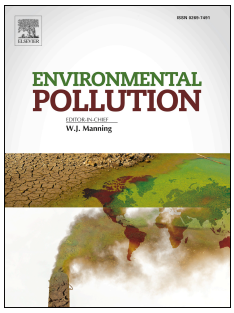
Journal Pre-proof



VIRS based detection in combination with machine learning for mapping soil pollution

Xiyue Jia, David O'Connor, Zhou Shi, Deyi Hou

PII: S0269-7491(20)36534-9

DOI: https://doi.org/10.1016/j.envpol.2020.115845

Reference: ENPO 115845

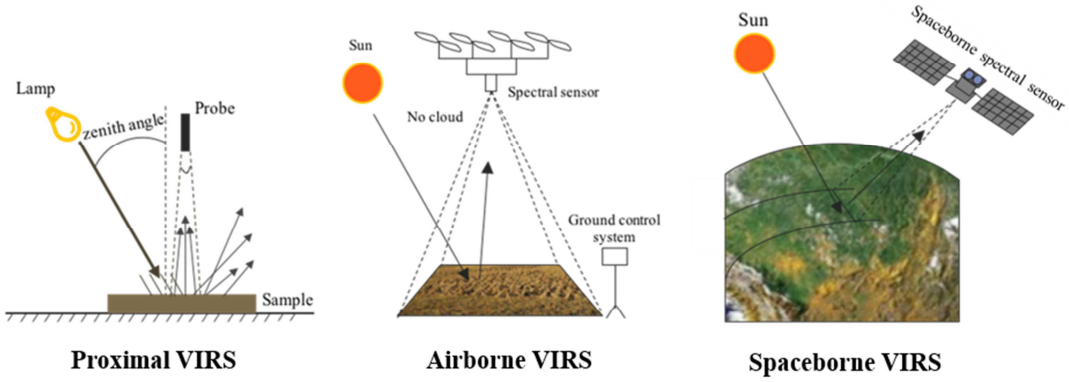
To appear in: *Environmental Pollution*

Received Date: 8 July 2020

Revised Date: 24 September 2020

Accepted Date: 11 October 2020

Please cite this article as: Jia, X., O'Connor, D., Shi, Z., Hou, D., VIRS based detection in combination with machine learning for mapping soil pollution, *Environmental Pollution*, https://doi.org/10.1016/j.envpol.2020.115845.



Proximal VIRS **Airborne VIRS** **Spaceborne VIRS**

VIRS based detection in combination with machine learning for mapping soil pollution

Xiyue Jia [1], David O'Connor [1], Zhou Shi [2], Deyi Hou [1, *]

[1] School of Environment, Tsinghua University, Beijing 100084, China
[2] College of Environment and Resource Science, Zhejiang University, Hangzhou, Zhejiang, China

[*] Corresponding author: houdeyi@tsinghua.edu.cn

Abstract

Widespread soil contamination threatens living standards and weakens global efforts towards the Sustainable Development Goals (SDGs). Detailed soil mapping is needed to guide effective countermeasures and sustainable remediation operations. Here, we review visible and infrared reflectance spectroscopy (VIRS) based detection methods in combination with machine learning. To date, proximal, airborne and spaceborne carrier devices have been employed for soil contamination detection, allowing large areas to be covered at low cost and with minimal secondary environmental impact. In this way, soil contaminants can be monitored remotely, either directly or through correlation with soil components (e.g. Fe-oxides, soil organic matter, clay minerals). Observed vegetation reflectance spectra has also been proven an effective indicator for mapping soil pollution. Calibration models based on machine learning are used to interpret spectral data and predict soil contamination levels. The algorithms used for this include partial least squares regression, neural networks, and random forest. The processes underlying each of these approaches are outlined in this review. Finally, current challenges and future research directions are explored and discussed.

Keywords: Reflectance spectroscopy; Machine learning; Soil mapping; heavy metals; Soil pollution

1 Introduction

Soils, in many places throughout the world, have been contaminated as a result of anthropogenic activities or natural processes (Hou et al., 2020b). Soil pollution is exacerbated by soil erosion (Boardman et al., 2019; Liao et al., 2019; Patriche, 2019) and acidification (Abd El-Halim and Omae, 2019; Tao et al., 2019). Soil degradation is thus threatening human health (Zhang et al., 2020), crop growth (Jia et al., 2020), and ecological system (Wang et al., 2020c), which weakens global efforts towards the Sustainable Development Goals (SDGs) (O'Connor et al., 2020). In response, the United Nations' Environment Programme (UNEP) has called on its members to report on soil pollution (UNEA, 2018). China has committed to conducting a nationwide soil pollution survey every ten years; a 2014 survey reported that 16.1% of the nation's soils are contaminated, including 19.4% of arable soils (MEE, 2014).

Detailed soil mapping based on survey data is needed to inform and guide policymakers so that they can introduce effective soil protection measures (Hou and Ok, 2019), and design green and sustainable remediation strategies (Wang et al., 2020a; Wang et al., 2020b). Accurate soil mapping, however, poses a huge technical challenge. This is primarily because soils can be highly heterogeneous (Hu et al., 2017b), with contaminant concentrations sometimes differing by several orders of magnitude within only a few meters (Han et al., 2018). Subsamples collected from a single sampling location have rendered heavy metal concentrations (e.g., Pb) that range over orders of magnitude (Brewer et al., 2017). In regional scale investigations, it is often found that average heavy metal concentrations can vary by 1~2 orders of magnitude between adjacent sampling sites.

In conventional sampling, soil samples are physically collected from the surveyed land. This is conducted according to a sampling plan, which is typical, - but not exclusively - a non-targeted grid pattern for regional assessments and targeted samples for site-specific assessments (Hou et al., 2017). Collected soil samples are subjected to laboratory-based analytical chemistry. For heavy metals and metalloids (hereafter collectively termed as

'heavy metals'), this usually involves acid digestion, pollutant extraction and detection with an inductively coupled plasma-mass spectrometer (ICP-MS) (Zuzolo et al., 2018). Geostatistical methods can then be applied to derive a spatial structure, enabling us to predict contaminant concentrations at un-sampled locations (Cheng et al., 2018; Hou et al., 2017).

This approach, however, relies on the underlying assumptions of geostatistics (i.e. spatial autocorrelation), which can be incorrect in very heterogeneous soil environments, especially where there are a diverse range of pollution sources (Hou et al., 2017). Importantly, geostatistics cannot capture spatial distribution patterns smaller than the distance between adjacent sampling locations (Goovaerts, 1999). For instance, a national-scale soil quality investigation is currently being conducted across China in which the sampling grid pattern is typically set at either 500 x 500 m or 1 x 1 km (MEE, 2017). While this investigation is expected to provide valuable information regarding levels of soil contamination on a basin-scale (e.g. units of square kilometers), and could render important data for identifying potential pollution sources (e.g. via geostatistical and/or multivariate statistical analysis), it is not intended to provide accurate predictions of pollution levels on a parcel level (i.e., sub-hectare resolution) (SC, 2016). For instance, soil samples collected within 200m of highways can contain high heavy metal contents (Pb, Zn and Cu), but such details could be overlooked on such a large-scale sampling resolution (Martinez-Carvajal et al., 2019).

Recently, researchers have explored the use of innovative tools that make the detection of soil contaminants easier and faster, thus enabling higher resolution prediction of contamination levels (Chakraborty et al., 2015). An emerging method is known as visible and infrared reflectance spectroscopy (VIRS), which involves in-the-field measurement of contaminants from either a handheld portable device, unmanned aerial vehicles (UAVs), or even satellites, for fast remote sensing of large spatial areas (**Table 1**) (Gholizadeh and Kopackova, 2019; Gholizadeh et al., 2018). The visible reflectance spectrum (VIS, 380-750 nm), near-infrared spectrum (NIR, 750-1300 nm), short wave inferred spectrum (SWIR, 1300-2500 nm), mid-infrared spectrum (MIR, 0.25 - 2.5 um) and long-wave infrared spectrum (LWI, 8-12 um)

have all been applied for VIRS based soil monitoring (Shi et al., 2016). The use of this sensing technique can accelerate soil pollution mapping at high resolution with less expense and time than other soil sampling approaches.

