Check for updates

Predicting plant available phosphorus using infrared spectroscopy with consideration for future mobile sensing applications in precision farming

Stefan Pätzold, et al. *[full author details at the end of the article]*

Abstract

Phosphorus (P) fertilisation recommendations rely primarily on soil content of plant available P (P_{avl}) that vary spatially within farm fields. Spatially optimized P fertilisation for precision farming requires reliable, rapid and non-invasive P_{avl} determination. This laboratory study aimed to test and to compare visible-near infrared (Vis–NIR) and mid-infrared (MIR) spectroscopy for P_{avl} prediction with emphasis on future application in precision agriculture. After calibration with the conventional calcium acetate lactate (CAL) extraction method, limitations of Vis–NIRS and MIRS to predict P_{avl} were evaluated in loess topsoil samples from different fields at six localities. Overall calibration with 477 (Vis–NIRS) and 586 (MIRS) samples yielded satisfactory model performance (R^2 0.70 and 0.72; RPD 1.8 and 1.9, respectively). Local Vis–NIRS models yielded better results with R^2 up to 0.93 and RPD up to 3.8. For MIRS, results were comparable. However, an overall model to predict P_{avl} on independent test data partly failed. Sampling date, pre-crop harvest residues and fertilising regime affected model transferability. Varying transferability could partly be explained after deriving the cellulose absorption index from the Vis–NIR spectra. In 62 (Vis–NIRS) and 67% (MIRS) of all samples, prediction matched the correct P_{avl} content class. Rapid discrimination between high, optimal and low P classes could be carried out on many samples from single fields thus marking an improvement over the common practice. However, P_{avl} determination by means of IR spectroscopy is not yet satisfactory for determination of precision fertilizer dosage. For introduction into agricultural practice, a standardized sampling protocol is recommended to help achieve reliable spectroscopic P_{avl} prediction.

Keywords Fertiliser recommendation · Calcium acetate lactate extraction · Vis–NIRS · MIRS · Soil · Calibration transferability · Phosphate prediction · Cellulose absorption index · Proximal soil sensing · PLSR

Stefan Pätzold and Matthias Leenen have contributed equally to the manuscript.

Introduction

Optimised phosphorus (P) fertilisation is desired to enhance farm input efficiencies, improve economic returns and minimize ecologic disturbances. In this respect, reducing P inputs is an actual aim in view of the high P status of many arable soils, at least in Western Europe (Csatho et al. 2007). Nonetheless, optimisation might also result in increased P dosage to exploit soil yield potential. Plant availability of P is a complex outcome of P fertilisation history, but also of multiple and partly interacting soil properties. However, P_{avl} makes up only a small percentage of total soil P. Diverse extractants are approved to estimate P_{avl} and to calculate fertiliser dosage. Jordan-Meille et al. (2012) reviewed European soil P tests and fertilising recommendations. The principle is similar everywhere: P extraction with chemical reagents is followed by calibration of P_{avl} values on the basis of field trials in order to achieve fertiliser recommendations. In Germany, the calcium acetate lactate (CAL) extraction (Schüller 1969) is a standard soil test method to determine P_{avl} and forms the basis of P fertilisation recommendations (LWK 2015). Note that P_{avl} as determined by conventional soil tests is operationally defined and does not account for chemical P binding forms (Hartmann et al. 2019).

Available nutrients and other soil properties that directly or indirectly impact crop growth may vary considerably within fields (Patzold et al. 2008). Precision agriculture addresses this problem by adapting fertilisation in different field-zones (Gebbers and Adamchuk 2010); such optimisation can lead to economic benefits in the long term (Schulte-Ostermann and Wagner 2018). However, common agricultural practice is to take composite soil samples that are regarded as representative for the sampled area. In Germany, official advisory services recommend one composite soil sample per 3 ha area (LWK 2015). In most cases, P fertiliser is still uniformly applied because farmers feel uncertain about how to divide a distinct field into management zones. Zones with varying soil properties can be distinguished when conducting a soil survey with diverse sensing techniques. Sensor maps often unravel soil heterogeneity and explain variation in, e.g., grain yield or total biomass (Mertens et al. 2008; Sun et al. 2011). Such information is commercially provided by consultants and helps to define management or soil sampling zones. However, routine soil sensing such as electromagnetic induction most often yields information that is not directly related to available decision rules (Sylvester-Bradley et al. 1999); this is especially the case for phosphorus.

For physical reasons, valuable information about P_{avl} in soil is not directly readable from sensor signals including electromagnetic induction, gamma spectrometry and optical instruments. However, in several studies, infrared (IR) spectroscopy in the visible and near-infrared (Vis–NIRS) and in the mid-infrared (MIRS) have performed satisfactorily in elucidating relationships between soil spectra and nutrient status as reviewed in Kuang et al. (2012). Spectroscopic models require calibration with conventionally analysed soil samples before predictions for the parameter of interest in unknown samples can be done. For model calibration, multivariate statistics such as partial least squares regression (PLSR) are in widespread use. Even if the soil property of interest has no specific spectral feature, PLSR can perform well for model calibration, when it finds surrogate spectral features that retain satisfactory estimations of the studied soil properties (Gomez et al. 2008).

Reviews by Kuang et al. (2012) and Soriano-Disla et al. (2014) have pointed out that no precise, but only approximate, quantitative prediction of P_{avl} is possible using IR spectroscopy. The reason is that the low dipole moment between P and oxygen (O) inhibits direct detection of orthophosphate. However, when organically bound P forms

a tight relationship with other relevant soil properties, P quantification is generally possible, but requires further testing in soils with a wide range of properties (Abdi et al. 2016). In general, MIRS yields better results than Vis–NIRS where orthophosphate has no spectral response (Soriano-Disla et al. 2014; Kruse et al. 2015). In chemically pure systems, Ahmed et al. (2019) elucidated orthophosphate and molecular mechanisms for phosphate adsorption to goethite using MIRS. Mayrink et al. (2019) successfully predicted exchange resin-adsorbed P, but also without direct spectral influence of soil components. However, P binding mechanisms and partners in soil are manifold, and P_{avl} is not a chemically defined component. Consequently, Wijewardane et al. (2018) found poor P_{avl} (Mehlich) prediction in the MIR when analysing 4969 samples from a broad range of soils all over the USA.

The P_{avl} proportion of total P is partly controlled by organic matter (OM) content and properties that determine IR reflectance of soil. Consequently, IR spectra may contain indirect, i.e., surrogate, information on P fractions (Chang et al. 2001; Abdi et al. 2012; Soriano-Disla et al. 2014). Abdi et al. (2016) successfully predicted organic P in chernozem soil samples because of the close relationship between OM and organic P. Organically bound P (e.g., phytate) is not, or only to a minor degree, captured by conventional soil P tests (Steffens et al. 2010). Abdi et al. (2012, 2016) reported variable success in P_{avl} prediction depending on the extractant. The CAL method was designed to characterize P_{avl} (Schüller 1969) and extracts P from diverse binding partners in soil (Steffens et al. 2010; Hartmann et al. 2019). Until now, Vis–NIRS and MIRS have hardly been tested to predict CAL-extractable P (CAL-P).

Most published studies to predict P_{avl} with emphasis on precision agriculture have been conducted on very few or even single fields, and models were mostly not transferred to fields outside calibration. Mouazen et al. (2007) and Mouazen and Kuang (2016) predicted P_{avl} in the lab and field with accuracy that was regarded sufficient for practical purposes. Mayrink et al. (2019) successfully predicted Mehlich-3 extractable P by taking diffuse reflectance spectra (325–1075 nm) from resin exchange strips that had been used to extract P from soil. All soil samples came from a 20 ha field; prior to model building, samples were pre-selected to optimise results.

Application of Vis–NIRS directly in the field is generally possible, but taking Vis–NIR spectra in the laboratory still performs generally better, because measurements can be taken under controlled conditions (Rodionov et al. 2014a, b, 2016). Mouazen and Kuang (2016) published a case study on P_{avl} prediction with mobile Vis–NIRS. For the mid-infrared, portable spectrometers (pMIRS) are available and are increasingly tested in soil science (Soriano-Disla et al. 2017). For example, the recent study by Rogovska et al. (2019) tested the potential for nitrate estimation in precision agriculture. However, prediction of P_{avl} in situ using pMIRS failed (Ji et al. 2016).

Major challenges concerning P_{avl} prediction by IR spectroscopy are (i) overcoming the case study character of P_{avl} prediction both in the Vis–NIR and MIR region, (ii) transferring models from calibration sites to independent fields, and (iii) calibrating P_{avl} prediction models that are universally applicable in practical precision farming. To address these challenges, the ongoing joint research project "BonaRes-I4S" is developing a mobile sensor platform that integrates non-invasive sensors and soil sampling for chemical and spectroscopic analyses (Gebbers et al. 2016). The objective of this study was to test and to compare Vis–NIR and MIR models capable of P_{avl} prediction in samples from different fields. The study was conducted in a laboratory setting as needed to evaluate the potential for integrating a pMIRS-based P_{avl} prediction on the future "BonaRes-I4S" sensor platform.

