RESEARCH ARTICLE

Determining soil particle-size distribution from infrared spectra using machine learning predictions: Methodology and modeling

Elizabeth Jeanne Parent [ID] *, **Serge-Étienne Parent** [ID], **Léon Etienne Parent**

Department of Soils and Agrifood Engineering, Université Laval, Québec, Canada

* elizabeth.parent.1@ulaval.ca

Abstract

Accuracy of infrared (IR) models to measure soil particle-size distribution (PSD) depends on soil preparation, methodology (sedimentation, laser), settling times and relevant soil features. Compositional soil data may require log ratio (*ilr*) transformation to avoid numerical biases. Machine learning can relate numerous independent variables that may impact on NIR spectra to assess particle-size distribution. Our objective was to reach high IRS prediction accuracy across a large range of PSD methods and soil properties. A total of 1298 soil samples from eastern Canada were IR-scanned. Spectra were processed by Stochastic Gradient Boosting (SGB) to predict sand, silt, clay and carbon. Slope and intercept of the log-log relationships between settling time and suspension density function (SDF) (R^2 = 0.84–0.92) performed similarly to NIR spectra using either *ilr*-transformed (R^2 = 0.81–0.93) or raw percentages (R^2 = 0.76–0.94). Settling times of 0.67-min and 2-h were the most accurate for NIR predictions (R^2 = 0.49–0.79). The NIR prediction of sand sieving method (R^2 = 0.66) was more accurate than sedimentation method(R^2 = 0.53). The NIR 2X gain was less accurate (R^2 = 0.69–0.92) than 4X (R^2 = 0.87–0.95). The MIR (R^2 = 0.45–0.80) performed better than NIR (R^2 = 0.40–0.71) spectra. Adding soil carbon, reconstituted bulk density, pH, red-green-blue color, oxalate and Mehlich3 extracts returned R^2 value of 0.86–0.91 for texture prediction. In addition to slope and intercept of the SDF, 4X gain, method and pre-treatment classes, soil carbon and color appeared to be promising features for routine SGB-processed NIR particle-size analysis. Machine learning methods support cost-effective soil texture NIR analysis.

Introduction

Soil particle-size distribution (PSD) is of prime importance for plant growth and soil management [1]. Mimicking particle sedimentation in natural water bodies, PSD has been traditionally quantified using the sieve-pipette method that determines particle mass, and the sieve-hydrometer or sieve-plummet balance method that measures changes in suspension density [2]. Sedimentation techniques are thought to overestimate the concentration of plate-like clay particles that do not fit into Stokes' law [3]. Reynolds' number should be less than 0.05,

otherwise, the drag force for sedimentation increases faster than predicted by Stokes' law [3]. Because organic matter binds soil particles [4], organic matter is often destroyed using a peroxide pre-treatment as method modifier to disperse soil particles. Laser techniques are much faster than sedimentation techniques but tend to underestimate clay-size particles due to insufficient particle dispersion despite sonication pre-treatment, and this may require using conversion equations [5,6].

On the other hand, visible and near infrared (VIS–NIR: 350–2500 nm) spectroscopy is an efficient soil quality and fertility screening tool [7] because spectra correlate well with several chemical, physical and mineralogical properties [8]. The VIS represents the visible light range between 350 and 780 nm, from violet to red. Because most routine soil and plant laboratories are equipped with NIR spectrometers for forage analysis, they may contribute to documenting soil characteristics at low cost. Where VIS is not available, soil RGB (red-green-blue) can be assessed from the Munsell color chart using computer models ("munsell2rgb" from the "aqp" package). Mid-infrared (MIR) is generally more accurate than NIR [9] but requires more sample preparation, limiting its application as routine determination method.

Infrared spectroscopy (IRS) more accurately predicts clay than sand and silt contents [10] because the IR spectrum is sensitive to clay mineralogy [10–12] and total reflectance decreases as grain size increases [13,14]. Light absorption is also influenced by soil features such as particle roundness [15–18], soil pH, and the Mehlich-3 soil test for Ca, Mg and Mn [19]. The IRS detects Al-OH (2200 nm) and Fe-OH (2290 nm) [20,21] that in turn have an impact on soil structure [22] and IR reflectance [23]. The VIS-NIR spectra are sensitive to soil moisture and C content [24]. Organic matter, multi-nutrient extraction and reconstituted bulk density from scooped soil samples are also common features quantified in routine laboratories.

The sedimentation methods provide percentages of sand, silt, and clay from the log-log relationship between settling time and suspension density. The slope and intercept return proportions of sand, silt and clay at pre-selected settling times that may vary between laboratories [1], thus affecting the accuracy of IRS models. Providing more flexibility using the slope and intercept of the log-log relationship and reducing the arbitrariness of the settling time selection may allow increased reliability of IRS models calibrated against sedimentation methods.

There are unattended sources of error in IRS calibration. There is systematic negative covariance between sand, silt and clay fractions due to resonance within the ternary diagram [25]. Indeed, there are D-1 degrees of freedom in a D-part composition [26]. Not considering the problem of closure to 100% in statistical analysis, confidence intervals about means of proportions may take values outside of the compositional space, i.e. < 0 or > 100% [27], and the measures of distance and dissimilarity are non-Euclidian [28]. To return unbiased statistical results, orthonormal balances among subsets of components can be computed as D-1 isometric log ratios (*ilr*) [29]. The back-transformed *ilr* values allow recovering proportions of sand, silt, and clay totalling exactly 100% within the limits of the ternary diagram.

For both the sedimentation and laser techniques there are several soil pre-treatments (peroxide, sodium hypochlorite, sodium hexametaphosphate, sonication intensity), soil features, calibration techniques, options (NIR 2X, NIR 4X, settling times or suspension density function for sedimentation; pump and stirrer spin, refractive index of the medium, real or imaginary refractive index, density for laser) and expressions (percentages, ratios) that influence results of particle-size distribution. Machine learning (ML) is an emerging data mining technique of artificial intelligence that can unravel patterns and rules in large data sets [30] and predict target variable from input data [31]. Machine learning methods can account for numerous independent variables that may impact on NIR spectra to assess accurately soil particle-size distribution.

This paper is presented in two parts, one focusing on methodology and the other on modeling. In the methodology section, we hypothesized that methods and pre-treatments return different results and thus cannot be combined to run a single IRS calibration set without input information on methodologies. In the modeling section, we hypothesized that: 1. IRS cannot accurately predict PSD from laser methods due to underestimation of the clay fraction; 2. IRS accurately predicts sand fraction determined using sand sieving methods; 3. IRS more accurately predicts PSD from sedimentation methods after accounting for soil features; 4. IRS prediction accuracy for PSD is improved using the slope and intercept of the log-log relationship and isometric log ratios compared to raw percentages and pre-determined settling times; and 5. MIR prediction accuracy for PSD is higher compared to the NIR spectra. The objective of this study was to calibrate infrared determination methods against routine PSD methods, methodological modifiers and soil features to reach high IR-ML model accuracy for routine soil texture determination.

Materials and methods

Characterization of soils

The data set of 1298 soil samples collected in the arable layer (0–20 cm) was obtained from several research institutions in Québec, Canada (Fig 1). The main crops were maize (*Zea mays*) cereals and forages on coarse- to fine-textured soils, potatoes (*Solanum tuberosum*) on sandy loams and loamy sands, and cranberries (*Vaccinium macrocarpon*) on sandy soils. The soils were mainly Inceptisols and Spodosols.

