in Agriculture, Agronomy 44* (pp. 729–784). Madison, WI, USA: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America. <https://doi.org/10.2134/agronomogr44.c26>
- Minasny, B., & McBratney, A. (2006). A Conditioned Latin Hypercube Method for Sampling in the Presence of Ancillary Information. *Computers & Geosciences*, 32, 1378–1388. <https://doi.org/10.1016/j.cageo.2005.12.009>
- Naimi, S., Ayoubi, S., Di Loreto, L. A., & Melo Dematte, J. A. (2022). Quantification of some intrinsic soil properties using proximal sensing in arid lands: Application of Vis-NIR, MIR and pXRF spectroscopy. *Geoderma Regional*, 28, e00484. <https://doi.org/10.1016/j.geodrs.2022.e00484>
- Nawar, N., Delbecq, Y., Declercq, P., De Smedt, P., Finke, A., Verdoodt, M., et al. (2019). Can spectral analyses improve measurement of key soil fertility parameters with X-ray fluorescence spectrometry? *Geoderma*, 350, pp. 29–39
- NEN-ISO 10390 (2005). *Soil quality -Determination of pH*
- NEN-EN 15936 (2012 Sludge). : treated bio waste, soil and waste –Determination of total organic carbon (TOC) by dry combustion
- NEN-EN 16168, *Sludge, treated biowaste and soil -Determination of total nitrogen using dry combustion method, September -Soil samples*
- Pinheiro, É. F., Ceddia, M. B., Clingensmith, C. M., Grunwald, S., & Vasques, G. M. (2017). Prediction of soil physical and chemical properties by visible and near-infrared diffuse reflectance spectroscopy in the central Amazon. *Remote Sensing*, 9(4), 293
- Quinlan, J. R. (1993). Combining instance-based and model-based learning. In: Proceedings of the 10th International Conference on Machine Learning, San Mateo, California. *Conference Proceedings* p. 236–243. San Mateo, CA, USA: The International Machine Learning Society. <https://doi.org/10.5555/3091529.3091560>
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*, Cambridge, MA, USA: The MIT Press, Massachusetts Institute of Technology, ISBN 026218253X. <https://doi.org/10.7551/mitpress/3206.001.0001>
- Reutemann, P., & Vanschoren, J. (2012). Scientific Workflow Management with ADAMS. In Proceedings of the Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), Part II, LNCS 7524, 833–837. https://doi.org/10.1007/978-3-642-33486-3_58
- Roudier, P., & Hedley, C. B. (2013). Smart sampling to assist on-farm nutrient management. In: *Accurate and efficient use of nutrients on farms* (Eds L.D. Currie -C L. Christensen). Occasional Report No. 26. Fertilizer and Lime Research Centre, Massey University, Palmerston North, NZ
- Seidel, M., Hutengs, C., Ludwig, B., Thiele-Bruhn, S., & Vohland, M. (2019). Strategies for the efficient estimation of soil organic carbon at the field scale with vis-NIR spectroscopy: spectral libraries and spiking vs. local calibrations. *Geoderma 354 Article*, 113856, <https://doi.org/10.1016/j.geoderma.2019.07.014>

- Seybold, C. A., Ferguson, R., Wysocki, D., Bailey, S., Anderson, J., Nester, B., et al. (2019). Application of Mid-Infrared Spectroscopy in Soil Survey. *Soil Science Society of America Journal*, 83, 1746–1759. <https://doi.org/10.2136/sssaj2019.06.0205>
- Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., et al. (2014). Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Science China Earth Sciences*, 57, 1671–1680. <https://doi.org/10.1007/s11430-013-4808-x>
- Soriano-Disla, J. M., Janik, L. J., Rossel, V., Macdonald, R. A., L. M. and, & McLaughlin, M. J. (2014). The Performance of Visible, Near- and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical and Biological Properties. *Applied Spectroscopy Reviews*, 49, 139–186. <https://doi.org/10.1080/05704928.2013.811081>
- Stevens, A., Nocita, M., Tóth, G., Montanarella, L., & van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PLoS One*, 8, e66409. <https://doi.org/10.1371/journal.pone.0066409>
- Terra, F. S., Demattê, J. A. M., & Viscarra-Rossel, R. A. (2015). Spectral libraries for quantitative analysis of tropical Brazilian soils: Comparing VIS-NIR and MIR reflectance data. *Geoderma*, 255–256, 81–93. <https://doi.org/10.1016/j.geoderma.2015.04.017>
- Towett, E. K., Shepherd, K. D., Sila, A., Aynekulu, E., & Cadisch, G. (2015). Mid-Infrared and Total X-Ray Fluorescence Spectroscopy Complementarity for Assessment of Soil Properties. *Soil Science Society of America Journal*, 79, 1375–1385. <https://doi.org/10.2136/sssaj2014.11.0458>
- Ulusoy, Y., Tekin, Y., Tümsavaş, Z., & Mouazen, A. M. (2016). Prediction of soil cation exchange capacity using visible and near infrared spectroscopy. *Biosystems Engineering*, 152, 79–93. <https://doi.org/10.1016/j.biosystemseng.2016.03.005>
- Vinzi, V. E., Trinchera, L., & Amato, S. (2010). PLS path modelling: From foundations to recent developments and open issues for model assessment and improvement. In E. Vinzi, V. Chin, W. W. Henseler, & J. Wang, H. (Eds.), *Handbook of Partial Least Squares: Concepts, Methods and Application* (pp. 47–82). Berlin Heidelberg Germany: Springer
- Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, 131, 59–75. <https://doi.org/10.1016/j.geoderma.2005.03.007>
- Viscarra Rossel, R. A., Jeon, Y. S., Odeh, I. O. A., & McBratney, A. B. (2008). Using a legacy soil sample to develop a mid-IR spectral library. *Australian Journal of Soil Research*, 46, 1–16. <https://doi.org/10.1071/SR07099>
- Viscarra Rossel, R. A. (2009). The Soil Spectroscopy Group and the development of a global soil spectral library NIR News, 20 (4), 14–15. <https://doi.org/10.1255/nirn.1131>
- Viscarra Rossel, R. A., & Webster, R. (2012). Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. *European Journal of Soil Science*, 63, 848–860. <https://doi.org/10.1111/j.1365-2389.2012.01495.x>
- Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., et al. (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews*, 155, 198–230. <https://doi.org/10.1016/j.earscirev.2016.01.012>
- Wolf, A., & Beegle, D. (Eds.). (2009). Recommended soil tests for macro and micronutrients. In: Northeastern Regional Publication: *Recommended soil testing procedures for the Northeastern United States*. 3rd Edition. Northeastern Regional Publication No. 493. Cooperative Extension, University of Delaware, Newark, USA. pp 39–48

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11119-022-09942-y>.

Authors and Affiliations

V. Vona¹ · S. Sarjant³ · B. Tomczyk³ · M. Vona² · R. Kalocsai¹ · I. M. Kulmány¹ · G. Jakab⁴ · A. Ver¹ · G. Milics⁵ · Cs. Centeri⁵

✉ G. Milics

milics.gabor@uni-mate.hu

- ¹ Faculty of Agricultural and Food Sciences, Széchenyi István University, Mosonmagyaróvár, Hungary
- ² Csernozjom Ltd, Nagykörű, Hungary
- ³ AgroCares, Wageningen, The Netherlands
- ⁴ Research Network. Research Centre for Astronomy and Earth Sciences, Eötvös Loránd, Geographical Institute, Budapest, Hungary
- ⁵ Hungarian University of Agriculture and Life Sciences, Gödöllő, Hungary