Materials and methods

Sample set

This study was conducted on archived sets of arable topsoil samples from different research sites. All samples had been analysed in different projects with the CAL method to characterise P_{avl} and were re-evaluated for this study. Six different localities with soils from loess or loess-derived sediments were sampled (Table 1). Klein-Altendorf (KA; 50°36′57″ N, 6°59′37″ E; four fields), Poppelsdorf (PO; 50°43′34″ N, 7°05′09″ E), Dikopshof (DI; 50°48′29″ N, 6°57′11″ E), and Hilberath (HI; 50°34′46″ N, 6°59′27″ E) were situated close to Bonn. The field at Merklingsen (ME; 51°34′07″ N; 8°00′05″ E) was located 110 km NE of Bonn. The Görzig (GÖ; 51°40′08″ N; 12°00′52″ E) field was located 360 km ENE of Bonn and 20 km N of Halle. The topsoils at KA, DI, ME, and GÖ formed from Weichselian loess. In HI, Pleistocene periglacial slope deposits (PPSD) from loess and underlying weathered Lower Devonian silt and sandstones led to expressed topsoil heterogeneity as described by Heggemann et al. (2017). Soils at PO developed from silty Holocene fluvial deposits and were similar to loess soils with regard to texture, pH, pedogenic iron and OM content. Only the HI soils partly revealed stagnic properties. All samples fell in texture classes of silt loam and loam. Mineralogical composition was not investigated because loess in the Bonn region was considered homogeneous. For ME and GÖ, divergent mineralogical composition could not be excluded with respect to the regionally deviating loess origin, although the clay fraction generally reveals little difference in mineralogy throughout the German loess belt (Gehrt 2000). Soil samples for chemical and spectroscopic investigation were taken from the plough layer (0–0.3 m depth) and were carbonate-free except at GÖ where 18 of 109 samples contained 0.1 to 0.9% carbonate. The pH values hardly varied between locations except at HI where pH was approximately one unit lower than in the other fields (Table 1).

To achieve a maximum range of P content at comparable soil properties, two long-term fertilising experiments (LTFE) at PO and DI with 68 and 120 plots were included in the study. For the DI-LTFE, effects of fertilisation, pre-crop and sampling date on spectral properties could be detected; therefore, sample sets were not only assigned to the treatments, but also re-grouped as listed in Table 1. The GÖ field of 30 ha comprises a variable-rate fertilising experiment on 72 m wide strips that was sampled on a 36 m grid. At KA, four fields with similar tilling and fertilising practices (only mineral fertiliser) were sampled. The large sample numbers from small fields rely on barley variety plot trials with variable N and uniform P fertilisation. The HI field had been cropped with cereals over the previous 4 years when sampled; before, it had been a pear orchard with grassed tracks for at least 30 years. The sampling points were located on a regular grid with 21 m width.

P analyses and classification of CAL-P contents

In all samples, P_{avl} was determined according to the CAL method (Schüller 1969). The CAL extraction (buffered at pH 4.1) is recommended as standard method by the Association of German Agricultural Analytical and Research Institutes (VDLUFA) and applied by official advisory services (LWK 2015). Fertilising recommendation in North Rhine-Westphalia, e.g., relies on five classes of CAL-P contents from very low (A) to very high (E) with allowance for soil texture. Wuenscher et al. (2015) compared various P extraction

Table 1 Set of soil samples under study. At the localities Dikopshof and Klein-Altendorf, sample sub-sets [experimental blocks (same crop), treatments (different crops)], and fields, respectively] were investigated (sub-sets in parentheses)

Locality	Block/treatment/field	Abbreviation	Size (ha)	N	Sampling date (months after harvest)	Soil surface/pre-crop	CAL-P (mg kg⁻¹) Min–Max	Mean±SD	Median	P class[a]	Organic carbon (%) Mean±SD	C:N ratio Mean±SD	pH Mean±SD
Dikopshof[f]		DI	3.7	120	3/2009		15–275	116±61	107	A–E (D)	1.01±0.20	11.6±0.63	
	Block I	(DI-I)		24	(5)	Bare[e]/Persian clover	18–275	123±68	111	B–E (D)	0.97±0.19	11.4±0.37	
	Block II	(DI-II)		24	(5)	Bare/potato	23–235	121±60	108	B–E (D)	0.97±0.19	11.0±0.89	
	Block III	(DI-III)		24	(3–4)	Winter wheat/sugar beet	23–222	121±59	109	B–E (D)	1.03±0.20	11.7±0.35	
	Block IV	(DI-IV)		24	(7)	Winter rye/winter wheat	15–234	112±62	104	A–E (D)	1.02±0.21	11.8±0.41	
	Block V	(DI-V)		24	(7)	Bare/winter rye	17–196	105±52	96	B–E (D)	1.05±0.19	12.0±0.40	6.39[b]
	With FYM	(+FYM)		60			53–275	154±56	160	C–E (E)	1.19±0.07	11.7±0.34	6.47
	Without FYM	(−FYM)		60			15–163	78±39	86	A–E (C)	0.83±0.07	11.5±0.81	6.30
	With mineral P	(+min P)		80			50–275	143±53	135	C–E (E)	1.02±0.19	11.6±0.56	6.58
	Without mineral P	(−min P)		40			15–142	63±36	54	A–D (C)	0.98±0.20	11.6±0.74	6.15
Poppelsdorf Klein-Altendorf		PO	0.7	67	10/2011 (3)	Bare/cereals	82–216	140±38	136	C–D (D)	1.22±0.15	12.6±1.36	6.6±0.3
		KA		210									6.5–7.0[c]
	Field 1	(KA-1)	0.24	69	10/2011 (3)	Stubble/summer barley	58–105	81±11	81	C–D (C)	1.07±0.04	10.1±0.40	
	Field 2	(KA-2)	0.75	32	11/2011 (1.5)	Beet leave/sugar beet	98–274	174±47	166	D–E (E)	1.57±0.21	10.9±0.43	
	Field 3	(KA-3)	0.24	77	3/2012 (7)	Bare/grain maize	72–206	117±32	108	C–E (D)	1.27±0.11	10.3±0.38	

Table 1 (continued)

Locality	Block/treatment/field	Abbreviation	Size (ha)	N	Sampling date (months after harvest)	Soil surface/pre-crop	CAL-P (mg kg^{-1})				Organic carbon (%)	C:N ratio	pH
							Min–Max	Mean ± SD	Median	P class[a]	Mean ± SD	Mean ± SD	Mean ± SD
	Field 4	(KA-4)	0.75	32	3/2012 (7)	Bare/winter wheat	34–64	47 ± 6	46	B–C (C)	1.11 ± 0.09	10.7 ± 0.64	
Hilberath		HI	2	42	1/2013 (5)	Stubble/winter wheat earlier: pear orchard	3–49	19 ± 13	15	A–C (B)	1.45 ± 0.27	12.4 ± 1.61	5.4 ± 0.4
Merklingsen		ME	6	38	11/2011 (5)	Stubble/oilseed rape	46–177	79 ± 26	68	C–E (C)	1.69 ± 0.21	10.6 ± 0.32	6.7[d]
Görzig		GÖ	30	110	07/2016 (1)	Stubble/winter wheat	8–164	55 ± 31	52	A–E (C)	1.67 ± 0.17	11.8 ± 0.53	6.3 ± 0.4

All samples were subject to Vis–NIR and MIR spectral investigations except samples from Görzig that were investigated by MIRS only

[a] Range of P classes occurring at the respective field, mean P class in parentheses; CAL-P classes according to LWK (2015); "A"—very low, "B"—low, "C"—optimal, "D"—high, "E"—very high; for P values, see Table 6

[b] pH values only available for block V

[c] No pH analyses were conducted in the breeding experiments on fields KA-1–4; the experimental farm generally maintains pH values between 6.5 and 7.0

[d] Only organic carbon and pH values for composite sample from entire field available

[e] Bare surface after ploughing (no residues at the surface)

[f] Since 1904 differentiated P fertilising regime: with farmyard manure (FYM, 221 kg P ha^{-1} 5 a^{-1}) or without FYM; with mineral P=plots receiving mineral P fertiliser (155–287 kg P ha^{-1} 5 a^{-1} depending on treatment, P fertilisation until 2000 as Thomas slag, from 2000 as triple super phosphate); without mineral P=no mineral P fertiliser

methods and found a close correlation between CAL-P and Olsen-P (r = 0.801***); yet CAL extracted slightly higher P amounts than Olsen.

Spectra acquisition

Vis–NIRS

Spectra were recorded in the laboratory using an AgriSpec Vis–NIR spectrometer with 350–2500 nm range (Analytical Spectral Devices Inc.; Boulder, CO, USA). Prior to spectra acquisition, soil samples were air-dried, sieved and pestled. Samples were placed and levelled in a petri dish and three mean spectra of 30 single spectra each in slightly different positions were taken as described by Rodionov et al. (2014b). Spectra were taken from a total of 477 samples. No Vis–NIR spectra could be recorded from the archived GÖ samples because not enough sample material was available after conventional soil analyses for the spectral investigation that was not initially planned.

MIRS

Mid-infrared spectra were recorded with a benchtop Bruker Tensor 27 spectrometer (Bruker Optik, Ettlingen, Germany). The laboratory spectrometer was equipped with an automated high throughput device (Bruker HTS-XT) as described by Patzold et al. (2008). Soil samples were air-dried, sieved, and milled prior to spectra acquisition as recommended by Stumpe et al. (2011). The recorded wavelengths were 1250–25,000 nm. Five composite measurements, comprising 120 scans at a time, were conducted for each sample. A total of 586 soil samples were examined.