Soil samples were air-dried then passed through a 2-mm sieve [32,33]. The 2-mm sieved soil was 3-mL scooped and then weighed to determine the reconstituted bulk density as performed routinely in soil testing laboratories. Soil color was assessed on dry samples using the Munsell chart and then transformed into RGB percentages using "munsell2rgb" in R. The
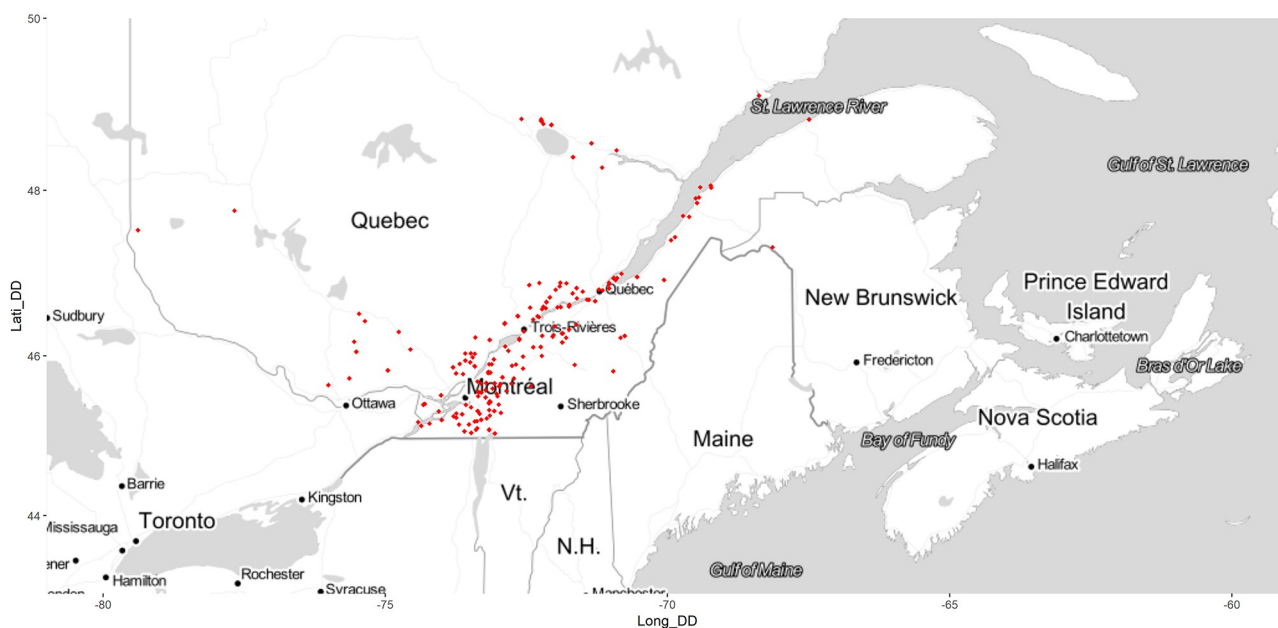

Fig 1. Repartition of soil sampling areas on eastern provinces in Canada. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL. Contains information from OpenStreetMap and OpenStreetMap Foundation, which is made available under the Open Database License. http://maps.stamen.com/#toner/12/37.7706/-122.3782.

https://doi.org/10.1371/journal.pone.0233242.g001

total carbon (Ct) was quantified using the Leco CNS analyzer (Leco Corporation, St. Joseph, Michigan). For pH determination, 10 g of soil was mixed with 20 mL 0.01M $CaCl_2$. For oxalate extracts, a 0.5 g sample of soil was mixed with 20 mL of oxalate solution (0.2 M of ammonium oxalate and 0.2 M of oxalate acid) and agitated for 4 h in the dark [34]. The mixture was centrifuged at 2000 rpm for 5 min and then filtered through Whatman no. 40 paper. Concentrations of P, Fe, Ca, Al, Mn and Si were quantified by ICP-OES. Soils were also extracted using the routine Mehlich3 method [35]. Concentrations of P, Ca, Mg, Fe, Al, Mn, Zn and Cu were quantified by ICP-OES.

Particle-size analysis

Particle size distribution was analyzed in 50–100 g samples using the sedimentation method [36]. A separate batch of soils was pre-treated with peroxide for comparison with the no-peroxide pre-treatment. Samples were mixed with 0.05 M hexametaphosphate and agitated at 300 rpm for 16 h. The mixture was transferred to a 1 L cylinder and hand-shaken for 30 sec. Suspension density readings (g L^{-1}) taken after 0.75-, 5-, 120-, 420-, and 1440- min were referred to as the sedimentation multi-point method. Samples reporting suspension density after 0.67 and 120 min as originally suggested [37] were assigned to the 2-point 2-h sedimentation method. The clay fraction was also recorded after 7-h settling time, close to the 6-h settling time used by Gee and Bauder (1979) [38]. Although the clay fraction was overestimated as compared to the 7-h sedimentation multi-point method, the 2-point 2-h sedimentation method was selected as our reference because it has been widely used as a proximate method in soil surveys combined with tactile assessment. After taking the last reading, the entire hydrometer contents from sedimentation method were passed through 1-, 0.5-, 0.25-, 0.10- and 0.05- mm sieves under tap water to clean the coarser particles of any adhering finer particles and to determine the sand-size distribution. The suspension density was transformed into particle-size percentages by mass using standard equations [37,38]. The sedimentation curve relating suspension density to settling time was log-transformed (ln) to determine the slope and intercept as model parameters.

Samples were also analyzed using the Mastersizer 2000 Laser particle size analyzer (Malvern Instruments, Worcestershire, UK, measurements at 633 nm and 466 nm) combined with Hydro 2000G (800-mL tap water volume, 500-rpm stirrer and 2000-rpm pump) with or without ultrasonic action at nominal 40 kHz frequency for 2 min [39]. Only samples within 10 to 20% of obscuration were retained in the database. The refractive index of the medium was set at 1.5.

Calculation

The three particle-size fractions and carbon content were isometric log ratio (ilr) transformed as follows [29]:

$$ilr_{[carbon|clay.silt.sand]} = \sqrt{\frac{2}{3}}ln\left(\frac{\sqrt{clay \times silt \times sand}}{carbon}\right) \tag{1}$$

$$ilr_{[clay|silt.sand]} = \sqrt{\frac{2}{3}}ln\left(\frac{\sqrt{silt \times sand}}{clay}\right) \tag{2}$$

$$ilr_{[silt|sand]} = \sqrt{\frac{1}{2}}ln\left(\frac{silt}{sand}\right) \tag{3}$$

Spectral data acquisition

The air-dried and 2-mm sieved samples [40] were placed into a 5-cm quartz cup then scanned using a Nicolet Antaris FT-NIR analyzer (Thermo Electron Corp., Ann Arbor, Michigan). Absorbance was measured with gains of 2X or 4X. Triplicated spectra were scanned 30 times in the range of 9090 to 4000 cm^{-1} (1100 to 2500 nm) at a resolution of 2 cm^{-1} (0.3 nm at 1250 nm) [41]. A library of 3069 NIR spectra with 2X gain and 668 spectra with 4X gain were obtained after probabilistic quotient normalization and first derivation [42]. All samples were stored in white plastic storage containers.

The soil and KBr samples were ground to less than 74 μm using a mortar and pestle [43], then oven-dried at 105°C for 3 h [44]. A 2 g sample was weighed and mixed with 200 g KBr powder. The mixture was pressed at 517 MPa for 3 min using a manual hydraulic press (Carver, Model 4350.L, Carver Inc., Wabash, Indiana) to yield clear 13-mm-diameter pellets. The mid-infrared spectra were scanned using a transmission DTGS detector Varian 1000 FT-IR Scimitar series spectrometer (Varian Inc., Palo Alto, California). Spectra were in the range of 4000 to 400 cm^{-1} (2500 to 25 000 nm) at 4 cm^{-1} (0.6 nm) resolution. Each pellet was scanned 10 times and rotated once manually at 90°. A library of 413 MIR spectra was obtained.