As with most analytical detection techniques, VIRS requires calibration to render accurate contaminant concentrations (Kemper and Sommer, 2002). However, this method requires considerable data processing before acceptable accuracy can be achieved (Kooistra et al., 2001). Recently, machine learning algorithms have been developed for this purpose (Liu et al., 2019a; Shan et al., 2018) which enable the measurement of heavy metals as well as organic contaminants (Douglas et al., 2018b; Liu et al., 2017). Therefore, the overall process for conducting soil surveys with VIRS detection is rather complicated, as shown in **Figure 1**.

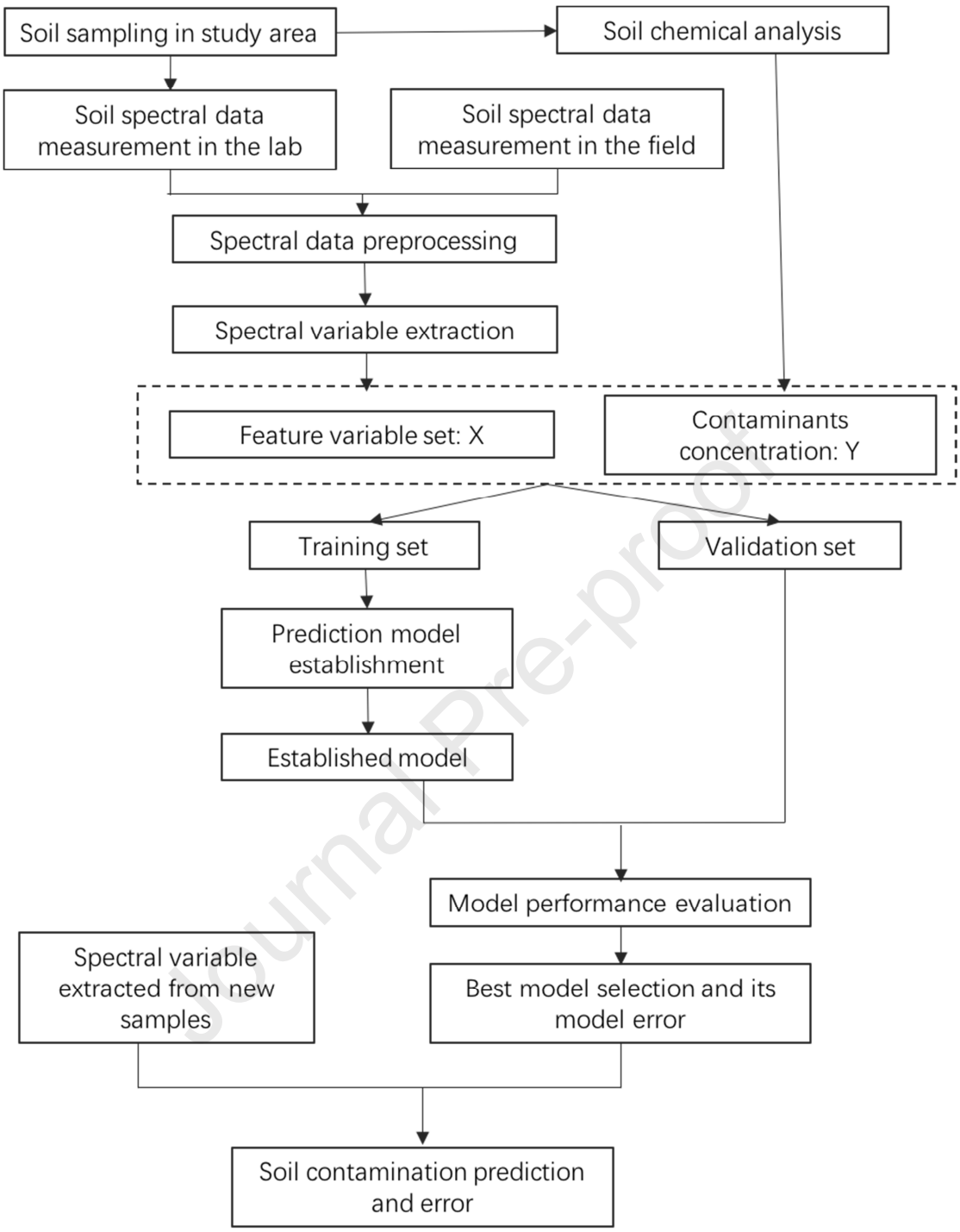


Figure 1. VIRS based detection and machine learning process

A number of recently published reviews have described different aspects of VIRS technology in detail (e.g. proximal, airborne and spaceborne spectrum) and its suitability for the detection of different types of contaminant (Gholizadeh and Kopackova, 2019; Gholizadeh et al., 2018; Shi et al., 2018). However, a detailed overview of how machine learning is used in

combination with VIRS has been lacking till now. Accordingly, the following topics are reviewed: 1) an overview of the mechanisms underlying VIRS detection of soil contamination; 2) machine learning algorithms for interpreting VIRS data; 3) application attributes.

Table 1 Soil surveys that have used VIRS

Reference	Country / Region	# of locations	Area (km^2)	Land use	Contaminant (concentration range (mg/kg))	Lab/field/remote detection	Sensing method	Wavelength range (nm)	Statistical analysis method	R^2
(Chakraborty et al., 2015)	USA	108	--	Oil production	TPH (0-326294.48)	Lab	VNIR	350~2500	PSR; RF	0.78
(Chakraborty et al., 2017)	India	200	8.1	Vegetable farming	As (2.42-10.37)	Lab	VNIR	350~2500	ENET	0.97
(Chen et al., 2015)	China	60	--	Wheat farming	Cd (0.37-5.6)	Lab	VNIR	325~1075	PLSR; BPNN	0.82
(Choe et al., 2009)	Spain	49	--	Gold mining	Pb (56.8-152.5); Cu (21.9-252.6); As (52.4-1493.8)	Lab / remote	VNIR	350~2500 / 450~2500	MLR	0.88
(Douglas et al., 2018a)	Nigeria	85	--	Oil production	TPH (16.07-252.59)	Lab	VNIR	350~2500	PLSR; RF	0.68
(Kooistra et al., 2001)	Netherlands	69	--	Flood plaints	Cd, Zn	Lab	VNIR	400~2500	PLSR	0.95
(Lassalle et al., 2018)	France	--	--	Oil production	TPH (0-140000); PAH (0-1600)	Lab	VNIR	350~2500	LDA	--
(Liu et al., 2011)	China	120~160	--	Rice farming	Cu (mean: 54.78), Cd (mean: 0.35)	Field	VNIR	350~2500	FNN	0.78
(Al Maliki et al., 2014)	Australia	31	--	Various	Pb	Lab	VNIR	400~2500	PLSR	0.46
(Okparanma et al., 2014a)	Nigeria	137	--	Oil production	PAH	Lab	VNIR	350~2500	PLSR	0.89
(Pascucci et al., 2009)	Italy	--	--	Industrial	Red mud	Field	VNIR FTIR	350~2500 8000~14000	--	--
(Peng et al., 2016)	Qatari	300	11,437	Various	As (0.4-7.9); Cr (1.9-64.9); Ni (2.3-76.1); Zn (2.8-130.9); Cu (0.6-28.8); Pb (0.5-14.1)	Remote	Landsat 8 images	450~2290	Cubist	0.74
(Ren et al., 2009)	China	33	--	Rice farming	As (19.33-403.77), Cu (31.83-190.51)	Lab	VNIR	350~2500	PLSR	0.62
(Shi et al., 2014b)	China	100	4.5	Rice farming	As (10.3-133.4)	Lab Field	VNIR VNIR	350~1200 350~2500	PLSR PLSR	0.59 0.50
(Song et al., 2012)	China	61	--	Rice farming	Cd (0.081-1.441), Cr (30.990-108.900); Pb (11.120-89.680), Cu (9.900-55.500); Hg (0.040-0.269); As (4.000-16.600)	Lab	VNIR	400~2500	PLSR	0.99
(Sun and Zhang, 2017)	China	74	--	Farming	Zn (60.44-4946.60)	Lab	VNIR	350~2500	PLSR	0.64
(Tayebi et al., 2017)	Iran	120	295	Iron mining	Fe (4436.25-271375)	Lab	VNIR	400~2450	PLSR, PCR	0.29~0.54
(Todorova et al.,	Southern	62	5151	Farming	Zn (8.54-410.46); Cu (1.68-263.56); Pb (5.60-	Lab	NIR	700~2500	PLSR	0.38~