Spectra analyses and model building

Partial least squares regression (PLSR) with leave-one-out cross-validation was carried out for model calibration using the software "ParLeS" (Viscarra Rossel 2008). Spectra were truncated to 500–2400 nm (Vis–NIR) and 2500–16,667 nm (MIR) wavebands to eliminate noise at the edges. Reflectance [R] was transformed to log (1/R). To account for site-to-site variance, all Vis–NIR spectra were mean-centered prior to modelling. After testing diverse data pre-treatments, the MIR spectra were subjected to PLSR as first derivative. For composite models from different sample sets, data were randomly divided into calibration (2/3) and validation sets (1/3). An exception was the entire DI-LTFE dataset, where three blocks with 24 plots each were used for calibration and two blocks with 24 plots each for test-set validation to achieve a balanced representation of the fertilising treatments.

    The samples were taken from various previous projects (see above); in consequence, CAL-P values were not optimally distributed in some of the sample sub-sets (see Supplementary Material, Tables S1–S3). Nevertheless, the full dataset was utilized to encompass the expected variability of future prediction datasets and to ensure the representativeness of samples. For practical usefulness, prediction models should be applicable to independent, i.e., unknown sample sets. Therefore, diverse combinations of calibration and independent test-set validation sample sub-sets were utilized. Calibrated models were transferred to sites that did not contribute to model building. For this test-set validation, complete datasets of the validation site were used. To improve model transferability between localities, the spiking approach as reviewed by Stenberg et al. (2010) was tested for some exemplary

cases. Prior to calibration, ten and eight randomly selected samples from other sites were added to samples sets from one (KA) and four (KA, DI, HI, ME) localities, respectively. Statistical parameters of CAL-P distribution for the diverse calibration and validation data sets and sub-sets are listed in Tables S1–S3.

The number of latent variables, herein denoted as "factors" (Viscarra Rossel 2008), was limited to ten to avoid overfitting. Quality of calibration models and validation was evaluated using R^2, ratio of performance to deviation (RPD) and root mean square error (of cross-validation or of prediction, resp., $RMSE_{CV}$ or $RMSE_P$). Applicability of RPD was generally questioned by Bellon-Maurel et al. (2010), and RPD threshold values have been controversially discussed (Bellon-Maurel and McBratney 2011). Nevertheless, RPD is one of the most popular parameters to evaluate model performance (Bellon-Maurel et al. 2010) and still applied in recent studies (Wijewardane et al. 2018; Mayrink et al. 2019). Accordingly, this study reports RPD and model quality categories as proposed by Chang et al. (2001): (A) excellent models (RPD > 2); (B) acceptable models (1.4 < RPD < 2); (C) non-reliable models (RPD < 1.4). Values for RMSE are important for evaluation of model quality and so were reported for all cross-validations and predictions.

For Vis–NIRS and the MIRS models comprising the entire dataset, the most relevant spectral regions were identified as described in detail by Gomez et al. (2008) to ensure that soil constituents being potentially relevant for phosphate bonding were adequately considered. When (i) the variable importance in the projection (VIP; Viscarra Rossel 2008) exceeded the value 1 and (ii) at the same time, the PLS regression coefficient (b-coefficient) was greater than its standard deviation, the respective wavelength was considered significant.

Calculation of cellulose absorption index (CAI)

In order to better differentiate the various sample sets with respect to their spectral properties, the CAI was calculated. This index is used in Vis–NIRS analyses to detect the presence of plant litter. For MIR, the use of the CAI has not yet been reported in the literature. Values > 0 generally point to the presence of lignin and cellulose, and CAI values of ageing litter decrease over time due to lignin and/or cellulose decomposition (Nagler et al. 2000). Further, CAI can indicate presence of pre-crop residues (Aguilar et al. 2012). The CAI was derived from Vis–NIR spectra as $CAI = 0.5 \cdot (R_{2025} + R_{2215}) - R_{2110}$ from the sums of reflectance factors R at the wavebands 2000–2050 nm (R_{2025}), 2089–2130 nm (R_{2110}) and 2190–2240 nm (R_{2215}), respectively (Rodionov et al. 2016).

Results and discussion

Testing the general performance of CAL-P prediction on different fields

Vis–NIRS

Quantification of CAL-P with Vis–NIRS was successful when the entire dataset (n = 477) was used for cross-validation (Fig. 1a). Prediction quality was acceptable with RPD = 1.84 and $R^2 = 0.703$ though $RMSE_{CV}$ equalled 30.3 and was not satisfactory for precise fertiliser dosage. Soils with high P demand would not have been fertilised and experience yield loss
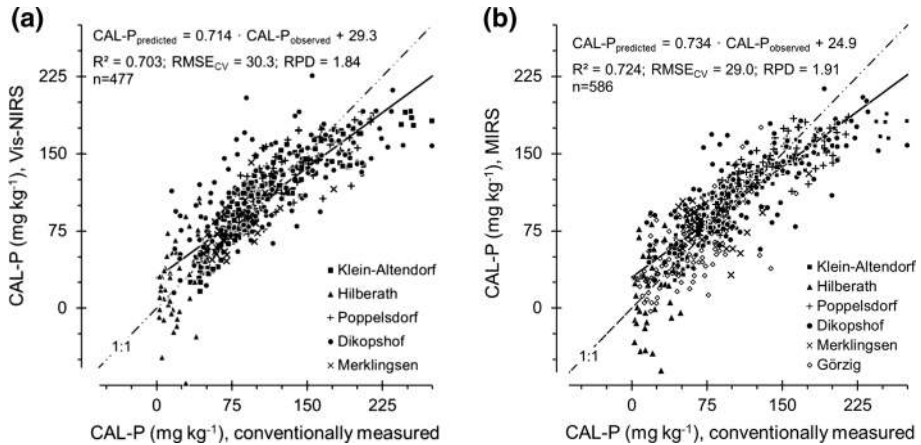
Fig. 1 Prediction of plant-available P (calcium acetate lactate extractable, CAL-P) with **a** Vis–NIR spectroscopy and **b** MIR spectroscopy in topsoil samples from different fields in Germany (laboratory spectra, PLSR leave-one-out cross validation)

whereas soils without P demand would have unnecessarily received P fertiliser and experienced increased risk of loss to the environment.

The fundamental spectral response in the MIR region and the electron transitions and overtones in the Vis–NIR region pointed to iron oxides, clay minerals and diverse functional groups of SOM (Viscarra Rossel and Behrens 2010) as major binding partners for CAL-P. The exact spectral regions that were selected for model calibration by the ParLeS software are depicted in Figs. S1 and S2. Neighbouring components and functional groups can lead to spectral shifts in the NIR.

Calibrating individual models for single fields led to highly variable results, although the same data were used as for the overall calibration (Table 2). Excellent CAL-P prediction was obtained for two of the five sample sets and three of the 13 subsets. Figure 2a illustrates the reliable (RPD = 4.31, category A) and precise (RMSE$_{CV}$ = 11.1 mg P kg^{-1} soil) result for field KA-2. Sample sets from other fields yielded worse, but acceptable, results (Fig. 2b for PO, Table 2). In some cases, field-individual prediction failed completely (Fig. 2c).

One reason for the partly poor results in P prediction might be the data range that differed among the fields (see Table 1). On the KA fields, fertilisation was not systematically varied, and CAL-P variability consequently relies on natural soil heterogeneity. In contrast, the LTFEs at DI and PO were established on homogeneous fields and the fertilizer treatments have been established for more than 100 years. Consequently, some differences in OM and pH have developed (Table 1). In addition, indirect effects such as variable crop growth have influenced SOM content and P binding and availability. As CAL-P is not defined by a specific binding partner but correlated to various soil properties (Wuenscher et al. 2015), the fertiliser-induced variability of soil properties makes CAL-P prediction complex and difficult. This is perhaps the reason why P prediction was much better on fields KA-2 and KA-3, where fertilisation has not been systematically varied (Table 1). Conceptually, the same is true for field KA-1 where P prediction was acceptable and KA-4 where prediction failed (Fig. 2). However, KA-1 was sampled shortly after barley harvest and the small CAL-P range in KA-4 probably hampered a better calibration result.