Spectral data modeling

Statistical analyses were conducted using the R-3.6.1 version [45] with the packages "tidy-verse", "reshape2", "stringi", "signal", "mvoutlier", "caret", "compositions", "soiltexture", "aqp", "patchwork", "broom", "ggExtra", "ggmap", "neuralnet" and "outliers". The machine learning (ML) method was GBM (Stochastic Gradient Boosting) that outperformed other ML models including neural networks. Neural networks was tested with a hidden of 3 and an activation function settled at "logistic" (S2 Table). The NIR and MIR spectra were normalized and passed through a binning process with 10 cut points and a Savitsky-Golay smoothing filter. After first derivation, spectra compressed into scores using principle component analysis (PCA) [46]. Because of the diversity of the soil samples, we did not remove outliers. Sand, silt and clay percentages, oxalate and Mehlich-3 extracts composed of 0.2N CH3COOH-0.25N NH4N03-0.015NNH4F-0.013NHN03-0.001M EDTA were *ilr*-transformed. Every set was divided into training (70%) and testing (30%).

Results

Soil characterization

There was a large spectrum of soil properties (Table 1). In comparison, soils of the region have been reported to range from 5 to 137 g organic matter kg^{-1}, 560 to 19 992 mg Fe$_{oxalate}$ kg^{-1} and 270 to 36 638 mg Al$_{oxalate}$ kg^{-1}, and 4–494 g sand kg^{-1}, 40–734 g silt kg^{-1}, and 6–796 g clay kg^{-1} [47]. Particle-size distribution is presented in Fig 2. Orthonormal balances were normally distributed except for [clay | silt,sand] using the laser method (Fig 3). Indeed, the clay content was systematically underestimated, and the silt content systematically overestimated by the laser method.

Methodologies

Paired t-tests comparing methods are presented in Table 2. Comparisons involving the peroxide pre-treatment must be interpreted with care due to the smaller number of observations (26 to 38). Mean *ilr* differences between the reference 2-point 2-h sedimentation method and other methods are illustrated in Fig 4. The difference between the laser and the 2-point 2-h

Table 1. Ranges of soil properties (0–20 cm) in the data set (particle-size distribution according to the multi-point 7-h sedimentation method using 45-sec settling time for sand and 7-h settling time for clay).

Site	Mean	Standard deviation	Minimum	Maximum
pH (0.01 M CaCl$_2$)	5.36	0.67	3.37	7.90
		g kg^{-1}		
Sand total	613	24	1	986
• **1–2 mm**	41	50	0	282
• **0.5–1.0 mm**	108	115	1	541
• **0.25–0.5 mm**	191	126	5	580
• **0.1–0.25 mm**	235	154	3	766
• **0.25–0.01 mm**	118	103	0	587
Silt	256	15	5	816
Clay	131	13	5	839
Total C	25.5	2.4	0.1	443
Total N	1.9	1.6	0.7	19
Total S	0.3	0.4	0	10.3
		mg kg^{-1}		
Fe$_{oxalate}$	5599	2449	447	10400
Al$_{oxalate}$	4000	2693	528	15346
Mn$_{oxalate}$	309	304	7	2287
Ca$_{oxalate}$	30	14	3	153
P$_{oxalate}$	803	464	68	2266
Si$_{oxalate}$	829	578	88	5585
P$_{Mehlich3}$	118	62	6	362
Ca$_{Mehlich3}$	1776	1437	113	5103
Mg$_{Mehlich3}$	191	224	5	1572
Fe$_{Mehlich3}$	232	108	72	715
Al$_{Mehlich3}$	1265	431	310	2241
Mn$_{Mehlich3}$	25	20	5	206
Zn$_{Mehlich3}$	8	9	0	130
Cu$_{Mehlich3}$	4	3	0.4	30
		%		
Green	47	8	28	68
Blue	34	10	11	58
Red	54	7	31	75
		g cm^{-3}		
Reconstituted bulk density	1.06	0.16	0.71	1.46

sedimentation methods depended on the *ilr* coordinate (also shown in Fig 3). After 2-min of sonication, the [silt | sand] balance differed slightly between methods while the [clay | silt, sand] balance differed markedly between methods, indicating insufficient dispersion of clay-size particles.

As expected, the multi-point 7-h sedimentation method returned a smaller proportion of clay-size particles compared to the 2-point 2-h sedimentation method due to longer settling time. The peroxide pre-treatment for the multi-point 7-h sedimentation method tended to increase clay and silt fractions. The [silt | sand] balance resulting from the no-peroxide pre-treatment preceding sedimentation methods was comparable across methods, with the exceptions of the laser and pretreated multi-point 7-h sedimentation methods. Excluding the 0-min sonicated laser and comparable methods, the amount of sand was slightly lower compared to

Texture triangle: Canada (CA)
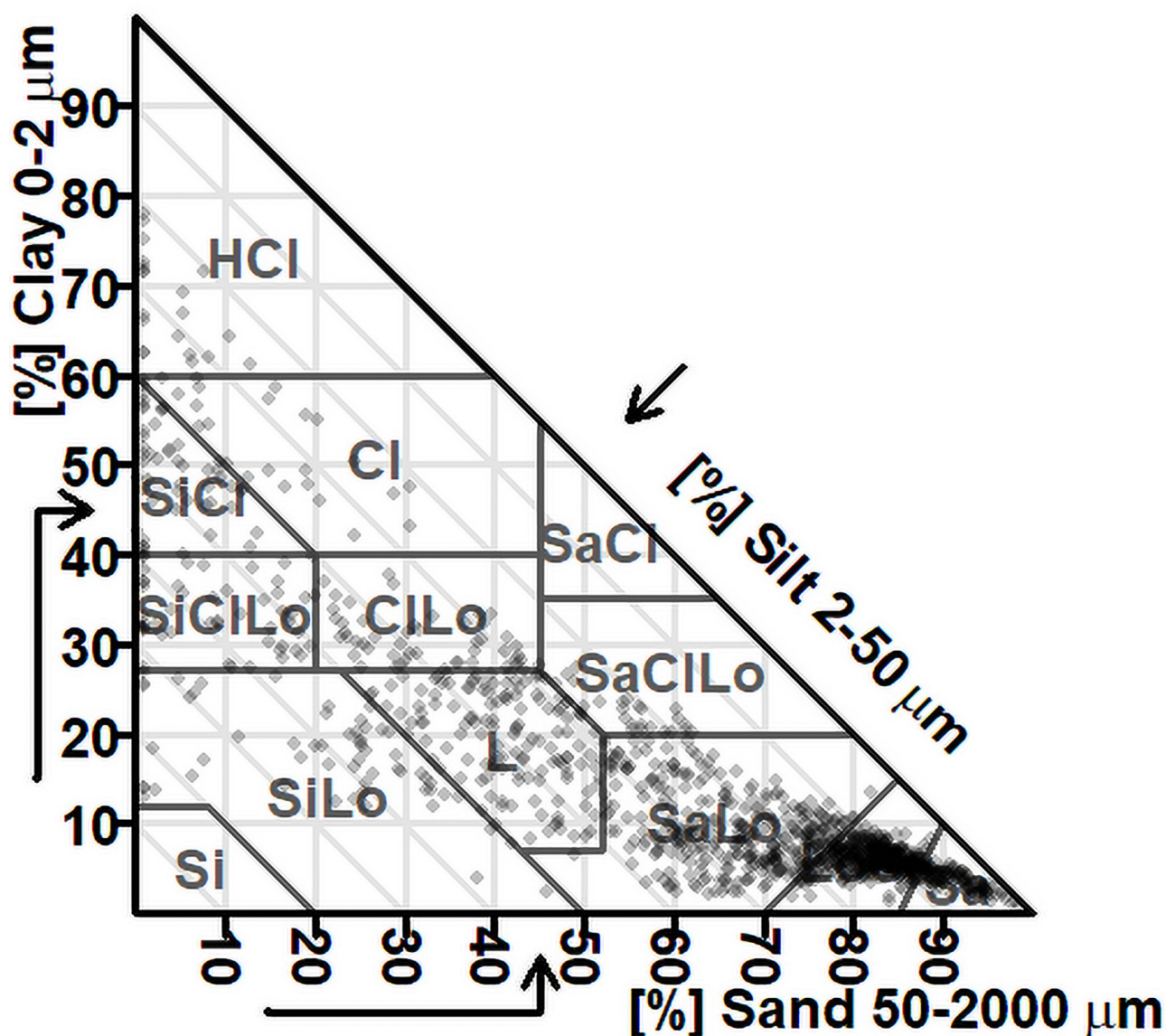
Fig 2. Particle size distribution of studied soils in the Canadian textural diagram.