Reference	Country / Region	# of locations	Area (km^2)	Land use	Contaminant (concentration range (mg/kg))	Lab/field/remote detection	Sensing method	Wavelength range (nm)	Statistical analysis method	R^2
2014)	Bulgaria				82.49); Cr(3.90-150.82); Ni (1.09-118.62)					0.89
(Wang et al., 2014)	China	100	--	Farming	As (1.91-21.90); Pb (9.01-37.60); Zn (29.32-117.49); Cu (8.30-26.38)	Lab	VNIR	350~2500	PLSR	0.49~ 0.69
(Webster et al., 2016)	Italy, Australia, Nigeria	194	--	Various	TPH (0-60000)	Lab	IR	6000~650 cm^{-1}	PLSR	0.99
(Wu et al., 2005)	China	120	--	--	Hg (0.04-1.26)	Lab	VNIR	380~2500	PCR	0.69
(Zhao et al., 2018)	China	75	179700	Various	Hg (0.018-0.615)	Lab	VNIR	340~2511	MLR, BPNN	0.92
(Stazi et al., 2014)	Italy	135	108	Farming	As (25-1045)	Lab	VNIR	500-800	PLSR, SVM	r: 0.82
(Pelta et al., 2019)	Israel	--	--	--	Oil	Field	VNIR	400 - 2500	LDA	Recall: 0.93

Acronyms: Statistical analysis: BPNN= back propagation neural network; ENET=elastic net regression; FNN=fuzzy neural network; MLR=multiple linear regression; PCR=principal component regression; PLSR=partial least squares regression; PSR=penalized spline regression; RF=random forest regression; SVM=support vector machine; LDA= linear discriminant analysis; Sensing: VNIR= visible near-infrared reflectance; Contaminants: As=arsenic; Cd=cadmium; Cu=copper; Cr=chromium; Hg=mercury; Ni=nickel; PAH=polycyclic aromatic hydrocarbon; Pb=lead; TPH=total petroleum hydrocarbon; Zn=zinc;

2 Predictors and underlying mechanisms

The use of VIRS relies on the fact that atoms and molecules absorb and emit electromagnetic radiation because of electron transition and molecular vibration (Shi et al., 2018). Identification and quantification of different chemicals can be achieved based on emission and absorption spectra. In soil contamination monitoring, VIRS captures reflectance energy from the land surface with the reflectance spectra informing us of the soil composition (Shi et al., 2014a).

Certain organic soil contaminants, such as polycyclic aromatic hydrocarbons (PAH) and petroleum hydrocarbons (collectively termed total petroleum hydrocarbons (TPH)), are often detectable in visible and infrared reflectance spectra (Chakraborty et al., 2010; Douglas et al., 2018b). In the case of heavy metals, direct monitoring can only be achieved at concentrations that rarely occur in the field (e.g., 4000 mg/kg in the case of Cd) (Liu et al., 2017; Wu et al., 2007; Xia et al., 2007). Fortunately, interactions between trace levels of heavy metals and more abundant soil components (e.g. clay, organic matter and Fe oxides) provides an opportunity to detect them indirectly (Wu et al., 2005; Zhao et al., 2018). Another way of detecting trace levels of metals is to monitor vegetation spectra because of the influence contaminants exert on plant physiology (Shi et al., 2016). Specific mechanisms for predicting soil pollutants are introduced in this section.

2.1 Molecular vibration

In the case of organic compounds, stretching and vibrations of aliphatic (alkyl) compounds and certain functional groups can often be observed in NIR and MIR spectra (Douglas et al., 2018a; Forrester et al., 2013). The first overtone of TPH is observed in the wavelength range of 1600-1820 nm, and the second at 1100-1500 nm. Observation of the second overtone is more difficult if TPH concentrations are

relatively low (Hauser et al., 2013). In the case of PAHs, the first overtone of C-H stretching and deformation of C-H combination, and the second overtone of C-H stretching in aromatic C-H are observed at wavelengths of 1675 nm, 1417 nm and 1097 nm, respectively (Okparanma et al., 2013). In MIR region, the peaks around 1630-1580 cm^{-1}, 1930-1840 cm^{-1} and 2060-1930 cm^{-1} are associated with aromatic functions (Hobley et al., 2014; Ng et al., 2017).

The concentration of TPH in soil samples collected from oil-contaminated sites can be determined by Vis-NIR spectrophotometry, with absorption peaks around 1712 nm, 1758nm and 2207 nm (Douglas et al., 2018a). The 1712nm and 1758 nm peaks are in the first overtone region, which are attributed to the stretching of terminal CH_3 and saturated CH_2 in alkyl (Workman and Workman, 2007); the 2207 nm peak is associated with either amide (C=O) or the stretch and bending caused by crude oil (Rossel and Behrens, 2010). Okparanma et al. (2014) demonstrated that PAHs in soil are detectable at a wavelength of 1670 nm, which was attributed to aromatic C-H. The calibration R^2 value for their PAH prediction model was 0.89, and the PRD reached 3.12.

Observed spectra for organic contaminants may overlap with soil organic matter (SOM), but the presence of SOM would not normally influence TPH detection (Ng et al., 2017). This is because TPH consists of medium length chains, whereas SOM mainly composes of long $–CH_2$ chains, and relatively low amounts of $–CH_3$ (Forrester et al., 2013). For example, it has been found that spiking TPH contaminated soils with SOM has little effect on observed NIR absorption spectra, but it may affect the MIR region (especially 1980, 1870, and 1790 cm^{-1} peaks) (Ng et al., 2017). Forrester et al., (2013) noted several characteristic absorption peaks in the spectrum of TPH contaminated soil with the presence of SOM, which were attributed to the vibrational overtone of terminal methyl in the MIR region. The presence of such peaks can fortuitously aid TPH detection.

2.2 Soil properties

2.2.1 Soil organic matter

Soil organic matter (SOM) derives from the breakdown of plant and animal debris. Many studies have shown that the combination of molecular vibration and overtones in SOM, including O-H, C-H, C=O groups, can be identified in Vis-NIR spectra (Kooistra et al., 2001). Because humic and fulvic acids in SOM bind with heavy metal cations, through COOH, OH, and C=O interactions (Piccolo & Stevenson, 1982), correlation between SOM and heavy metals levels has been observed (Egli et al., 1999).

Several studies have exploited SOM spectral bands to predict heavy metal concentrations in soil. For example, at an agricultural site contaminated by polluted irrigation water, it was found that Cd levels were positively correlated with SOM. Measurement of 410, 581-626, and 670-690 nm wavelengths were found to be effective for predicting Cd levels (Chen et al., 2015). Chakraborty et al. (2017) used VIS-NIR spectroscopy to determine As concentrations using the absorption feature associated with O-H and C-H bonds in SOM at a wavelength of around 1290-1310 nm.

2.2.2 Fe-oxides

Iron oxides and hydroxides are widely found in the earth's surface, especially iron oxyhydroxide (goethite), which forms from weathered iron-rich minerals (Shi et al., 2014a; Wu et al., 2007). Because Fe-oxides are characterized by high surface charge, large surface area and strong adsorption capacity, they play a crucial role in the fate and transport of heavy metals in the subsurface (Shuman, 1982). For this reason, concentrations of soil heavy metals often correlate to those of Fe-oxides (Wu et al., 2007). VIRS detection is possible because various peaks, including 565, 435, 500 nm

and bands between 650 and 760 nm, have been associated with Fe-oxides, with significant correlations identified with soil heavy metals (Xia et al., 2007).