Table 2 Vis–NIRS calibration models (leave-one-out cross validation, CV) for different entire sample sets

	Sample set and sub-set	N	F	R^2	$RMSE_{CV}$	RPD^a	$Model^b$ [slope; intercept]	Performance
1	DI	120	10	0.72	32.7	1.88	0.774; 23.3	B
2	DI-I	24	10	0.54	49.8	1.40	0.732; 27.9	B
3	DI-II	24	10	0.76	31.6	1.95	0.887; 8.8	B
4	DI-III	24	10	0.33	54.0	1.12	0.562; 50.2	C
5	DI-IV	24	9	0.81	27.3	2.33	0.858; 14.4	A
6	DI-V	24	10	0.69	29.8	1.77	0.787; 26.0	B
7	DI + FYM	60	6	0.578	36.1	1.55	0.626;57.2	B
8	DI − FYM	60	10	0.44	31.1	1.26	0.625; 30.0	C
9	DI + mineral P	80	10	0.74	27.3	1.95	0.792; 30.5	B
10	DI − mineral P	40	9	0.65	21.8	1.66	0.757; 14.9	B
11	PO	67	10	0.79	17.4	2.17	0.835; 22.8	A
12	KA	210	10	0.90	14.8	3.21	0.907; 9.6	A
13	KA-1	69	10	0.60	7.1	1.57	0.680; 25.9	B
14	KA-2	32	10	0.95	11.1	4.31	0.981; 2.6	A
15	KA-3	77	10	0.93	8.4	3.81	0.944; 6.1	A
16	KA-4	32	10	0.04	7.5	0.80	0.209; 37.5	C
17	HI	42	10	0.27	11.6	1.10	0.461; 9.8	C
18	ME	38	5	0.45	19.5	1.35	0.536; 35.7	C

All spectra were transformed to absorbance (log [1/R] prior to PLSR). N = Number of samples, F = number of PLSR factors. CV for the complete sample set (N = 477) is depicted in Fig. 1. Statistical parameters for the sample sets are listed in Table S1

[a]RPD, ratio of performance to deviation (model reliability): A = excellent, RPD > 2; B = acceptable, 1.4 < RPD < 2; C = non-reliable, RPD < 1.4

[b]Model equation: CAL-P$_{predicted}$ = [slope] × CAL-P$_{observed}$ + [intercept]

Prediction failed in HI although the range in CAL-P was large. Here, the former long-term use as a pear orchard caused highly variable OM quality and P binding conditions due to wood residues in the former tree strips, and grassed inter-row spaces. Soil heterogeneity at HI also was pronounced with respect to relief and parent material (Heggemann et al. 2017).

MIRS

In general, MIRS cross-validations were comparable to those in the Vis–NIR range, with some exceptions (Fig. 1b and Table 3). For the entire dataset, RPD was slightly better, but still only acceptable. Although the data range was large, CAL-P prediction via MIRS was poor for the GÖ sample set (Fig. 2d). Excellent to acceptable performance was achieved for DI-all, PO and KA (except KA-4), as in the Vis–NIRS. Hence, in contrast to the statement of Kruse et al. (2015), MIRS did not generally yield generally better results than Vis–NIRS.
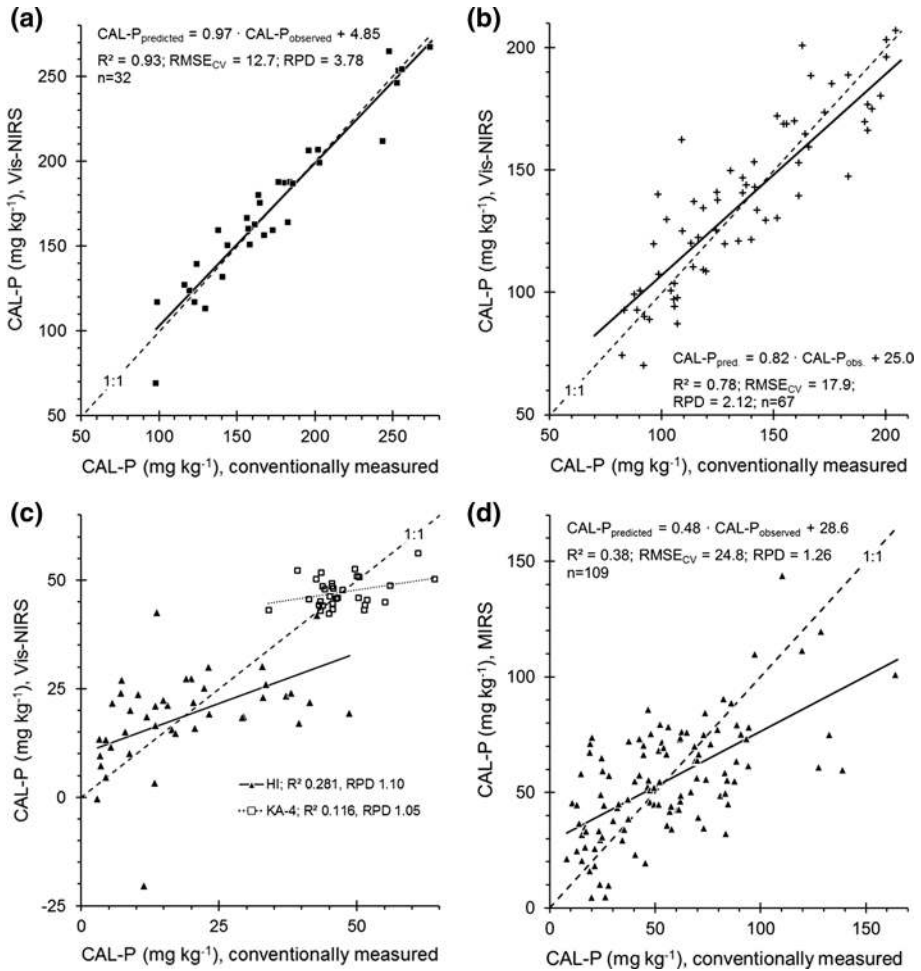
Fig. 2 Prediction of CAL-P with Vis–NIR spectroscopy (**a–c**) and MIR spectroscopy (**d**) in topsoil samples of single fields (leave-one-out cross-validation): **a** KA-2, **b** PO, **c** HI and KA-4, and **d** GÖ. For sample set designations, refer to Table 1

Transferring calibration models to leave-out fields

Transferability of models between fields from one locality

All Klein-Altendorf fields belong to the same experimental farm, reveal similar soils and are managed in the same way. The dataset for each of the fields was randomly divided into calibration and validation data. Reference data distributions for all sample sets are compiled in Tables S1–S3. In a first step, fields KA-1–3 were merged, and field KA-4 was excluded because here CV had failed. Calibration (CV) and test-set validation (P) performed "A" in the Vis–NIR as well as in the MIR range (Table 4, lines 1–2 and 13–14). Again, Vis–NIRS and MIRS performed comparably. These results and good performance

Table 3 MIRS cross-validation results for CAL-P (entire datasets)

Sample set and sub-set	N	F	R^2	$RMSE_{CV}$	RPD^a	Model[b] [slope; intercept]	Performance
DI	120	10	0.81	26.2	2.22	0.909; 2.70	A
DI-I	24	8	0.36	59.3	1.18	0.597; 54.9	C
DI-II	24	5	0.41	49.4	1.25	0.597; 52.8	C
DI-III	24	5	0.64	36.1	1.68	0.709; 38.0	B
DI-IV	24	3	0.31	52.4	1.22	0.381; 68.9	C
DI-V	24	9	0.69	28.9	1.83	0.749; 25.1	B
DI + FYM	60	8	0.42	43.7	1.28	0.548; 70.6	C
DI − FYM	60	10	0.35	32.9	1.19	0.503; 39.4	C
DI + mineral P	80	10	0.77	25.5	2.08	0.821; 26.4	A
DI − mineral P	40	6	0.65	21.3	1.70	0.736; 16.8	B
PO	67	8	0.75	18.8	2.00	0.747; 35.9	A–B
KA	210	8	0.91	14.6	3.26	0.911; 9.14	A
KA-1	69	6	0.55	7.51	1.49	0.616; 31.0	B
KA-2	32	6	0.83	19.5	2.46	0.829; 29.2	A
KA-3	77	7	0.85	12.3	2.61	0.877; 14.1	A
KA-4	32	10	0.02	8.68	0.69	− 0.141; 53.8	C
HI	42	9	0.34	10.2	1.24	0.389; 11.1	C
ME	38	10	0.25	23.1	1.14	0.346; 49.5	C
GÖ	110	6	0.39	24.8	1.26	0.478; 26.6	C

All spectra were subjected to PLSR as 1st derivative. F = factors in leave-one-out cross validation (CV). For annotations, see Table 2. Comprehensive CV for the entire sample set (N = 586) is shown in Fig. 1. Statistical parameters for the sample sets are listed in Table S1

[a]RPD, ratio of performance to deviation (model reliability): A = excellent, RPD > 2; B = acceptable, 1.4 < RPD < 2; C = non-reliable, RPD < 1.4

[b]Model equation: CAL-P$_{predicted}$ = [slope] × CAL-P$_{observed}$ + [intercept]

for the single datasets (Tables 2 and 3) gave rise to the expectation of satisfactory transferability tests.

However, calibrating individual models (CV) for each KA field and transferring them to other KA fields for test-set validation (P) yielded mixed results. In most cases, no CAL-P prediction was possible with Vis–NIRS calibration models, although model calibration had yielded $RMSE_{CV}$ values that were acceptable (Table 4). For subsequent prediction, R^2 as well as RPD were small and most $RMSE_P$ values were too large to serve as a basis for precise fertiliser dosage when compared to the recommended soil P values (Table 6). This result was insofar unexpected as the ranges in values of calibration and validation data were wide and overlapping (Table S2). Only CAL-P prediction for field KA-2 was acceptable after calibrating a model with KA-3 data. For the MIR range, some predictions were slightly better (Table 4), but the overall results were comparable to those of Vis–NIRS. Consequently, in this respect MIRS was not superior to Vis–NIRS.