sand sieving. Sand contents did not differ significantly between the peroxide pre-treatments or multi-points calculation without pre-treatment. Except for the peroxided pre-treated dataset, the [carbon | clay,silt,sand] balances were differentially influenced by methodologies.

Spectral data modeling

To conduct predictions to details of methodologies and features, the data set was divided into eight subsets (Table 3). Prediction models were run for each subset. In general, prediction accuracies were weakest for the silt fraction and were similar whether data were crude or *ilr-*

Fig 3. Distribution of *ilr* textural variables across methodologies.

https://doi.org/10.1371/journal.pone.0233242.g003

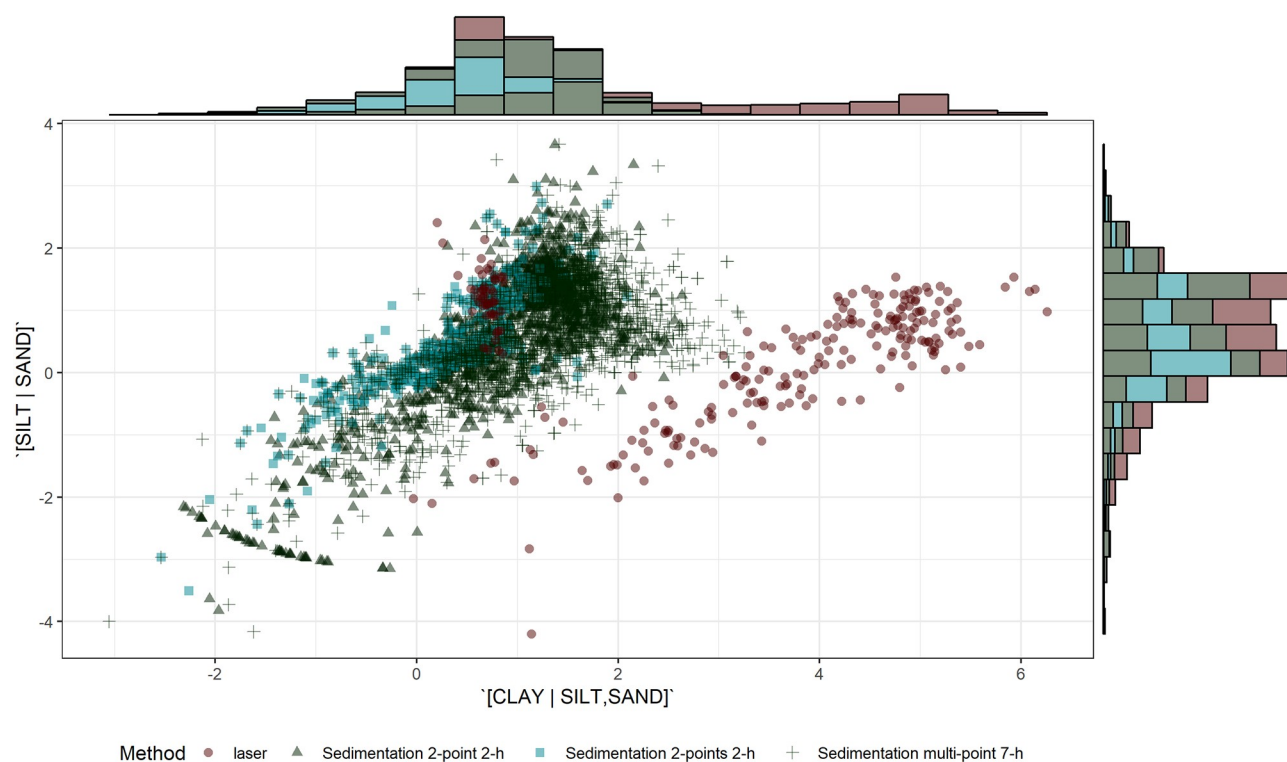**Table 2. Comparison of methods (method1 minus method2) using paired t-test and confidence intervals (p ≤ 0.05).**

Target variable	Method1	Method2	p.value	N
**[Carbon	Clay,Silt,Sand]**	**Sedimentation No peroxide**	**Sedimentation peroxide**	ns	38
**[Carbon	Clay,Silt,Sand]**	**Sedimentation No peroxide**	**Laser 0 min**	**	46
**[Carbon	Clay,Silt,Sand]**	**Sedimentation No peroxide**	**Laser 2 min**	**	106
**[Carbon	Clay,Silt,Sand]**	**Sedimentation 2-pointpoint No peroxide**	**Sedimentationmulti-pointpoint No peroxide**	**	763
**[Carbon	Clay,Silt,Sand]**	**Sedimentation 2-pointpoint peroxide**	**Sedimentation multi-points peroxide**	ns	227
[Clay	Silt,Sand]	**Sedimentation No peroxide**	**Sedimentation peroxide**	**	38
[Clay	Silt,Sand]	**Sedimentation No peroxide**	**Laser 0 min**	**	46
[Clay	Silt,Sand]	**Sedimentation No peroxide**	**Laser 2 min**	**	106
[Clay	Silt,Sand]	**Sedimentation 2-point No peroxide**	**Sedimentation multi-point No peroxide**	**	763
[Clay| Silt,Sand]	**Sedimentation2-point peroxide**	**Sedimentation multi-point peroxide**	**	227
**[Silt	Sand]**	**Sedimentation No peroxide**	**Sedimentation peroxide**	ns	38
**[Silt	Sand]**	**Sedimentation No peroxide**	**Laser 0 min**		46
[Silt| Sand]	**Sedimentation No peroxide**	**Laser 2 min**	*	106
**[Silt	Sand]**	**Sedimentation 2-point No peroxide**	**Sedimentation multi-point No peroxide**	**	763
**[Silt	Sand]**	**Sedimentation 2-point peroxide**	**Sedimentation multi-point peroxide**	**	227
Sand	**Sedimentation No peroxide**	**Sieving**	*	746
Sand	**Sedimentation peroxide**	**Sieving**	ns	26
Sand	**Laser 0 min**	**Sieving**	*8	34
Sand	**Laser 2 min**	**Sieving**	*	77
Sand	**Sedimentation multi-point No peroxide**	**Sieving**	ns	358
Sand	Sedimentation **multi-point peroxide**	**Sieving**	*	84

ns, *,**: Non-significant and significant at the 0.05 and 0.01 levels, respectively.