Kemper and Sommer (2002) found that As closely correlated with the reflectance of Fe oxide related bands at ~550 nm wavelength. Wu et al. (2011) reported that Ni concentrations can exhibit a negative correlation with iron oxides, especially in the 480-580 nm wavelength region. Chakraborty et al. (2017) reported that As concentrations had a strong correlation with Fe oxides, meaning that high levels of regression fitness with diffuse reflectance data could be achieved.

2.2.3 Clay Minerals

Hydroxyl absorption associated with molecular water can be detected at 1400 nm, 1900 nm and 2200 nm, which is associated with clay minerals (Ibrahim et al., 2008; Kemper and Sommer, 2002; Zhao et al., 2018). Bands at 538 nm wavelength correspond with the Si-O and Si-O-Al bonds in clay minerals (Song et al., 2012). This is important for soil contamination surveys because the cation exchange capacity (CEC) of clays minerals are often high, meaning that heavy metals cations can easily replace clay mineral cations. Heavy metals tend to sorb to clays by Van der Waals forces and hydrogen bonds (Kumpiene et al., 2007).

Concentrations of heavy metals in mine tailings can correlate with bands at 1400 nm, 1900 nm and 2200 nm (Kemper and Sommer, 2002). Choe et al. (2009) found that As levels had a statistically significant ($p = 0.006$) correlation with reflectance at 2200 nm. The calibration R^2 value was 0.56. Song et al. (2012) found that Cu displayed the highest correlation at 538 nm, which was related to Si-O bands, with an R^2 value of 0.551 ($p < 0.001$). A positive correlation between Hg concentration and adsorption at 2210 nm was reported by Wu et al. (2005).

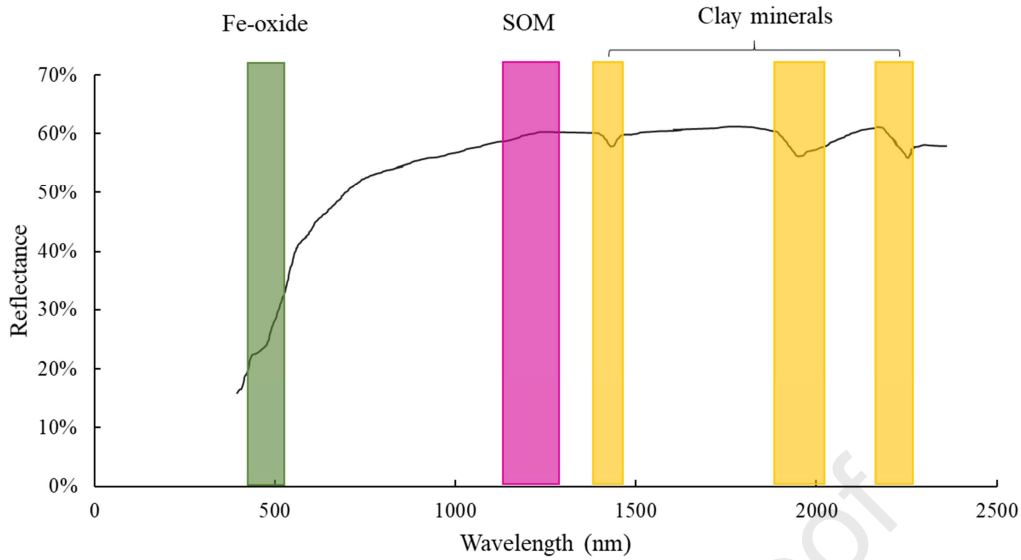


Figure 2 Key wavelengths for soil contamination prediction based on VIRS

2.3 Vegetation

Wavelengths around 540, 690, 730, and 780 nm are closely associated with chlorophyll-a/-b contents in plant leaves and pigment composition (Blackburn, 1998). Leaf anatomical features, including mass per area and structure differences (i.e., cell morphology and parenchyma structure) can present significant correlation with NIR peaks (Ourcival et al., 2010). By combining VIS, NIR and short wave infrared, the water content in vegetation can be monitored (Cao et al., 2013). Because pigments, anatomical features, and plant water content relate to plant health (Shi et al., 2016), vegetation reflectance can be used for assessing soil contamination levels (Huang et al., 2009). Changes to the physicochemical and biological properties of soils also cause an effect on vegetation reflectance (Jiang et al., 2010; Lassalle et al., 2018; Rosso et al., 2005).

Shi et al. (2014b) explored the reflectance of rice plants to predict soil As concentrations. It was found that 768, 939, 953, 1132, and 1145 nm wavelengths correlated to As levels, while 768, 939 and 953 nm wavelengths were related to the

leaf area index and chlorophyll density, and 1132 and 1145 nm wavelengths were associated with the cellular structure, which could be used for indirect measurement of As levels. A partial least squares regression (PLSR) model was developed with an R^2 of 0.77 (Shi et al., 2014b). Two-band and three-band vegetation indices have been used to predict As levels by leaner and polymeric regression models. The three-band index $(R_{716} - R_{568})/(R_{552} - R_{568})$ is the more effective of these (Shi et al., 2016).

It should be noted that environmental factors unrelated to soil contaminant levels (e.g., nutrient availability) may affect the health of plants and should be considered when relying on vegetation reflectance data (Lassalle et al., 2018). Moreover, the sensitivity to contaminant exposure is different for different plant species (Lassalle et al., 2018; Sanches et al., 2013).

2.4 Factors affecting VIRS detection

Several factors, including contaminant concentrations and other soil components (e.g. SOM and clay minerals), affect VIRS detection. Because the contaminant concentration determines how much energy is absorbed and emitted, the higher the soil contamination level the easier it is to interpret reflectance spectrum directly (Okparanma and Mouazen, 2013; Somsubhra et al., 2014). Wu et al. (2007) noted that when concentrations of Cr and Cu were higher than 4000 mg/kg, adsorption could be discriminated at wavelengths of around 610 and 830 nm, respectively. However, detection was not possible at concentrations below 1000 mg/kg. Moreover, when contaminant concentrations are limited, spectral peaks may shift from their usual wavelength positions (Somsubhra et al., 2014).

Because most soil heavy metals only exist in trace amounts, they must be monitored indirectly. The predictability of trace levels of heavy metals varies depends on their *in situ* behavior. Kemper and Sommer (2002) found that Pb could be predicted with a

high R^2 value (0.940), followed by Hg and As with R^2 values of 0.929 and 0.858, respectively. The R^2 values for Cd, Cu and Zn were much lower (Kemper and Sommer, 2002). Hou et al. (2019) monitored six heavy metals through hyperspectral VIRS detection, finding that prediction accuracy decreased in the order of Ni>Zn>Pb>Cu>Cr>Cd.

In general, detectability relates to the affinity of contaminants to different soil components (Stazi et al., 2014; Xia et al., 2007). Song et al. (2012) found that the detection of heavy metals in soil was correlated to their affinity to Fe_2O_3, Al_2O_3 and SOM. (Wu et al., 2007) reported that goethite detection (at 500 nm) was positively correlated with various heavy metals.

All the predictors mentioned above have certain advantages and disadvantages. Molecular vibration are widely used to predict organic pollutants, but the key wavelengths will shift while the pollutant concentration changes (Somsubhra et al., 2014). Soil-property-related predictors (i.e. Fe-oxide and clay minerals) are mainly used to predict heavy metals in soil, since heavy metals tend to sorb to them and significant statistic relationship are identified between them (Wu et al., 2007). However, Douglas et al. (2018a) have observed that correlation between the contents of soil-property-related predictors (e.g. organic matter and clay) and TPH were not significant, and have obtained similar conclusion. Therefore, the use of soil-property-related predictor in predicting soil organic concentration is limited. Vegetation can potentially indicate the extent of soil pollution, which is a crucial indicator especially when spaceborne spectrometers are employed (Shi et al., 2014b). However, compared to molecular vibration and soil components, vegetation, incorporates more unstable factors when used to predict soil pollution, such as their different ability on indicating pollutants contents (Lassalle et al., 2018). Further exploration are requisite to interpreting the relationship between vegetation spectrum and soil pollution concentration.