Figure 3a illustrates Vis–NIRS prediction for KA-2 based on the KA-3 model. Prediction of KA-3 completely failed when KA-2 were used for calibration (Fig. 3b). It is assumed that soil properties connected to CAL-P or the IR-active soil components related to P extractability varied between the two fields. Sampling date, soil surface cover and pre-crop were different for KA-2 and KA-3 (Table 1). Field KA-2 was sampled in November

Table 4 Testing model transferability within a farm at one locality: results of model calibration (leave-one-out cross-validation, CV) at different sites and subsequent application of prediction model to different but comparable fields (P; test-set validation); field KA-4 was excluded from this test because single-field CV failed (see Tables 2 and 3); spectra pre-treatment Vis–NIRS: log (1/R), MIRS: 1st derivative. Statistical parameters for the sample sets are listed in Table S2

	Step	Sample sub-set	N	F[b]	RMSE$_{CV}$	R^2	RMSE$_P$	RPD[a]	Model[a] [slope; intercept]	Performance[a]
Vis–NIRS										
1	CV	KA-1–3	124	10	13.8	0.894		3.08	0.903; 10.7	A
2	P	KA-1–3	54	9		0.895	16.3	3.03	0.821; 22.2	A
3	CV	KA-1 and KA-2	101	9	12.0	0.947		4.35	0.950; 5.6	A
4	P	KA-3	77			0.157	32.7	0.98	0.174; 82.4	C
5	CV	KA-1 and KA-3	146	10	12.4	0.834		2.46	0.854; 14.7	A
6	P	KA-2	32			0.771	45.9	1.05	0.342; 82.0	C
7	CV	KA-2 and KA-3	109	9	12.3	0.926		3.69	0.934; 8.3	A
8	P	KA-1	69			0.136	126.9	0.09	0.959; -121	C
9c	CV	KA-2	32	10	11.1	0.946		4.31	0.981; 2.6	A
10.1	P	KA-3	77			0.168	75.7	0.42	0.293; 13.7	C
10.2	P	KA-3d	32			0.209	85.1	0.35	0.366; 7.4	C
11c	CV	KA-3	77	10	8.4	0.930		3.81	0.944; 6.1	A
12.1	P	KA-2	32			0.636	32.5	1.48	0.448; 107	B
12.2	P	KA-2e	26			0.607	29.1	1.04	0.594; 85.7	C
MIRS										
13	CV	KA-1–3	124	7	14.4	0.884		2.95	0.885; 12.6	A
14	P	KA-1–3	54			0.894	16.7	2.95	0.804; 25.5	A
15	CV	KA-1 and KA-2	101	6	13.9	0.928		3.76	0.925; 8.18	A
16	P	KA-3	77			0.597	25.5	1.26	0.451; 49.9	C
17	CV	KA-1 and KA-3	146	8	11.5	0.857		2.65	0.867; 13.3	A
18	P	KA-2	32			0.724	29.8	1.61	0.600; 54.6	B
19	CV	KA-2 and KA-3	109	5	16.4	0.876		2.85	0.875; 16.4	A
20	P	KA-1	69			0.000	20.9	0.54	0.004; 72.9	C
21c	CV	KA-2	32	7	19.5	0.830		2.46	0.829; 29.2	A

Table 4 (continued)

	Step	Sample sub-set	N	F[b]	R^2	$RMSE_{CV}$	$RMSE_P$	RPD[a]	Model[a] [slope; intercept]	Performance[a]
22	P	KA-3[d]	32		0.251		56.8	0.52	0.386; 56.8	C
23	CV	KA-3	77	5	0.843	12.6		2.54	0.859; 16.4	A
24	P	KA-2	32		0.795		29.2	1.64	0.985; −14.5	B

[a]Refer to Table 2

[b]Number of factors selected for model calibration

[c]See Fig. 3

[d]Dataset randomly reduced to N=32 to adapt to range and size of calibration dataset

[e]Dataset reduced to samples with CAL-P content <203 mg kg^{-1} in order to avoid extrapolation
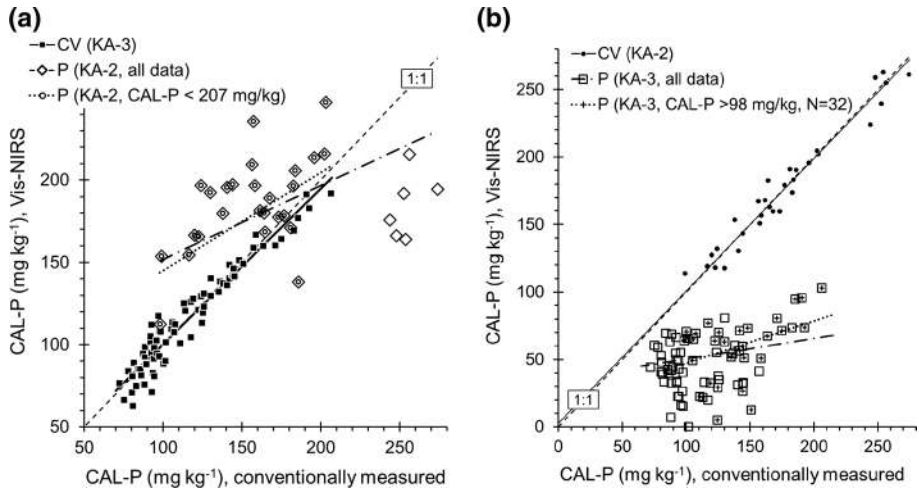
Fig. 3 Transferring Vis–NIRS calibration models between two fields of the same locality (i.e., on the same farm): calibration (CV, leave-one-out cross validation) and CAL-P prediction (P, test-set validation); **a** CV for field KA-3 and P for field KA-2 (entire dataset and for samples < 207 mg $P_{observed}$ kg^{-1} to avoid extrapolation); **b** CV for field KA-2 and P for field KA-3 (entire dataset and reduced dataset with > 98 mg $P_{observed}$ kg^{-1} and N = 32 to avoid extrapolation and imbalanced CV/P sample numbers. Values for RMSE$_{CV}$ and RMSE$_P$ are given in Table 4

approximately 6 weeks after sugar beet harvest prior to tillage, i.e., fresh beet leaf residues were present. In contrast, KA-3 sampling was conducted in March, approximately 7 months after grain maize harvest with subsequent ploughing, cultivation and wheat sowing (October). Conventionally measured CAL-P contents varied between the fields, but the CAL-P ranges largely overlapped.

Transferring models between localities

The performance of calibration and test-set validation was almost identical for the Vis–NIR and the MIR range. Therefore, only the Vis–NIR results are described here and shown in Table 5. The respective MIR results can be found in the supplementary material (Table S4). Calibrating a model for the entire dataset of KA-1–4 (n = 210) yielded, as expected from the results already shown, a highly performing cross-validation (Table 5, line 1). Transferring this model to the localities PO, DI, HI and ME revealed non-satisfactory test-set validations (lines 2–5). In expectation of better validation results, farmyard manure (FYM) treated plots were excluded from the PO and DI sample sets (PO–FYM, DI–FYM) because no FYM had been applied to KA fields. However, prediction performance further decreased (Table 5, lines 6–7) thus indicating a need for widely varying calibration samples when model transfer is envisaged. On the other hand, restricting the DI samples to those receiving mineral fertiliser slightly improved prediction (line 8). Spiking the model by adding ten randomly selected DI samples to the calibration set improved independent P prediction and performed "B" instead of "C" (lines 9–10). Further combinations of single-site CV and prediction were tested, but no satisfactory results were achieved. Omitting KA-4 samples from the CV sample set led to worse P predictions in the test-sets, although KA-4 had failed in the single-field calibration. Overall, the results indicated that calibration sets

Table 5 Testing vis–NIRS model transferability between localities: results of model calibration (leave-one-out cross-validation, CV) on large datasets from different calibration localities and prediction (P; test-set validation) for a single hold-out field (complete dataset of each field). For annotations, see Tables 2, 3, 4

	Step	Sample sub-set	N	Factors	R^2	$RMSE_{CV}$	$RMSE_P$	RPD	model [slope; intercept]	Performance
One locality calibration; without spiking										
1	CV	KA	210	10	0.902	14.8		3.21	0.907; 9.6	A
2	P	PO (all)	67		0.270		44.8	0.84	0.321; 63.7	C
3	P	DI (all)	120		0.475		61.9	0.99	0.414; 111	C
4	P	HI	42		0.090		323	0.04	1.240; 315	C
5	P	ME	38		0.092		61.3	0.43	0.259; 113	C
6	P	PO – FYM	35		0.061		43.4	0.74	0.152; 74.9	C
7	P	DI – FYM	60		0.192		67.8	0.58	0.301; 112	C
8	P	DI +mineral P	80		0.511		43.6	1.22	0.467; 99.2	C
One locality calibration; spiked sample set										
9	CV	KA +10 samples from DI +mineral P	220	10	0.879	16.8		2.88	0.887; 12.1	A
10	P	DI +mineral P – 10 samples	70		0.628		32.7	1.63	0.588; 63.3	B
Multiple locality calibration; without spiking										
11	CV	KA, DI, HI, ME	410	10	0.713	29.9		1.87	0.727; 26.3	B
12	P	PO	67		0.345		31.0	1.22	0.443; 78.2	C
Multiple locality calibration; spiked sample set										
13	CV	KA, DI, HI, ME + 8 samples PO	418	10	0.721	29.5		1.89	0.740; 25.3	B
14	P	PO – 8 samples	59		0.379		33.3	1.15	0.427; 66.3	C

Statistical parameters for the sample sets are listed in Table S3

with expressed variability—also concerning quality of OM as binding partner for P_{avl}—may lead to better model transferability.