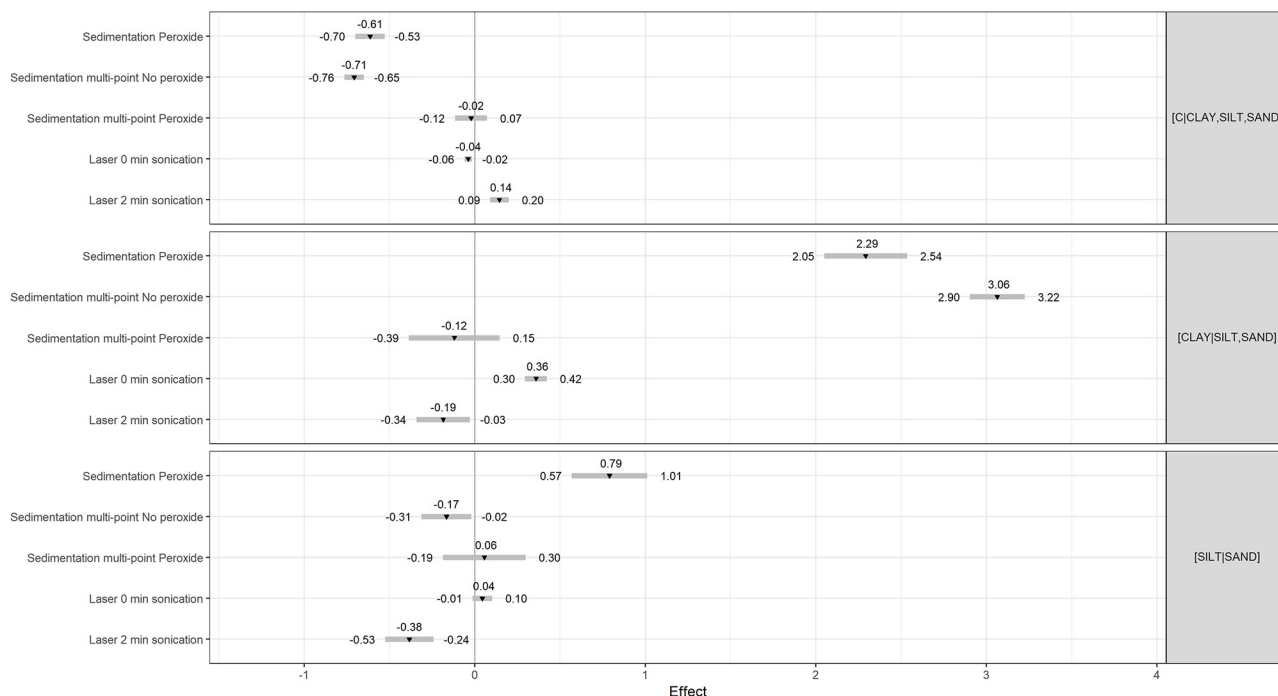N: Sample size.

https://doi.org/10.1371/journal.pone.0233242.t002

Fig 4. Mean differences (p ≤ 0.025) in textural balances between several methodologies against the reference 2-point 2-h sedimentation method without pre-treatment.

transformed. As expected, the clay fraction determined by laser in Set1 was poorly related to the NIR spectra (R^2 = 0.45–0.64) whether the PSD was expressed as *ilr* or % (Table 3). Combining laser and sedimentation methods in Set2 improved the clay predictions with R^2 values of 0.85–0.91. In Set3, PSD predictions were more accurate with the 2-point 2-h sedimentation method (R^2 = 0.49–0.79) than with the multi-point sedimentation method (R^2 = 0.44–0.70). Modeling Set4 showed that NIR was less accurate (R^2 = 0.40–0.71) compared to MIR (R^2 = 0.45–0.80). Accuracy was higher with NIR-4X (R^2 = 0.87–0.95) than with NIR-2X (R^2 = 0.69–0.92) (Set5). Models for Set6 showed that data expressed as *ilr* or % (R^2 = 0.76–0.94) for interpolated PSD were similarly accurate to the slope and intercept from the relationship between settling time and suspension density (R^2 = 0.84–0.92). Several features in Set7 improved model accuracy (R^2 = 0.86–0.91) compared to no features at all (R^2 = 0.82–0.97). The total carbon content contributed the most to the increase in accuracy, followed by colors and oxalate extracts. Finally, the carbon content showed higher performance with raw data (R^2 = 0.63–0.89) than with *ilr*-transformed data (R^2 = 0.28–0.84). Compared to a model without features (R^2 = 0.97), the reconstituted bulk density (R^2 = 0.99), color (R^2 = 0.96), oxalate (R^2 = 0.98) and Mehlich-3 (R^2 = 0.98) extracts were similar to the accuracy of the NIR model. For the sand fraction, the NIR model calibrated against the sand-sieving method showed higher accuracy (R^2 = 0.66) compared to the sedimentation 2-point 2-h method (R^2 = 0.53).

Discussion

Methodologies. The IR technologies are generally calibrated against a single methodology. In this paper, several methodologies were IR-tested on comparable soil samples. The present paper presented IR-SGB results across several methodologies and features that are rarely addressed altogether in the literature. Machine learning integrated the available

Table 3. Accuracy, slope, intercept, and number of observations of models.

Set	Dependent Variables	Independent variables	Laboratory method for calibration	Sand	Silt	Clay	C	Sand	Silt	Clay	C	Value interpretation*	N
				Adjusted R²				RMSE					
Set1	*Ilr*	2X	Laser	0.80	0.79	0.45	0.28	10.80	10.64	1.32	0.67	-	92
	%			0.86	0.80	0.64	0.63	9.11	10.35	1.06	0.48	-	
Set2	*Ilr*	Lab method, PT, NIR-2X	2-point 2-h sedimentation + Laser	0.90	0.85	0.80	0.58	9.05	6.94	5.27	0.72	+	485
	%			0.89	0.91	0.85	0.71	9.30	7.17	4.21	0.62	++	
Set3	*Ilr*	NIR-2X	Multi-point 7-h sedimentation	0.63	0.44	0.70	0.40	16.17	13.22	6.94	0.92	-	667
			2-point 2-h sedimenttion	0.76	0.49	0.79	0.36	13.42	11.50	6.60	0.95	-	
Set4	*Ilr*	MIR	2-point 2-h sedimentation	0.75	0.45	0.80	0.21	14.75	11.34	7.88	0.77	-	222
		NIR-2X		0.71	0.40	0.69	0.02	15.78	11.81	9.75	0.86	-	
Set5	*Ilr*	NIR-4X	2-point 2-h sedimentation	0.94	0.87	0.95	0.85	7.26	6.15	3.62	0.48	++	311
		NIR-2X		0.83	0.69	0.92	0.76	12.42	9.40	4.78	0.62	+	
Set6	*Ilr*	PT, NIR-2X	2-point 2-h sedimentation	0.90	0.81	0.93	0.84	8.75	6.98	3.94	0.73	+	860
	%			0.91	0.76	0.94	0.89	8.40	7.74	3.84	0.61	+	
	Slope, Intercept			0.90	0.84	0.92	-	8.63	6.34	4.29	-	++	
Set7	*Ilr*	All features, NIR-2X	2-point 2-h sedimentation	0.91	0.86	0.90	0.69	9.27	5.86	6.06	0.36	++	156
		No feature, NIR-2X		0.94	0.82	0.97	0.97	7.78	7.67	3.07	0.71	++	
		Carbon, NIR-2X		0.96	0.88	0.97	0.95	6.59	6.34	3.36	0.97	++	
		Bulk density, NIR-2X		0.89	0.63	0.94	0.99	10.59	10.84	4.66	0.52	-	
		pH, NIR-2X		0.92	0.78	0.96	0.94	8.84	8.39	3.53	1.06	+	
		Color, NIR-2X		0.94	0.86	0.95	0.96	7.55	6.64	4.10	0.84	++	
		Oxalate, NIR-2X		0.95	0.84	0.96	0.98	7.19	7.17	3.71	0.54	++	
		Mehlich3, NIR-2X		0.89	0.70	0.93	0.98	10.62	9.83	4.71	0.64	+	
Set8	%	NIR-2X	Sand sieving	0.66	-	-	-	12.15	-	-	-	+	746
			2-point 2-h sedimentation	0.53	-	-	-	14.73	-	-	-	-	

C: Carbon; PT: Pre-treatments (no peroxide or peroxide); MIR: MIR scores; RMSE: Root mean square error; N: Sample size.