3 Remote sensing techniques

Proximal, airborne and spaceborne carrier devices have been employed for VIRS based soil contamination detection, allowing large areas to be covered at low cost and with minimal secondary environmental impacts (**Figure 3**).

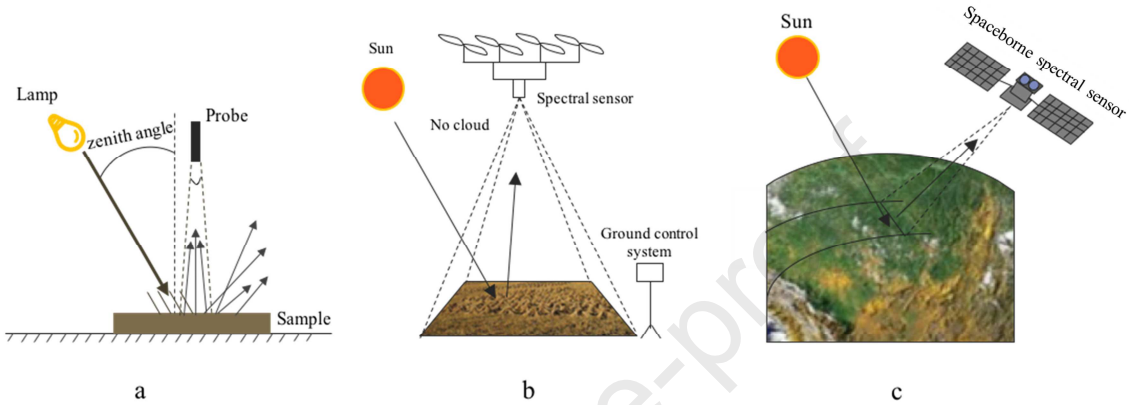


Figure 3 Schematic diagrams of (a) proximal sensing, (b) airborne imaging and (c) a spaceborne spectrometer (Shi et al., 2014a; Shi et al., 2018)

3.1 Proximal sensing techniques

Proximal soil sensing refers to the collection of soil information close to soil (i.e., within 2 m) (Rossel et al., 2011). Various proximal VIRS sensors have been developed that collect physical, chemical and biological information in this way, with the most common detectors listed in **Table 2**. The spectral resolutions of those spectrometers are in the range 0.05-10 nm and 2-8 cm^{-1} in the Vis-NIR and MIR range, respectively. In general, higher detection accuracy is achieved with smaller spectral resolution.

In laboratory-based proximal sensing studies, field soil is transported to the laboratory for scanning. Chakraborty et al. (2017) evaluated As levels with a portable Vis-NIR spectroradiometer, reporting a calibration R^2 of 0.97. Webster et al. (2016) measured MIR spectra to assess TPH levels in soil, reporting R^2 values of up to 0.99. However,

outdoor field environments are more complex than those in the laboratory. Multiple factors affect spectra, including air and soil moisture content. Therefore, R^2 values reported in field studies are usually much lower than those obtained in the laboratory (Shi et al., 2014b). Nevertheless, Shi et al. (2014b) was able to predict soil As contamination from vegetation using a portable spectroradiometer. Linear discriminant analysis (LDA) of vegetation spectra has also been shown to discriminate water-deficient or oil-contaminated soils in the field (Lassalle et al., 2018).

Table 2 Commonly used portable proximal spectrophotometers

Spectrophotometer	Manufacturer	Wavelength range	Spectral resolution	Reference
PSR-3500® VisNIR spectroradiometer	Spectral Evolution, USA	350-2500 nm	3.5 to 10 nm	(Chakraborty et al., 2017)
Perkin-Elmer Lambda 900 spectrophotometer	Perkin-Elmer, Germany	400-2500 nm	UV-VIR: < 0.05 nm NIR: < 0.20 nm	(Song et al., 2012)
FieldSpec HandHeld	Analytical Spectral Devices, Inc., USA	325-1075 nm	3.5 nm at 700 nm	(Zhang et al., 2016)
LabSpec® 2500	Analytical Spectral Devices, Inc., USA	350-2500 nm	10 nm at NIR	(Douglas et al., 2018b)
FieldSpec® 3	Analytical Spectral Devices, Inc., USA	350-2500 nm	3 nm at 700 nm 10 nm at 1400/2100 nm	(Wang et al., 2014)
FieldSpec® 4	Analytical Spectral Devices, Inc., USA	350-2500 nm	3 nm @ 700 nm 8 nm at 1400/2100 nm	(Hou et al., 2019)
FTIR TENSOR 37	Bruker Optics, Ettlingen, Germany	2500–25,000 nm	4 cm^{-1} between 3996 to 599 cm^{-1} at	(Ng et al., 2017)
Nicolet 6700 FT-IR spectrometer	Thermo Scientific, USA	2500-20000 nm	1.928 cm^{-1}	(Song et al., 2012)
Hand-held 4100 ExoScan FTIR spectrometer	Agilent Technologies, USA	6000–650 cm^{-1}	8 cm^{-1}	(Webster et al., 2016)

3.2 Airborne imaging

Airborne spectrometry is a promising approach to remote sensing. Various hyperspectral sensors have been equipped to aircraft and unmanned aerial vehicles (UAVs) (**Table 3**). For example, airborne imaging was used to predict Pb, Zn and As by Choe et al. (2008). The distribution of red mud dust has also been observed by MIVIS based airborne imaging (Pascucci et al., 2009). The spatial resolution of airborne imaging is determined by the field-of-view (FOV) and altitude of the sensor, which can be adjusted in accordance with practical demands. It should be noted that

airborne imaging is affected by factors such as air moisture and vegetation cover, meaning that successful field surveys are currently limited (Shi et al., 2014a).

Table 3 Hyperspectral imaging sensors for airborne imaging

Hyperspectral sensor	Spectral range (nm)	Spectral bands	Spectral resolution (nm)	Spatial resolution	Reference
HyMap	450-2480	126	16	5 m	(Choe et al., 2008; Franke et al., 2009)
ASIA	400-2500	481	3.3	1.5 m (altitude: 1.3 km)	(Hillnhuetter et al., 2011)
AVIRIS	400-2500	224	10	20 m (altitude: 20 km)	(Chabrillat et al., 2002)
CASI	450-2500	--	10-15	3 m	(Dutkiewicz et al., 2009)
MIVIS	430-12700	102	9-540	--	(Forzieri et al., 2012)

3.3 Spaceborne spectrometers

Spaceborne spectrometry is an efficient, economical and increasingly accessible approach to soil mapping (Guan et al., 2019). Earth observation satellites with high resolution sensors and high numbers of spectral bands have been launched, including the European Space Agency's (ESA's) Sentinel satellites, NASA's Landsat program, and China's HJ-1 (Gholizadeh et al., 2018; Roy et al., 2014). NASA's Landsat-8 is equipped with an operational land imager and thermal infrared sensor, covering the 11 wavelength bands (4 visible, 1 near-infrared, 2 shortwave infrared, 1 panchromatic, 1 cirrus and 2 thermal infrared) (Roy et al., 2014). ESA's Sentinel-2 satellite has 13 bands, covering 443 to 2190 nm wavelength (BERGER et al., 2012). The HJ-1 satellite is equipped with a hyperspectral sensor with 115 spectral bands in the range of 450-950 nm (Liu et al., 2015).

Peng et al. (2016) and Guan et al. (2019) used the spectral bands of the Landsat-8 satellite, involving brightness and normalized difference vegetation index (NDVI)) with land features (e.g. elevation and slope), to predict the concentrations of As, Cr, Ni, Pb and Zn. Liu et al. (2019b) used spectral data from the HJ-1 satellite to predict soil concentrations of Cd through multiple nonlinear regression, achieving an R^2 of 0.81.