Looking for a more universal applicability, the complete datasets of KA, DI, HI and ME were used for calibration and the model was applied to the PO dataset. This combination was chosen because PO yielded a well performing single-locality cross-validation; despite the large variability of the calibration set, test-set validation failed. Spiking the calibration data set with eight randomly selected samples from PO did not improve the result (Table 5, lines 11–14).

Reasons for limited transferability of calibration models in the Vis–NIR and MIR range

The DI-LTFE offered good opportunities to study the effects of pre-crop and sampling date as possible reasons for limited model transferability. All treatments were present in each of the five blocks (balanced sample set). Consequently, CAL-P contents covered similar ranges in each block (Table 1).

Vis–NIRS

Model calibration (CV) yielded satisfactory P prediction, when all 120 plots were considered. The same applied (except for DI-III) when calibration was performed block-wise as already shown in Table 2. Remarkably, selecting calibration and validation data with respect to pre-crop (block-wise) or, alternatively, to treatment led to differentiated results. Again, this points to better performance of heterogeneous sample-sets. Best results were achieved for block-wise calibration (i.e., integrating all treatments) and at least 5 months after harvesting the pre-crop. Note that visible straw and other particles had been thoroughly removed before spectra acquisition. Nevertheless, different combinations of Vis–NIR calibration and validation sample sub-sets led to contrasting results.

Predicting CAL-P in DI-I samples with the DI-II model was satisfactory (performance B), but failed the other way around (not shown). This was insofar surprising as both blocks were sampled at bare surface and 5 months after pre-crops with easily degradable residues. More calibration–validation combinations were tested, but no general rule could be found. It seems that crop residues, perhaps at the molecular scale, hamper model transferability. Separating the DI sample-set into the treatments "with" and "without FYM", i.e., reducing the variability of OM quality, performed only "D" (Table 2, lines 7–8); transferability was consequently not further tested. Much better results were achieved when samples were separated in "with" or "without mineral P" (+ mineral P and − mineral P, respectively; Table 2, lines 9–10). However, validation with the respective opposite treatment failed.

These results point to the complexity of boundary conditions for successful CAL-P prediction. Further, they likely explain why published results on P prediction are somehow contradictory as reviewed by Stenberg et al. (2010) and Soriano-Disla et al. (2014). Over time, decreasing amounts of freshly added phosphates are being extracted by the CAL method (Hartmann et al. 2019). Lozier et al. (2017) found variable concentrations of water-extractable soil P in different cover crops and crop residues over time during the nongrowing season. Accordingly, Delin (2016) reported that release of soluble P from different organic amendments disappeared within 2 months. During breakdown of straw, leaves and root residues, diverse intermediate organic substances formed dissolved OM while revealing seasonal dynamics (Kalbitz et al. 2000). In the MIR, spectral features of dissolved OM

originating from maize and rye or wheat differed (Ellerbrock and Kaiser 2005). Consequently, it is assumed that crop, pre-crop, fertilization regimes, residence time of P released from fertilizer or crop residues and further factors affected CAL-P at the sampling date and model calibration. Such differences were obviously not, or not fully, mirrored in the spectra. Temporarily changing spectral features and their—perhaps incoherent—attribution to CAL-P by PLSR obviously led to highly variable prediction quality. The elemental analyses (organic carbon content, C:N ratio) could not explain the results (Table 1).

To summarize, the influences on P_{avl} were numerous, complex and differed from field to field. Mayrink et al. (2019) observed that lesser factors were sufficient for modelling when taking spectra from exchange resin with adsorbed P instead of scanning soil samples directly. They suggest that this fact is responsible for their very good results. Indeed, the rather high factor number that was needed for model building in this study underlines the problem complexity that probably has decreased model transferability.

MIRS

Overall, for MIRS, the same general statements can be summed up as for Vis–NIRS. However, MIRS-CV results for the single DI blocks and for the treatments $\pm$ FYM were mostly worse than using Vis–NIRS (Tables 2 and 3) and performed only "B" and "C". Consequently no test-set validation was conducted. As for Vis–NIRS, separation into "with" and "without mineral P" yielded very good CV results, but test-set validation with the opposite treatment failed. The factors determining CAL-P are, as discussed above, mostly related to OM quality, and Vis–NIRS obviously better captures the complex situation.

Cellulose absorption index (CAI) as elucidation

The CAI potentially provides information about the presence of variable amounts of degrading plant material (Nagler et al. 2000). Aguilar et al. (2012) showed that CAI is sensitive to partly degraded residues insofar as CAI of soil with pea residues was affected if wheat residues from the previous year were present. Here, CAI was calculated to point out differences in OM quality between the sample sets. Note that CAI is only one possibility to better characterise spectral variability of soil samples; it is shown here as exemplary indicator. Figure 4 shows mean CAI's for different sample sets and sub-sets.

CAIs vary in a large range between the localities but also within some sample sub-sets. Between treatments of DI-LTFE, variation is small. However, pre-crop seemed to affect CAI, because considerable variation arose when the LTFE dataset was sorted by blocks (compare blocks DI-I and DI-III). High standard deviation occurs because sorting by treatments integrates different crops and, in turn, each block comprises all treatments. Accordingly, CAI differences between the KA fields occurred. After sugar beet harvest, the CAI was lower than after cereal harvest at other KA fields, although soil samples were taken only 6 weeks after beet but ≥ 3 months after cereals. For the two fields with the highest CAI values, single-field CV failed: KA-4 was sampled shortly after cereal harvest (Table 2, line 16) and, at the ancient orchard HI (line 17), wooden residues (roots, branches) were still present. This underlines that type of harvest residue—perhaps also at the molecular scale—and sampling date influenced spectra and sometimes hampered CAL-P prediction.

Calibration for DI after separation into treatments $+$ FYM and $-$ FYM failed (Table 2, lines 7–8), although CAIs were similar. Thus, more factors than CAI alone must have been relevant. Note the small variability of organic carbon contents within the sample sets
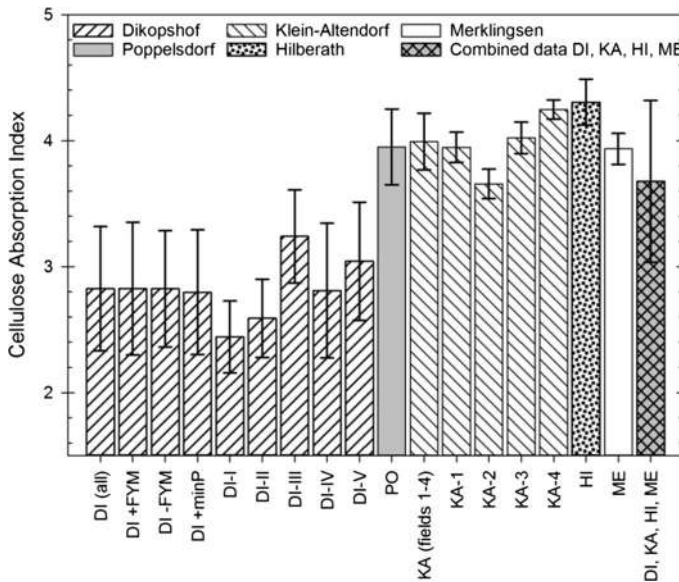
Fig. 4 Cellulose absorption index for different sample sets and sub-sets. Error bars represent standard deviation

+FYM and − FYM compared to any other grouping within DI (Table 1). The better performance of the models + minP and − minP compared to the FYM grouping was perhaps due to the larger organic carbon content range. Accordingly, Abdi et al. (2012) were not able to predict P_{avl} with Vis–NIRS and concluded that the low correlation of the P related properties and soil C were the reason. However, the authors combined soil samples from 6 years of a plot experiment, and the results here additionally turned out the effect of precrop and harvest residues on P prediction (see above).

Of course, CAI can only serve as a proxy for changing OM properties and related P binding. This becomes obvious by the weak transferability of the universal calibration from DI–KA–HI–ME to the validation site PO. The CAI mean and standard deviation of the DI–KA–HI–ME dataset include CAI for PO (Fig. 4), suggesting good transferability. Nevertheless prediction (without or with spiking) was not satisfactory (Table 5). Consequently CAI alone is not an appropriate parameter to decide upon model application to unknown samples. Yet, in this study, neither the details of P binding nor spectral features of OM were elucidated. Further, more factors controlling CAL-P (e.g., iron oxides) and related spectral features should be considered but were not investigated in this study.

Utilisation of IR-based P prediction in fertilisation practice

In Germany, P dosage recommendation relies on soil CAL-P content, in some regions also on soil texture, and crop yield (Table 6, see also Jordan-Meille et al. 2012). In view of the primary goal to integrate spectroscopic P determination into agricultural practice, a correct estimation of the P class is sufficient and marks a step towards variable rate P dosage.