*: Value interpretation on sand, silt and clay contents.

-: Vulnerable estimations; +: Approximated estimations; ++: Research application estimations; +++: Quality control estimations [48,49].

information on pre-treatments, methods and features to predict soil texture at low cost in routine laboratories.

Sedimentation methods. The peroxide pre-treatment tended to increase the clay and silt fractions at the expense of the sand fraction, indicating that clay- and silt-size particles were freed from the clay- or silt-organic matter complexes in micro-aggregates [50]. While metals such as Fe are sequestered by soil organic matter [51] or associated with it [52], iron may be freed from organic matter using hydrogen peroxide to form reactive iron hydroxide [53] that may, conversely, positively impact soil aggregation [54]. Peroxide-treated soils may thus require additional dispersion prior to sedimentation [55]. Nevertheless, the 2-point 2-h sedimentationwith and without peroxide pre-treatments showed comparable results or minor differences for the [clay | silt,sand] and [silt | sand] balances.

Laser method. The Mie theory considers soil particles to be spherical (Malvern Instruments Ltd., 2007). Parameters such as the refractive index of the medium, the real or imaginary refractive index and the density could be included to improve accuracy. The 2-min sonicated laser method and the sand sieving method returned closed results. The 40 kHz sonication for a duration of 2 min appeared to be suitable for sandy soils [56]. Variants include 36 kHz for 3

min [57] and 30 kHz for 30 min [2]. Although fragile quartz grains may be broken using ultra-sonication [58], the [silt | sand] balance for the sedimentation method showed results close to the 2-min sonicated laser method.

According to Storti and Balsamo (2010) [59], the high-strength materials are not as affected by the procedures as the low-strength materials. Using water as the dispersing agent and 1750-, 700- rpm for the pump and stirrer speeds, Sochań et al. (2012) [60] obtained R^2 values of between 0.67 and 0.95 for high-strength materials in silt loamy to sandy soils, values close to our raw data results (R^2 = 0.64–0.86). For clayed soils, a surfactant or solvent may replace tap water [61]. As small particles increase in number and combine into sand-sized aggregates, pre-treatments may be required and adjusted to soil specificities. There is no standard procedure to disperse soil samples for the laser method because pre-treatment is soil specific.

Spectral data modeling

In this study, we used subsets $\geq$ 92 samples and GBM as the machine learning method for particle-size prediction compared to suggested subsets $\geq$ 130 samples for cLHS and FCMS [62]. Here, only Set1 for laser samples did not meet that criterion. The Boosted Regression Trees is another suitable ML method for complex predictions [41] and is relatively fast compared to random forest and neural network algorithms. Model accuracies in validation were in the range reported in the literature where R^2 values have been found to vary between 0.46 and 0.94 for clay [10,63,64] and between 0.53 and 0.82 for sand. More generally, the R^2 values may vary widely between 0.05 and 0.84 [65,66].

Prediction accuracy.   Because the 2-point 2-h sedimentation with and without pre-treatment showed comparable results, they could be combined for IRS calibration with method and treatment information inputs. Hence results of soil surveys could provide a large database for IRS calibration purposes.

Laser methods produced results with lower accuracy compared to the sedimentation methods, due to underestimation of the clay fraction. The [silt | sand] balance from the laser method, however, was close to that of the 2-point 2-h sedimentation method, despite higher silt and sand percentages for the laser method, indicating advantage for log-ratioing. Higher accuracies from the laser method were obtained by Blott and Pye (2006) [67] across a wide range of soils, sediments and powders. Zobeck (2004) [68] related results from a LS-230 laser diffraction particle size analyzer to those of the pipette method and obtained R^2 values of 0.97, 0.99, and 0.99 for the $<$ 2-, $<$ 50-, and $<$ 100 –μm in non-calcareous soils, using a shape factor of 0.2 compared to the default 1.0 shape factor for the laser method. Nevertheless, we found that the laser method did not produce results as consistent as the sedimentation method for clay predictions by IRS, indicating regional, soil-specific, calibration.

Mehlich-3 extracts, reconstituted bulk density and pH did not improve prediction accuracy. Soil features such as total carbon content, colors and amorphous materials (oxalate extracts) increased model prediction accuracy. Prediction accuracy of carbon content could perform with features as reconstituted bulk density as well as Mehlich-3, colors and oxalate extracts.

Using suspension density function parameters instead of arbitrary settling times did not increase the accuracy of PSD predictions but provided a uniform base to run NIR models as various settling times have direct impact on predictions. Then, in Set3, we observed that 0.67-min and 2-h settling times were more accurate than 0.75-min and 7-h periods for sand and clay NIR predictions.

In the present study, MIR spectra were more accurate than NIR for sand, silt and clay determinations. The NIR method can provide accurate prediction for clay as it has been also found to be accurate for cation exchange capacity (CEC) with R^2 values of 0.82 [41] and 0.81 [63]. On

the same direction, Viscarra Rossel et al. (2006) [9] concluded that MIR was more suitable than NIR for texture and carbon determination, due to higher incidence of spectral bands combined with higher intensity and specificity of the signal compared to NIR. To further support NIR calibration and model accuracy, the R-coded GBM machine learning model used in the present study came across several soil textural classes, carbon contents and features that have not been addressed simultaneously in past research.

Conclusions

In this paper, IRS inaccurately predicted the PSD's clay fraction using laser methods. However, IRS accurately predicted PSD against sedimentation and sieving methods after adding soil features such as color, total carbon content and concentration of amorphous materials related to soil genesis and classification. Soil pre-treatments and the need for dispersing agents could be adjusted to the nature and concentration of binding agents for silt-, clay-size particles and fine particles adhering to sand particles. Combined with method and treatment as features, post-screening total carbon content and color routinely determined in service laboratories can improve IRS accuracy for mineral soils. Features as reconstituted bulk density and Mehlich-3 extracts could be added as features for higher-C soils.

The GBM returned similar results whether particle-size data were analyzed as raw pre-determined settling times (percentages) or as *ilr*-transformed percentages. The GBM returned similar accuracies using the slope and intercept of the log-log relationship between settling time and suspension density, *ilr*-transformed percentages or raw percentages. The MIR and NIR 4X gain methods performed better than the NIR 2X gain method. However, additional features increased NIR 2X predictability. Modeling the log-log relationship between settling time and suspension density provided greater flexibility in the choice of soil-specific settling times. The GBM model proved to be a powerful tool to process the results of several analytical methods used worldwide to determine soil grain-size distribution [8]. More methodologies could be included in the future such as the pipette methods, and chemical and sonication methods to disperse microaggregate.

Supporting information

S1 Table. Slope, intercept, and number of observations of models.
(DOCX)

S2 Table. Accuracy of GBM and neural network models using k-fold cross validation (k = 10).
(DOCX)

S1 File.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Elizabeth Jeanne Parent, Serge-Étienne Parent.

Formal analysis: Elizabeth Jeanne Parent, Serge-Étienne Parent.

Funding acquisition: Léon Etienne Parent.

Investigation: Elizabeth Jeanne Parent, Léon Etienne Parent.

Methodology: Elizabeth Jeanne Parent, Serge-Étienne Parent, Léon Etienne Parent.

Project administration: Léon Etienne Parent.

Resources: Léon Etienne Parent.

Software: Elizabeth Jeanne Parent, Serge-Étienne Parent.