Advanced hyperspectral satellites that will provide higher accuracy are due to be launched in the coming years, including the HyspIRI satellite with 214 spectral bands, the CCRSS satellite with 328 spectral bands and the EnMAP satellite with 242 spectral bands (Gholizadeh et al., 2018).

4 Spectral data analysis by machine learning

4.1 Regression

Regression algorithms are often used to interpret spectral data (Chakraborty et al., 2017). For example, univariable regression is used to predict independent variables from a single dependent variable. Shi et al. (2016) used this approach to predict soil As levels (R^2 = 0.56). However, as multiple dependent variables can usually be extracted from spectral data, multiple linear regression (MLR) is more commonly used. Compared to other advanced multivariate algorithms, MLR is easier to perform and interpret. However, MLR prediction accuracy is reduced when predictor variables involve non-linear relationships. Kemper and Sommer (2002) employed MLR to predict the concentration of heavy metals from spectral data (R^2 = 0.234-0.957). Ng et al. (2017) used this approach to predict soil TPH levels (R^2 = 0.71).

The most used techniques for interpreting spectral data are principle component regression (PCR) and partial least squares regression (PLSR). PCR is a two-step technique in which predictor variables are transformed into principal components by principal component analysis (PCA), which are then inputted as predictors into MLR (Wu et al., 2005). The first step allows multi-linear problems to be solved. As an enhancement to PCR, PLSR has a similar structure but also takes response variables into account in the PCA step. Therefore, PLSR not only handles multi-linear data but also allows for the number of variables to exceeds that of the samples (Shi et al., 2014b; Wang et al., 2014). Douglas et al. (2018b) and Webster et al. (2016) used

PLSR to predict soil TPH levels with Vis-NIR and MIR spectral data, reporting R^2 values of 0.63 and 0.99, respectively.

Other regression approaches include elastic net regression (ENR) and penalized spline regression (PSR). ENR overcomes the problem of overfitting, whereas PSR is able to solve problems of high-dimensional data analysis. Both ENR and PSR have been utilized to predict soil As levels with reported R^2 values of 0.97 and 0.89, respectively (Chakraborty et al., 2017).

4.2 Neural network

The neural network is composed of artificial neurons which form layers that further link into connections, thus mimicking the human brain (Laberge et al., 2000). This non-linear method has attracted extensive interest in multiple fields (Abedinia et al., 2018; Park et al., 2011). In soil surveys, back-propagation neural network (BPNN) has obtained attention for its ability to interpret spectral data more effectively than partial least squares regression (PLSR) (Chen et al., 2015; Zhao et al., 2018).

Algorithm optimization of BPNN has been explored to improve predictive accuracy. For example, Zhao et al. (2018) used BPNN with a genetic algorithm (GA) to predict soil Hg levels. A combination of particle swarm optimization and BPNN (PSO-BPNN) mitigates slow convergence and avoids trapping in local minima. (Liu et al., 2019b) used PSO-BPNN to predict concentrations of Hg, Cd and As with higher accuracy than primary BPNN. Tian et al. (2019) optimized BPNN with the combination of GA and the ant colony algorithm to predict heavy metal concentration, with a reported R^2 value for Cr detection (0.87) higher than for primary BPNN (0.55).

4.3 Random forest

Random forest (RF) evolved from the decision tree algorithm, a classical and intuitive algorithm that exploits top-down and binary splits to handle regression and classification problems (Ellis et al., 2014). Because this can lead to high variance and overfilling, bagging (bootstrapping aggregation) has been included. A variety of decision trees are “trained” on extracted subsamples and the average value of splitting points. However, trees generated by bagging may correlate with each other because they are trained with similar samples. RF was developed to de-correlate trees, which selects a subsample of a feature set for each tree, compelling trees to consider all features (Svetnik et al., 2003). It has been increasingly used in environmental applications and achieved superior results in comparison with other predictive techniques (Zhu et al., 2020a; Zhu et al., 2020b).

Douglas et al. (2018a) used RF to predict the concentration of TPH in soil using Vis-NIR data. The reported R^2 value and PRD were 0.68 and 1.85, which was higher than that for PLSR (0.54 and 1.51, respectively). Wei et al. (2019)used RF to determine soil As concentrations (R^2 = 0.95). Chakraborty et al. (2017) reported that the performance of RF in predicting As levels was higher than PSR.

4.4 Other algorithms

Support vector machine (SVM) and linear discriminate analysis (LDA) algorithms have also been explored in several studies (Lassalle et al., 2018; Shan et al., 2018). SVM is an effective and classical classification algorithm, which can also be used for regression. The LDA method is like linear regression but involves data classification. Stazi et al. (2014) used SVM to predict the concentrations of As in agricultural soil with 18 variables (R^2 = 0.82 and PRD = 2.03). Wei et al. (2019) reported As detection with an R^2 value of 0.91 using 5 feature variables.

5 Data acquisition

5.1 Soil data collection

Because VIRS requires a calibration model, both soil contaminant concentrations (e.g., traditional physical sampling) and corresponding VIRS spectra data are simultaneously collected to build a calibration model. Since soil properties can vary significantly, soil samples may be needed to build unique calibration models for each location studied. In some studies, soil samples have been prepared in the laboratory with spiked soil samples (Pelta and Ben-Dor, 2019). There are no existing studies to indicate the effect of soil sampling depth, however, it should be noted that airborne and spaceborne spectrometers will only observe surface soils. Therefore, soil sampling depth is usually limited to less than 20 cm (Chakraborty et al., 2010; Hou et al., 2019).

5.2 Spectral measurement

Proximal sensing in the laboratory requires soil samples to be processed. Firstly, debris, organisms and large gravel are removed before sieving (typically 2 mm mesh) (Antonucci et al., 2012; Song et al., 2012). Some studies involved grinding soil to 38-840 μm particle size (Liu et al., 2019b). The soil is then dried for 1-14 days, either at room temperature or at a constant oven temperature (i.e., 40 □) (Liu et al., 2019b; Song et al., 2012). Some studies applied higher temperatures to speed up drying (e.g. 65 □ or 105 □), but it should be noted that this could remove any volatile content from the soil (Douglas et al., 2018b; Stazi et al., 2014).

Operational parameters used when scanning soil samples in the laboratory are detailed in **Table 4.** In this process, the sample is placed on smooth surface (e.g., a glass slide or petri dish) to diffuse reflection and gain a good signal-to-noise ratio (Okparanma et al., 2013). Samples can be smoothed by saturating with distilled water to make a

slurry before drying (usually at 40 °C) (Wu et al., 2005)or simply smoothed over manually (Liu et al., 2019b). Measurements are conducted in a darkroom with a light source. In the case of Vis-NIR spectral measurements, the light source could be a tungsten filament lamp or a tungsten halogen lamp with a wavelength of 320-2500 nm (Sridhar et al., 2011). The light source is normally placed 30-70 cm above the soil (Shen et al., 2019) and the detector a distance of 10-120 cm (Pelta and Ben-Dor, 2019; Wei et al., 2019). Keeping the light source and detector in specific distances from soil ensures that the light can evenly irradiate the surface of the measured object, and maintains the sample in the FOV of the detector (Shi et al., 2014a). Before measurement, background adsorption is carried out with a white reference material, such as Spectralon, polytetrafluoroethylene (PTFE) or $BaSO_4$ (Kooistra et al., 2001). Additionally, each sample should be measured 3-10 times to reduce error (Douglas et al., 2018a).

Table 4 Operational parameters used in the laboratory

Soil particle size (μm)	Lamp power (W)	lamp to soil (cm)	Detector to soil (cm)	Parallel test	Spectral calibration	Reference
840	500	--	--	5	white Spectralon	(Liu et al., 2019b)
--	500	40	15	10	white Spectralon panel	(Wang et al., 2014)
150	1000	50	15	--	--	(Hou et al., 2019)
< 2000	50	--	--	10	--	(Zhao et al., 2018)
--	50	60	--	5	white $BaSO_4$ panel	(Todorova et al., 2014)
2000	--	30	--	10	white $BaSO_4$ panel	(Ren et al., 2009)
--	5	--	--	4	NIST certified white reference	(Chakraborty et al., 2017)
149	50	40	--	10	--	(Chen et al., 2015)

There are two main approaches for spectral measurement in the field: portable spectral devices and airborne devices (see Section 3). Solar light intensity plays an important role in the quality of spectral data in the field. Because clouds reflect and absorb light at certain wavelengths, cloud cover should be minimal. Most studies are conducted with zero cloud cover and good visibility (e.g., 60 km) (Götze et al., 2016). Rainfall and high humidity (which condenses into water films) should be avoided (Soriano-Disla et al., 2014).