Applying the universal Vis–NIRS and MIRS models, > 60% of the predictions met the correct CAL-P class (Table 6). The CAL-P status prediction was better for soils in classes

Table 6 Classes of CAL-soluble phosphorus contents in arable topsoils, P fertilization recommendation for North Rhine-Westphalia (LWK 2015), and percentage of correct and wrong fertiliser recommendation on the basis of spectroscopic P prediction

Texture class	CAL-soluble P content[a] (mg P/kg soil)				
Light and medium soils except shallow sand (not present in this study)	≤1 A	15–41.5 B	41.5–81 C	81–142 D	≥142 E
Heavy soils[b] and shallow sandy soils	≤24 A	24–59 B	59–107 C	107–168 D	≥168 E
	% of crop removal				
Officially recommended fertilisation	166	133	100	50	0
	% of samples				
Vis–NIRS, complete sample set (N = 477), universal model					
IR-based P recommendation too high/**correct**/too low	0/**45**/55	9/**56**/35	5/**62**/33	16/**77**/7	62/**38**/0
MIRS, complete sample set (N = 586), universal model					
IR-based P recommendation too high/**correct**/too low	0/**39**/61	24/**48**/28	10/**67**/23	16/**78**/6	62/**38**/0

Percentage of cases with correct P recommendation are given in bold

Original table gives P as P_2O_5 and soil P content in mg 100 g^{-1}

[a] Original table lists CAL-P as P_2O_5. Here, P was recalculated as $P_2O_5 \cdot 0.4364$. Letters A–E correspond to P classes from A, very low to E, very high CAL-P content (see Table 1)

[b] Heavy soils are defined as having > 65% silt AND > 25% clay OR 50–65% silt AND > 30% clay OR 30–50% silt AND > 35% clay OR < 30% silt AND > 45% clay)

B (low P), C (optimal status), and D (high P status) than for soils with very low (class A) or very high (class E) CAL-P contents. In the latter cases, spectroscopic P estimation would often lead to wrong fertilising recommendations. Consequently, the procedure cannot yet fully be recommended for practical application. However, for some single locations, results were better than for the entire dataset (e.g., for KA 73%, not shown). Accordingly Stenberg et al. (2010) and Gholizadeh et al. (2013) reported that local calibrations outperform global ones, although spectral libraries from numerous sites can be advantageous for calibrating models (Wijewardane et al. 2018). Until now, Vis–NIR and MIR spectral libraries for P_{avl} in German arable soils are not available. Considering the results of this study will help to construct such libraries and to reduce costs for P analyses. The calculated acceptable costs for sensor-based variable-rate P fertilisation allow for future introduction of spectroscopic P_{avl} determination in precision farming practice as pointed out by Schulte-Ostermann and Wagner (2018) for the Görzig site.

Limitations of the present study

Currently, the results are only valuable for the German standard CAL method. Nevertheless, the findings can help to improve P model calibration also for other soil tests that are operationally and not chemically defined. However, individual calibration of spectral models is still mandatory for each P extraction method.

Surprisingly, no clear preference for Vis–NIRS or MIRS could be derived from the results, because performance was similar under laboratory conditions. Both techniques revealed pros and cons. Although MIRS has not so far been introduced into mobile field applications, the results of this study encourage testing portable MIR spectrometers in the field.

The overall unsatisfactory prediction results were perhaps the consequence of dataset composition: the calibration and validation samples were not pre-selected but taken from previous projects as entire sets. This decision had been taken in order to explore the value of archived samples and data for future model calibration on a regional basis. However, it is widely accepted that calibration and validation samples should be selected to cover a wide range of the parameter of interest (Stenberg et al. 2010; Soriano-Disla et al. 2014).

This study was based on the linear standard PLSR calibration. Non-linear machine learning approaches can overcome some modelling restrictions (Gholizadeh et al. 2013). Nawar and Mouazen (2019) successfully performed random forest SOC modelling on a spectral library that was spiked with selected samples from the target field. As a consequence, the authors recommended testing the approach with other soil properties. This present study confirms this statement.

While standardizing spectra recording (surface properties, light source, geometry, etc.) is known to be an important issue when building spectral libraries (Gholizadeh et al. 2013; Soriano-Disla et al. 2014), no related references were found concerning the soil sampling itself. This study showed that sampling season, pre-crop, etc. affect calibration and prediction performance. One reason was, presumably, the presence of different harvest residues of microscopic or molecular size. Such residues impeded successful P prediction when the test-set samples contained no, less, or other such substances. In this respect, spiking the calibration set resulted in no significant prediction improvement. Consequently, for more universal validity of P prediction models, attention can easily and must thoroughly be paid to composition of the sample sets. This can, e.g., necessitate strict co-ordination of sampling campaigns or standardization of sampling guidelines with respect to pre-crop and sampling date. Although not examined in

this study that focused on non-stagnic loess soils, it must be assumed that also parent material and pedogenic iron oxides should be taken into account. Further work is needed to systematically elucidate spectral features of factors that control P_{avl} extractability. Future research should focus on standardisation of the sampling procedure in order to enhance performance of Vis–NIRS, and MIRS prediction of P_{avl}. The development of a sampling protocol to build a broader and valuable data pool as a basis for further model improvement is recommended. This guideline should include standards for calibration sample collection including pre-crop species, sampling date and presence of harvest residues. Such sampling protocol may make it possible to improve transferability and universal validity of P_{avl} models. With this precondition, P_{avl} prediction using Vis–NIRS or MIRS will probably further develop towards practical applicability in precision agriculture.

Conclusions

Arable soil samples from multiple sites in Germany were used to evaluate performance of Vis–NIR and MIR spectroscopy for predicting CAL-P. In the laboratory, Vis–NIRS and MIRS proved to be potentially useful for developing future applications in mobile sensing. Performance of Vis–NIRS or MIRS conditions was similar and no preference could be derived. Analysis of the spectral data resulted in the rapid quantification of plant available CAL-P at a moderate level of precision (30.3 mg kg^{-1}). Unfortunately, P prediction often failed for independent test fields indicating a lack of transferability, occasionally even between neighbouring fields being similarly managed. Therefore, universal model validity and transferability to unknown sites are not yet satisfactory. A wide range of CAL-P alone did not ensure satisfactory modelling results. Variability in P extractability and spectral reflectance properties remained unexplained, but were ascribed to season, fertilising regime, pre-crop and harvest residues. The CAI—although originally designed to distinguish plants and soil in remote sensing—proved useful to explain varying transferability. However, CAI alone was not a reliable indicator to predict model transferability. Non-linear machine learning modelling instead of PLSR might help to better account for the complex problem of P_{avl} prediction. Anyway, Vis–NIRS and MIRS at their current state are already sufficient to outperform the actual practice of evaluating P requirement, because the models frequently predicted the correct P recommendation for soils classified as having low, optimal or high CAL-P contents. The application of a future sampling guideline will most probably lead to better P_{avl} predictions. The expected increase in prediction performance of Vis–NIRS and MIRS makes an introduction into precision agriculture seem possible.

References

Abdi, D., Cade-Menun, B. J., Ziadi, N., Tremblay, G. F., & Parent, L.-É. (2016). Visible near infrared reflectance spectroscopy to predict soil phosphorus pools in chernozems of Saskatchewan, Canada. *Geoderma Regional, 7,* 93–101.

Abdi, D., Tremblay, G. F., Ziadi, N., Bélanger, G., & Parent, L.-É. (2012). Predicting soil phosphorus-related properties using near-infrared reflectance spectroscopy. *Soil Science Society of America Journal, 76,* 2318–2326.

Aguilar, J., Evans, R., & Daughtry, C. S. T. (2012). Performance assessment of the cellulose absorption index method for estimating crop residue cover. *Journal of Soil and Water Conservation, 67,* 202–210.

Ahmed, A. A., Gypser, S., Leinweber, P., Freese, D., & Kühn, O. (2019). Infrared spectroscopic characterization of phosphate binding at the goethite–water interface. *Physical Chemistry Chemical Physics, 21,* 4421–4434.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends in Analytical Chemistry, 29,* 1073–1081.

Bellon-Maurel, V., & McBratney, A. (2011). Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils: Critical review and research perspectives. *Soil Biology & Biochemistry, 43,* 1398–1410.

Chang, C.-W., Lair, D. A., Mausbach, M. J., & Hurburgh, C. R., Jr. (2001). Near-infrared reflectance spectroscopy: Principal components regression analyses of soil properties. *Soil Science Society of America Journal, 65,* 480–490.

Csatho, P., Sisak, I., Radimszky, L., Lushaj, S., Spiegel, H., Nikolova, M. T., et al. (2007). Agriculture as a source of phosphorus causing eutrophication in Central and Eastern Europe. *Soil Use and Management, 23,* 36–56.

Delin, S. (2016). Fertiliser value of phosphorus in different residues. *Soil Use and Management, 32,* 17–26.

Ellerbrock, R. H., & Kaiser, M. (2005). Stability and composition of different soluble soil organic matter fractions: Evidence from $\delta^{13}C$ and FTIR signatures. *Geoderma, 128,* 28–37.

Gebbers, R., & Adamchuk, V. I. (2010). Precision agriculture and food security. *Science, 327,* 828–831.

Gebbers, R., Dworak, V., Mahns, B., Weltzien, C., Büchele, D., Gornushkin, I. et al. (2016). Integrated approach to site-specific soil fertility management. In *Proceedings of the 13th International Conference on Precision Agriculture* (unpaginated, online). Monticello, IL, USA: International Society of Precision Agriculture. Retrieved October 10, 2019 from https://www.ispag.org/proceedings/?action=abstract&id=2084.