Validation: Serge-Étienne Parent.

Visualization: Léon Etienne Parent.

Writing – original draft: Elizabeth Jeanne Parent.

Writing – review & editing: Léon Etienne Parent.

References

1. Bohn CC, Gebhardt K. Comparison of Hydrometer Settling Times in Soil Particle Size Analysis. J Range Manag. 1989; 42(1):81–3.

2. Fisher P, Aumann C, Chia K, O'Halloran N, Chandra S. Adequacy of laser diffraction for soil particle size analysis. PLoS ONE. 4 mai 2017; 12(5):1–20. https://doi.org/10.1371/journal.pone.0176510 PMID: 28472043

3. Cheng Nian-Sheng. Simplified Settling Velocity Formula for Sediment Particle. J Hydraul Eng. 1 févr 1997; 123(2):149–52.

4. Cambardella CA. Aggregation and organic matter. In: Encyclopedia of Soil Science. Taylor and Francis. Boca Raton, FL.; 2006. p. 52–5.

5. Yang XM, Drury CF, Reynolds WD, MacTavish DC. Use of sonication to determine the size distributions of soil particles and organic matter. Can J Soil Sci. 1 août 2009; 89(4):413–9.

6. Yang X, Zhang Q, Li X, Jia X, Wei X, Shao M. Determination of Soil Texture by Laser Diffraction Method. Soil Sci Soc Am J. 12/01 2015; 79(6):1556–66.

7. Stevens A, Nocita M, Tóth G, Montanarella L, van Wesemael B. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. PLOS ONE. 19 juin 2013; 8(6):e66409. https://doi.org/10.1371/journal.pone.0066409 PMID: 23840459

8. Viscarra Rossel RA, Behrens T, Ben-Dor E, Brown DJ, Demattê JAM, Shepherd KD, et al. A global spectral library to characterize the world's soil. Earth-Sci Rev. 1 avr 2016; 155:198–230.

9. Viscarra Rossel RA, Walvoort DJJ, McBratney AB, Janik LJ, Skjemstad JO. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma. 1 mars 2006; 131(1):59–75.

10. Stenberg B, Jonsson A, Börjesson T. Near infrared technology for soil analysis with implications for precision agriculture. Infrared Spectrosc Proc 10th Int Conf. 1 janv 2002;279–84.

11. Brown DJ, Shepherd KD, Walsh MG, Dewayne Mays M, Reinsch TG. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma. 1 juin 2006; 132(3):273–90.

12. Sørensen LK, Dalsgaard S. Determination of Clay and Other Soil Properties by Near Infrared Spectroscopy. Soil Sci Soc Am J. 1 janv 2005; 69(1):159–67.

13. Leu DJ. Visible and near—infrared reflectance of beach sands: A study on the spectral reflectance/grain size relationship. Remote Sens Environ. 1 janv 1977; 6(3):169–82.

14. Okin GS, Painter TH. Effect of grain size on remotely sensed spectral reflectance of sandy desert surfaces. Remote Sens Environ. 15 févr 2004; 89(3):272–80.

15. Barrett PJ. The shape of rock particles, a critical review. Sedimentology. 1980; 27(3):291–303.

16. Baumgardner MF, Silva LF, Biehl LL, Stoner ER. Reflectance Properties of Soils. In: Brady NC, éditeur. Advances in Agronomy [Internet]. Academic Press; 1986 [cité 11 nov 2019]. p. 1–44. http://www.sciencedirect.com/science/article/pii/S0065211308606720.

17. Bowman ET, Soga K, Drummond W. Particle shape characterisation using Fourier descriptor analysis. Géotechnique. 1 août 2001; 51(6):545–54.

18. Bullard JE, White K. Quantifying iron oxide coatings on dune sands using spectrometric measurements: An example from the Simpson-Strzelecki Desert, Australia. J Geophys Res Solid Earth. 2002; 107(B6): ECV 5-1–ECV 5–11.

19. Abdi D, Tremblay GF, Ziadi N, Bélanger G, Parent L-É. Predicting Soil Phosphorus-Related Properties Using Near-Infrared Reflectance Spectroscopy. Soil Sci Soc Am J. 1 nov 2012; 76(6):2318–26.

20. Clark RN, King TVV, Klejwa M, Swayze GA, Vergo N. High spectral resolution reflectance spectroscopy of minerals. J Geophys Res Solid Earth. 1990; 95(B8):12653–80.

21. Post JL, Noble PN. The Near-Infrared Combination Band Frequencies of Dioctahedral Smectites, Micas, and Illites. Clays Clay Miner. 1 déc 1993; 41(6):639–44.

22. Xu Y, Jimenez MA, Parent S-É, Leblanc M, Ziadi N, Parent LE. Compaction of Coarse-Textured Soils: Balance Models across Mineral and Organic Compositions. Front Ecol Evol [Internet]. 2017 [cité 11 nov 2019]; 5. Disponible sur: https://www.frontiersin.org/articles/10.3389/fevo.2017.00083/full.

23. Baranoski GVG, Kimmel BW, Chen TF, Miranda E. Influence of Sand-Grain Morphology and Iron-Oxide Distribution Patterns on the Visible and Near-Infrared Reflectance of Sand-Textured Soils. IEEE J Sel Top Appl Earth Obs Remote Sens. sept 2014; 7(9):3755–63.

24. Hummel JW, Sudduth KA, Hollinger SE. Soil moisture and organic matter prediction of surface and sub-surface soils using an NIR soil sensor. Comput Electron Agric. 1 août 2001; 32(2):149–65.

25. Aitchison J. The Statistical Analysis of Compositional Data. London, UK, UK: Chapman & Hall, Ltd.; 1986.

26. Aitchison J, Greenacre M. Biplots of compositional data. J R Stat Soc Ser C Appl Stat. 2002; 51(4):375–92.

27. Díaz-Zorita M, Perfect E, Grove JH. Disruptive methods for assessing soil structure. Soil Tillage Res. 1 févr 2002; 64(1):3–22.

28. Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ. The Principle of Working on Coordinates. In: Compositional Data Analysis [Internet]. John Wiley & Sons, Ltd; 2011 [cité 11 nov 2019]. p. 29–42. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119976462.ch3.

29. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric Logratio Transformations for Compositional Data Analysis. Math Geol. 1 avr 2003; 35(3):279–300.

30. Zhang D, Tsai JJP. 2007 Advance In Machine Learning Applications In Software Engineering (Du Zhang) [Internet]. [cité 8 avr 2020]. http://archive.org/details/2007AdvanceInMachineLearningApplicationsInSoftwareEngineeringDuZhang.

31. Qin Z, Myers DB, Ransom CJ, Kitchen NR, Liang S-Z, Camberato JJ, et al. Application of Machine Learning Methodologies for Predicting Corn Economic Optimal Nitrogen Rate. Agron J. 12/01 2018; 110(6):2596–607.

32. Day PR. Particle Fractionation and Particle-Size Analysis. Methods Soil Anal Part 1 Phys Mineral Prop Stat Meas Sampl. janv 1965; agronomymonogra(methodsofsoilana):545–67.

33. Nduwamungu C, Ziadi N, Tremblay GF, Parent L-É. Near-Infrared Reflectance Spectroscopy Prediction of Soil Properties: Effects of Sample Cups and Preparation. Soil Sci Soc Am J. 1 nov 2009; 73(6):1896–903.

34. Ross GJ, Wang C. Extractable Al, Fe, Mn and Si. In: Soil Sampling and methods of analysis. Lewis Publishers. Boca Raton, FL.: Carter M.R; 1993.

35. Mehlich A. Mehlich 3 soil test extractant: A modification of Mehlich 2 extractant. Commun Soil Sci Plant Anal. 1 déc 1984; 15(12):1409–16.