5.3 Soil spectral libraries

To expand the use of VIRS in soil monitoring, international efforts are being made to establish spectral libraries. The first was established by the US National Soil Survey Center in 2006, which contains 3768 samples from the US and 416 samples from countries in Africa (125), Europe (112), Asia (104) and the Americas (75) (Brown et al., 2006). Other institutions have published data including 21,500 spectra collected from 4000 soil profiles in Australia, and the spectra of 20,000 samples collected across Europe (Antoine et al., 2013; Rossel and Webster, 2012). In recent years, Viscarra et al., (2016) compiled the Vis-NIR spectra of 23,631 soil samples collected from 35 institutions around the world (Rossel et al., 2016).

6 Statistical analysis methods and modeling strategies

6.1 Data pre-processing

Data pre-processing is used to render data valid for model building. Kooistra et al. (2001) reported that prediction accuracy and model quality was vastly improved after pre-processing was carried out. Pre-processing usually involves outlier removal, noise minimization and curve smoothing (Stazi et al., 2014).

Data outliers may originate from the sample itself or from experimental operations. Removal of outliners is one of the keys to establishing stable and effective predictive models. Outliers ought to be identified using a systematic method, such as principal component analysis (PCA) (Shi et al., 2014b), with outliers identified by a score matrix. Chakraborty et al. (2017) used PCA to pre-process spectral data, identifying 10 outliers.

A normal distribution is a prerequisite for some statistical methods, such as Pearson correlation analysis. A normality test can be used to check data normality (e.g., a

Shapiro–Wilk test and Kolmogorov-Smirnov test showing a p of > 0.05). If the data is non-normal, transformations such as Box-Cox transformation and logarithmic transformation can be applied (Chakraborty et al., 2015; Chakraborty et al., 2010).

Noise in collected spectra will often relate to the roughness of the surveyed land or the observation angle (Zhao et al., 2018). Bands showing large amounts of noise can be removed (e.g., the initial and tail bands) (Hou et al., 2019; Pelta et al., 2019). Additionally, mathematical transformation methods can be adopted to reduce noise levels (**Table 5**).

Table 5 Spectral transformation methods

Name	Objective
Mean centering	Eliminate the absolute absorption value of the spectrum, increase the difference between the sample spectra, and improve the robustness and prediction ability of the model
Orthogonal signal correction	Filter out signals that are not related to the concentration of the target pollutant in the spectrum
Standard Normal Variate (SNV)	Eliminate spectral errors caused by solid particle size and surface scattering
Multiplicative Scatter Correction	Same as SNV
Savitzky-Golay smoothing filter	Smooth the spectral curve to reduce noise
Derivative (first derivation and second derivation)	Correct the spectral baseline, eliminate interference from other backgrounds, and improve spectral resolution.

The relative effectiveness of data pre-processing has been analyzed in several studies. Liu et al. (2017) reported that reflectance data processed by logarithm and continuous removal increased the level of correlation with heavy metals. Chen et al. (2015) compared six pre-processing methods, finding that orthogonal signal correction most effectively reduced noise and improved prediction accuracy. However, reported optimal preprocessing methods have varied among studies, owing to the specific features of the spectral data (Chen et al., 2015; Kooistra et al., 2001). In practice, the use of multiple data preprocessing methods may be needed to determine the optimal approach.

6.2 Variable construction

Variables can be classified as two types: 1) raw or preprocessed spectral feature bands; 2) combined spectral data. The first type provides the most representative information and higher model quality.

There are two methods for selecting feature bands: 1) linear regression; or 2) PCA. For linear regression, reflectance correlation coefficients are calculated, with the bands of highest value used as feature variables. The magnitude of correlation coefficients can depend on the pre-processing method applied (Liu et al., 2017). In PCA, uncorrelated principal variables are extracted explaining the highest variance. Calibration models such as PCR and PSLR are constructed based on principal components. Other prediction methods also use principal components as prediction variables (e.g. RF) (Douglas et al., 2018a). The number of feature bands used for modeling can range from one to hundreds (Liu et al., 2019b). Calibration R^2 values will tend to increase as the number of feature variable increases, but overfitting may occur at higher numbers. As a rule of thumb, the optimal number of variables is around one third of the number of samples.

Combined spectral data can also serve as variables. For this, correlation analysis can be used to select the most effective combination. Liu et al. (2019a) selected two combinations of spectral data to predict Cd, Hg and As levels in soil. Some commonly used spectral indexes for vegetation, such as the normalized difference vegetation index (NDVI) and the infrared percentage vegetation index (IPVI), have been identified as efficient predictors when using vegetation reflectance data (Shi et al., 2014a). For example, Shi et al. (2016) employed different vegetation indices to predict As levels (R^2 = 0.75).

6.3 Model selection

Various models for interpreting spectral data were introduced in Section 4. The model function should be considered firstly in model selection. If the underlying relationship is non-linear, algorithms such as PSR, neural network, RF should be used. PLSR and stepwise regression could help to diminish the risk of collinearity (Chen et al., 2015; Somsubhra et al., 2014). Model parameters should be considered carefully to avoid overfitting. For instance, in the neural network algorithm, the number of neurons in each hidden layer, the number of hidden lays and the selection of propagation functions can influence model accuracy. Overfitting can occur if the model is too complex. Models with different combinations of parameters should be built, tested and compared. Additionally, models can be optimized with other advanced algorithms, including the genetic, particle swarm optimization, least absolute shrinkage and selection operator algorithms (Liu et al., 2019a; Wang et al., 2014). Such algorithms help improve solution searching and avoid the problem of overfitting.

6.4 Model validation

Model validation is required to determine prediction error and evaluate model quality. After initial data pre-processing, it is useful to split the data into two separate sets: one set for model training and another for validation (Okparanma et al., 2014b). Usually, around 70% of data is used for training and 30% for validation (**Table 6**).

Model performance can be evaluated systematically using cross-validation techniques. In k-fold cross-validation, the data is randomly divided into k equal sized subsamples, with one subsample retained as validation data. The remaining k−1 subsamples are used as training data. The process is repeated k times, with each subsample used once as the validation subset. The average error serves as the performance parameter (Liu et al., 2017). The leave-one-out validation procedure is utilized when the number of

available samples is small (Ren et al., 2009). In this approach, n-1 samples are adopted to train the model and the remaining sample used for validation. The procedure is repeated n times and the root mean square error of cross-validation (RMSECV) serves as the performance parameter. Kooistra et al. (2001) used the leave-one-out approach to validate a PLS model with 69 samples.

Table 6 summary of data use in selected studies

Training/validation sets	Corresponding statistical methods	References
75% (n=81) / 25% (n=27)	PLSR, RF, PSR	(Chakraborty et al., 2015)
70% (n=133) / 30% (n=57)	RF, PSR, ENET	(Chakraborty et al., 2017)
75% (n=225) / 25% (n=75)	Cubist	(Peng et al., 2016)
66% (n=63) / 34% (n=32)	PLSR	(Shi et al., 2014b)
80% (n=96) / 20% (n=24)	PLSR, PCR	(Tayebi et al., 2017)
78%(n=107)/22%(n=30)	PLSR	(Okparanma et al., 2014a)
75% (n=101)/25%(n=34)	PLSR, SVM	(Stazi et al., 2014)
67% (n=50)/33% (n=75)	MLR, BPNN, GA-BPNN	(Zhao et al., 2018)

Acronyms: GA-BPNN= genetic algorithm optimization of back propagation neural network; ENET=elastic net regression; MLR=multiple linear regression; PCR=principal component regression; PLSR=partial least squares regression; PSR=penalized spline regression; RF=random forest regression; SVM=support vector machine;

6.5 Model quality assessment

Model quality assessment is a key process in machine learning. Determination coefficients (R^2), the root mean square error (RMSE), residual prediction deviation (RPD), the ratio of performance to inter-quartile distance (RPIQ), standard error (SE) and bias, can all be used to quantitatively assess model quality (**Table 7**).