Gehrt, E. (2000). Nord- und mitteldeutsche Lössbörden und Sandlössgebiete (North and Central German loess börde and sand loess areas; in German). In H.-P. Blume, P. Felix-Henningsen, H.-G. Frede, G. Guggenberger, R. Horn, & K. Stahr (Eds.), *Handbuch der Bodenkunde*. Weinheim, Berlin, Germany: Wiley.

Gholizadeh, A., Borůvka, L., Saberioon, M., & Vašát, R. (2013). Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. *Applied Spectroscopy, 67,* 1349–1362.

Gomez, C., Lagacherie, P., & Coulouma, G. (2008). Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma, 148,* 141–148.

Hartmann, T. E., Wollmann, I., You, Y., & Müller, T. (2019). Sensitivity of three phosphate extraction methods to the application of phosphate species differing in immediate plant availability. *Agronomy, 9,* 29.

Heggemann, T., Welp, G., Amelung, W., Angst, G., Franz, S. O., Koszinski, S., et al. (2017). Proximal gamma-ray spectrometry for site-independent in situ prediction of soil texture on ten heterogeneous fields in Germany using support vector machines. *Soil & Tillage Research, 168,* 99–109.

Ji, W., Adamchuk, V. I., Biswas, A., Dhawale, N. M., Sudarsan, B., Zhang, Y., et al. (2016). Assessment of soil properties in situ using a prototype portable MIR spectrometer in two agricultural fields. *Biosystems Engineering, 152,* 14–27.

Jordan-Meille, L., Rubæk, G. H., Ehlert, P. A. I., Genot, V., Hofman, G., Goulding, K., et al. (2012). An overview of fertiliser-P recommendations in Europe: Soil testing, calibration and fertiliser recommendations. *Soil Use and Management, 28,* 419–435.

Kalbitz, K., Solinger, S., Park, J.-H., Michalzik, B., & Matzner, E. (2000). Controls on the dynamics of dissolved organic matter in soils: A review. *Soil Science, 165,* 277–304.

Kruse, J., Abraham, M., Amelung, W., Baum, C., Bol, R., Kühn, O., et al. (2015). Innovative methods in soil phosphorus research: A review. *Journal of Plant Nutrition and Soil Science, 178,* 43–88.

Kuang, B., Mahmood, H. S., Quraishi, M. Z., Hoogmoed, W. B., Mouazen, A. M., & van Henten, E. J. (2012). Sensing soil properties in the laboratory, in situ, and on-line: A review. *Advances in Agronomy, 114,* 155–223.

Lozier, T. M., Macrae, M. L., Brunke, R., & Van Eerd, L. L. (2017). Release of phosphorus from crop residue and cover crops over the non-growing season in a cool temperate region. *Agricultural Water Management, 189,* 39–51.

LWK (Chamber of Agriculture of North Rhine-Westphalia). (2015). Düngung mit phosphat, kali, magnesium (Fertilisation with phosphate, potassium, magnesium). phosphat-kalium-magnesium-pdf.pdf. Retrieved September 30, 2019 from http://www.landwirtschaftskammer.de.

Mayrink, G. O., Valente, D. S. M., Queiroz, D. M., Pinto, F. A. C., & Teofilo, R. F. (2019). Determination of chemical soil properties using diffuse reflectance and ion-exchange resins. *Precision Agriculture, 20,* 541–561.

Mertens, F. M., Pätzold, S., & Welp, G. (2008). Spatial heterogeneity of soil properties and its mapping with apparent electrical conductivity. *Journal of Plant Nutrition and Soil Science, 171,* 146–154.

Mouazen, A. M., & Kuang, B. (2016). On-line visible and near infrared spectroscopy for in-field phosphorous management. *Soil & Tillage Research, 155,* 471–477.

Mouazen, A. M., Maleki, M. R., De Bardemaeker, J., & Ramon, H. (2007). On-line measurement of some selected soil properties using a VIS–NIR sensor. *Soil & Tillage Research, 93,* 13–27.

Nagler, P. L., Daughtry, C. S. T., & Goward, S. N. (2000). Plant litter and soil reflectance. *Remote Sensing of Environment, 71,* 207–215.

Nawar, S., & Mouazen, A. M. (2019). On-line Vis–NIR spectroscopy prediction of soil organic carbon using machine learning. *Soil & Tillage Research, 190,* 120–127.

Patzold, S., Mertens, F. M., Bornemann, L., Koleczek, B., Franke, J., Feilhauer, H., et al. (2008). Soil heterogeneity at the field scale: a challenge for precision crop protection. *Precision Agriculture, 9,* 367–390.

Rodionov, A., Pätzold, S., Welp, G., Cañada Pallares, R., Damerow, L., & Amelung, W. (2014a). Sensing of soil organic carbon using visible and near-infrared spectroscopy at variable moisture and surface roughness. *Soil Science Society of America Journal, 78,* 949–957.

Rodionov, A., Pätzold, S., Welp, G., Pude, R., & Amelung, W. (2016). Proximal field Vis–NIR spectroscopy of soil organic carbon: A solution to clear obstacles related to vegetation and straw cover. *Soil & Tillage Research, 163,* 89–98.

Rodionov, A., Welp, G., Damerow, L., Berg, T., Amelung, W., & Pätzold, S. (2014b). Towards on-the-go field assessment of soil organic carbon using Vis–NIR diffuse reflectance spectroscopy: Developing and testing a novel tractor-driven measuring chamber. *Soil & Tillage Research, 145,* 93–102.

Rogovska, N., Laird, D. A., Chiou, C.-P., & Bond, L. J. (2019). Development of field mobile nitrate sensor technology to facilitate precision fertilizer management. *Precision Agriculture, 20,* 40–55.

Schüller, H. (1969). Die CAL-methode, eine neue methode zur bestimmung des pflanzenverfügbaren phosphates in Böden (The CAL method, a new method for the determination of plant available phosphate in soils). *Zeitschrift Pflanzenernährung Bodenkunde, 123,* 48–63.

Schulte-Ostermann, S. & Wagner, P. (2018). Variable-rate-fertilization of phosphorus and lime: Economic effects and maximum allowed costs for smallscale soil analysis. In *Proceedings of the 14th International Conference on Precision Agriculture* (unpaginated, online). Monticello, IL, USA: International Society of Precision Agriculture. Retrieved October 10, 2019 from https://www.ispag.org/proceedings/?action=abstract&id=5354.

Soriano-Disla, J. M., Janik, L. J., Allen, D. J., & McLaughlin, M. J. (2017). Evaluation of the performance of portable visible-infrared instruments for the prediction of soil properties. *Biosystems Engineering, 161,* 24–36.

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., & 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*(2), 139–186.

Steffens, D., Leppin, T., Luschin-Ebengreuth, N., Yang, Z. M., & Schubert, S. (2010). Organic soil phosphorus considerably contributes to plant nutrition but is neglected by routine soil-testing methods. *Journal of Plant Nutrition and Soil Science, 173,* 765–771.

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., & Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. *Advances in Agronomy, 107,* 163–215.

Stumpe, B., Weihermüller, L., & Marschner, B. (2011). Sample preparation and selection for qualitative and quantitative analyses of soil organic carbon with mid-infrared reflectance spectroscopy. *European Journal of Soil Science, 62,* 849–862.

Sun, Y., Druecker, H., Hartung, E., Hueging, H., Cheng, Q., Zeng, Q., et al. (2011). Map-based investigation of soil physical conditions and crop yield using diverse sensor techniques. *Soil & Tillage Research, 112,* 149–158.

Sylvester-Bradley, R., Lord, E., Sparkes, D. L., Scott, R. K., Wiltshire, J. J. J., & Orson, J. (1999). An analysis of the potential of precision farming in Northern Europe. *Soil Use and Management, 15,* 1–8.

Viscarra Rossel, R. A. (2008). ParLeS: Software for chemometric analysis of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems, 90,* 72–83.

Viscarra Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma, 158,* 46–54.

Wijewardane, N. K., Ge, Y., Wills, S., & Libohova, Z. (2018). Predicting physical and chemical properties of US soils with a mid-infrared reflectance spectral library. *Soil Science Society of America Journal, 82,* 722–731.

Wuenscher, R., Unterfrauner, H., Peticzka, R., & Zehetner, F. (2015). A comparison of 14 soil phosphorus extraction methods applied to 50 agricultural soils from Central Europe. *Plant, Soil and Environment, 61*(2), 86–96.

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

Affiliations

Stefan Pätzold[1] [ORCID] **· Matthias Leenen[1] · Peter Frizen[1,2] · Tobias Heggemann[1,2] · Peter Wagner[3] · Andrei Rodionov[1,4]**

✉ Stefan Pätzold
    s.paetzold@uni-bonn.de

[1] Institute of Crop Science and Resource Conservation (INRES) – Soil Science and Soil Ecology, University of Bonn, Nussallee 13, 53115 Bonn, Germany

[2] Present Address: Chamber of Agriculture North Rhine-Westphalia, Nevinghoff 40, 48147 Münster, Germany

[3] Institute of Agricultural and Nutritional Sciences, Farm Management Group, Martin Luther University Halle-Wittenberg, Karl-Freiherr-von-Fritsch-Straße 4, 06099 Halle (Saale), Germany

[4] Present Address: Department Soil Ecology, University of Bayreuth, Dr.-Hans-Frisch-Straße 1–3, 95448 Bayreuth, Germany