36. Sheldrick BH, Wang C. Particle size analysis. In: Soil Sampling and Methods of Analysis. Lewis Publishers. Boca Raton: Carter M.R; 1993.

37. Bouyoucos GJ. Hydrometer Method Improved for Making Particle Size Analyses of Soils 1. Agron J. 10/01 1962; 54(5):464–5.

38. Gee GW, Bauder JW. Particle Size Analysis by Hydrometer: A Simplified Method for Routine Textural Analysis and a Sensitivity Test of Measurement Parameters 1. Soil Sci Soc Am J. 10/01 1979; 43 (5):1004–7.

39. Özer M, Orhan M. Determination of an Appropriate Method for Dispersion of Soil Samples in Laser Diffraction Particle Size Analyses. Int J Comput Exp Sci Eng. 26 sept 2015; 1(1):19–25.

40. Knadel M, Viscarra Rossel R, Deng F, Thomsen A, Greve M. Visible-Near Infrared Spectra as a Proxy for Topsoil Texture and Glacial Boundaries. Soil Sci Soc Am J. 1 mars 2013; 77:568–79.

41. Leblanc M, Parent E, Parent L. Lime Requirement Using Mehlich-III Extraction and Infrared-Inferred Cation Exchange Capacity. Soil Sci Soc Am J. 1 janv 2016; 80. https://doi.org/10.2136/sssaj2015.11.0414 PMID: 29657354

42. Wehrens R. Chemometric Applications. In: Wehrens R, éditeur. Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences [Internet]. Berlin, Heidelberg: Springer; 2011 [cité 11 nov 2019]. p. 235–67. (Use R). https://doi.org/10.1007/978-3-642-17841-2_11

43. Guillou F, Wetterlind W, Viscarra Rossel R, Hicks W, Grundy M, Tuomi S. How does grinding affect the mid-infrared spectra of soil and their multivariate calibrations to texture and organic carbon? Soil Res. 1 janv 2015; 53.

44. Madejová J, Komadel P. BASELINE STUDIES OF THE CLAY MINERALS SOCIETY SOURCE CLAYS: INFRARED METHODS. Clays Clay Miner. 1 oct 2001; 49(5):410–32.

45. RC team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Found. Stat. Comput.; 2017.

46. Wehrens R. Principal Component Analysis. In: Wehrens R, éditeur. Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences [Internet]. Berlin, Heidelberg: Springer; 2011 [cité 11 nov 2019]. p. 43–66. (Use R). https://doi.org/10.1007/978-3-642-17841-2_4

47. Pellerin A, Parent LE, Tremblay C, Fortin J, Tremblay G, Landry CP, et al. Agri-environmental models using Mehlich-III soil phosphorus saturation index for corn in Quebec. 2006 [cité 11 nov 2019]; https://pubag.nal.usda.gov/catalog/2775013.

48. Malley DF, Martin PD, Ben-Dor E. Application in analysis of soils. Roberts C.A., Workmann J. Jr and Reeves J.B III. Madison, Wisconsin: Agronomy; 1999. 729–784 p. (Near infrared spectroscopy in agriculture; vol. 44).

49. Draper NR, Smith H. On Worthwhile Regressions, Big F's, and R2. In: Applied Regression Analysis [Internet]. John Wiley & Sons, Ltd; 1998 [cité 29 mars 2021]. p. 243–50. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118625590.ch11.

50. Moni C, Derrien D, Hatton P-J, Zeller B, Kleber M. Density fractions versus size separates: does physical fractionation isolate functional soil compartments? Biogeosciences. 17 déc 2012; 9(12):5181–97.

51. Nuzzo A, De Martino A, Di Meo V, Piccolo A. Potential alteration of iron–humate complexes by plant root exudates and microbial siderophores. Chem Biol Technol Agric. 17 oct 2018; 5(1):19.

52. Kaiser K, Guggenberger G. Distribution of hydrous aluminium and iron over density fractions depends on organic matter load and ultrasonic dispersion. Geoderma. 15 juin 2007; 140(1):140–6.

53. Petigara BR, Blough NV, Mignerey AC. Mechanisms of Hydrogen Peroxide Decomposition in Soils. Environ Sci Technol. 1 févr 2002; 36(4):639–45. https://doi.org/10.1021/es001726y PMID: 11878378

54. Duiker SW, Rhoton F, Torrent J, Smeck N, Lal R. Iron (Hydr)Oxide Crystallinity Effects on Soil Aggregation. Soil Sci Soc Am J. 1 janv 2003; 67:606.

55. Beretta AN, Silbermann AV, Paladino L, Torres D, Bassahun D, Musselli R, et al. Soil texture analyses using a hydrometer: modification of the Bouyoucos method. Cienc E Investig Agrar. 2014; 41(2):263–71.

56. Périard Y, José Gumiere S, Rousseau AN, Caillier M, Gallichand J, Caron J. Assessment of the drainage capacity of cranberry fields: Problem identification using soil clustering and development of a new drainage criterion. Can J Soil Sci. 5 oct 2016; 97(1):56–70.

57. Šinkovičová M, Igaz D, Kondrlová E, Jarošová M. Soil Particle Size Analysis by Laser Diffractometry: Result Comparison with Pipette Method. IOP Conf Ser Mater Sci Eng. oct 2017; 245:072025.

58. Ferro V, Mirabile S. COMPARING PARTICLE SIZE DISTRIBUTION ANALYSIS BY SEDIMENTATION AND LASER DIFFRACTION METHOD. J Agric Eng. 30 juin 2009; 40(2):35.

59. Storti F, Balsamo F. Particle size distributions by laser diffraction: sensitivity of granular matter strength to analytical operating procedures. Solid Earth. 19 avr 2010; 1(1):25–48.

60. Sochań A, Bieganowski A, Ryzak M, Dobrowolski R, Bartmiński P. Comparison of soil texture determined by two dispersion units of Mastersizer 2000. In 2012.

61. Murray MR. Is laser particle size determination possible for carbonate-rich lake sediments? J Paleolimnol. 1 févr 2002; 27(2):173–83.

62. Ramirez-Lopez L, Schmidt K, Behrens T, van Wesemael B, Demattê JAM, Scholten T. Sampling optimal calibration sets in soil infrared spectroscopy. Geoderma. 1 août 2014; 226–227:140–50.

63. Vendrame PRS, Marchão RL, Brunet D, Becquer T. The potential of NIR spectroscopy to predict soil texture and mineralogy in Cerrado Latosols. Eur J Soil Sci. 2012; 63(5):743–53.

64. Wetterlind J, Stenberg B, Jonsson A. Near infrared reflectance spectroscopy compared with soil clay and organic matter content for estimating within-field variation in N uptake in cereals. Plant Soil. 1 janv 2008; 302(1):317–27.

65. Chang C-W, Laird DA, Mausbach MJ, Hurburgh CR. Near-Infrared Reflectance Spectroscopy–Principal Components Regression Analyses of Soil Properties. Soil Sci Soc Am J. 2001; 65(2):480.

66. Islam K, Singh B, Mcbratney A. Simultaneous Estimation of Several Soil Properties by Ultra-Violet, Visible, and Near-Infrared Reflectance Spectroscopy. Aust J Soil Res—AUST J SOIL RES. 1 janv 2003; 41.

67. Blott SJ, Pye K. Particle size distribution analysis of sand-sized particles by laser diffraction: an experimental investigation of instrument sensitivity and the effects of particle shape. Sedimentology. 2006; 53 (3):671–85.

68. Zobeck T. M. RAPID SOIL PARTICLE SIZE ANALYSES USING LASER DIFFRACTION. Appl Eng Agric. 2004; 20(5):633–9.