The R^2 value is the most widely used parameter for assessing model quality, which is the proportion of the variance in a predicted value (dependent variable) that is predictable from the independent variable. The closer R^2 is to 1, the better the fit of the model. Reported R^2 values in the reviewed literature ranged from 0.11 to 0.99. The RMSE value is the standard deviation of the residuals (prediction errors). The smaller the RMSE, the higher the accuracy of the model. PRD is a goodness-of-fit

parameter that is defined as the standard deviation divided by the RMSE, with values greater than 1.8 considered good (Douglas, Nawar, Alamar, et al., 2018; R. A. V. Rossel, Walvoort, Mcbratney, Janik, & Skjemstad, 2006). PRD values reported in the reviewed studies ranged from 0.51 to 6.23 (Somsubhra Chakraborty et al., 2017; Kemper & Sommer, 2002).

Table 7 Evaluation parameters for determining model quality

Parameters	Equations
r (correlation coefficient)	$\mathrm{r} = \frac{\sum_1^n (y_{i,pre} - \bar{y}_{pre})(y_i - \bar{y})}{\sqrt{\sum_1^n (y_{i,pre} - \bar{y}_{pre})^2}\sqrt{\sum_1^n (y_i - \bar{y})^2}}$
R^2	$R^2 = \frac{\sum_1^n (y_{i,pre} - \bar{y})^2}{\sqrt{\sum_1^n (y_i - \bar{y})^2}} = 1 - \frac{\sum_1^n (y_i - y_{i,pre})^2}{\sqrt{\sum_1^n (y_i - \bar{y})^2}}$
SE (standard error)	$\mathrm{SE} = \sqrt{\frac{1}{n-1}\sum_1^n (y_i - y_{i,pre})^2}$
RMSE (Root mean square effort)	$\mathrm{RMSE} = \sqrt{\frac{\sum_1^n (y_{i,pre} - y_i)^2}{n}}$
RPD	$\mathrm{RPD} = \frac{y_{i,pre} - y_i}{\sqrt{\frac{\sum_1^n (y_{i,pre} - y_i)^2}{n}}}$
RSD (relative standard deviation)	$\mathrm{RSD} = \frac{\mathrm{SD}}{\mathrm{mean}}$
Bias	$\mathrm{Bias} = \sum_1^n \frac{y_{i,pre} - y_i}{n}$

y_i is the observed value of sample i; $y_{i,pre}$ is the predicted value of sample i; $\bar{y}$ is the average of observed value; $\bar{y}_{pre}$ is the average of predicted value.

7 Summary and future research directions

Soil contamination has become a global issue, and sustainable remediation strategies rely upon detailed mapping of soil pollutants (Hou, 2020; Hou et al., 2020a). VIRS combined with machine learning has been identified as a promising approach for detecting soil contamination remotely. Organic contaminants, including TPH and PAH, can be detected by VIRS due to characteristic molecular vibration and stretching (Song et al., 2012; Webster et al., 2016). Heavy metals are detected by

proxy, exploiting their relationships with various soil constituents, including SOM, Fe-oxides and clay minerals. Because soil contaminants can affect plant physiology, vegetation spectra can also be used to predict soil contamination levels (Shi et al., 2014b; Wu et al., 2005).

Proximal, airborne and spaceborne sensors have all been used to collect VIRS spectral data, with the ability to assess large areas in little time (Gholizadeh et al., 2018). After collection, VIRS data requires preprocessing to diminish noise and remove the outliers, which can be achieved with various mathematical methods. Traditional physical sampling is also required for model calibration and validation. Various machine learning algorithms have been used in spectral data interpretation, including regression, neural network, and random forest. These methods can be improved by other advanced algorithms, such as genetic algorithms. However, challenges still exist, and further research is needed in various areas of VIRS based remote sensing in combination with machine learning for soil contamination mapping.

Data collected by VIRS strongly relates to local soil properties (Rathod et al., 2013). The need to calibrate with site specific samples is identified as a big drawback. Libraries of soil spectra collected throughout the world are being established by different institutions (Brown et al., 2006). As these libraries are furnished with greater abundance of spectra data, further research is needed to determine if VIRS analysis can be adequately calibrated with cataloged spectral data when surveying unsampled sites.

VIRS could also serve as a complementary method to small-scale site investigations. The sampling size could be reduced to predict contaminant levels at specific sites. Although models would need to be built for contaminant prediction, this would be attractive due to the low-cost of model building compared to traditional sampling

techniques. A drawback would be that VIRS is limited to surface monitoring, while contaminated land site investigations are often concerned with the subsurface.

Combining VIRS with complementary technology may prove a promising research direction. Recent studies that have lead in this direction include Hu et al. (2017a), who explored a method involving VIRS combined with X-ray fluorescence (XRF) to measure heavy metals rapidly. Xu et al. (2019) also utilized XRF and VIRS, combined with the strategy of outer-product analysis and Granger–Ramanathan averaging, to predict Cd contamination in soil and obtained an acceptable prediction accuracy. Chakraborty et al. (2017) combined Vis-NIR diffuse reflectance spectrometry with geostatistical analysis to identify As hotspots. Additionally, land feature variables correlating to pollutant pathways could be used in combination with spectral data. For instance, topography, land use type, distance to factories could be incorporated.

Although several studies have shown that VIRS could be used for soil contamination mapping, the majority have been conducted in the laboratory (Shi et al., 2014a). Field-based studies have not proved as accurate as those in the laboratory. The presence of vegetation, cloud and moisture can significantly influence the accuracy of VIRS (Shi et al., 2014b). This challenge might be solved by advanced sensors with higher signal-to-noise ratios and more effective spectral data preprocessing and calibration with machine learning algorithms.

Spaceborne spectrometry enables us to achieve long-term temporal and large-scale spatial monitoring of soil health. However, obstacles such as cloud and vegetation coverage need to be dealt with. There have been a limited number of studies that attempted to explore soil contamination through vegetation spectral data, most of those were based on a single vegetation species (Lassalle et al., 2018; Shi et al., 2016).

The information provided by spectral data may vary from both vegetation species and their growth stage, which should be further investigated.

Acknowledgements

This work was supported by National Key Research and Development Program of China (Grant No. 2019YFC1804900), and the Ministry of Ecology and Environment's National Soil Pollution Investigation Project.

References:

Abd El-Halim, A., Omae, H., 2019. Performance assessment of nanoparticulate lime to accelerate the downward movement of calcium in acid soil. Soil Use and Management 35, 683-690.

Abedinia, O., Amjady, N., Ghadimi, N., 2018. Solar energy forecasting based on hybrid neural network and improved metaheuristic algorithm. Computational Intelligence 34, 241-260.

Al Maliki, A., Bruce, D., Owens, G., 2014. Prediction of lead concentration in soil using reflectance spectroscopy. Environmental Technology & Innovation 1-2, 8-15.

Antoine, S., Marco, N., Gergely, T., Luca, M., Bas, v.W., YH, C.H., 2013. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. J Plos One 8.

Antonucci, F., Menesatti, P., Holden, N.M., Canali, E., Giorgi, S., Maienza, A., Stazi, S.R., 2012. Hyperspectral Visible and Near-Infrared Determination of Copper Concentration in Agricultural Polluted Soils. Communications in Soil Science and Plant Analysis 43, 1401-1411.

BERGER, Michael, MORENO, Jose, JOHANNESSEN, Johnny, A., LEVELT, Pieternel, F., HANSSEN, Ramon, F., 2012. ESA's sentinel missions in support of Earth system science. Remote Sensing of Environment 120, 84-90.

Blackburn, G.A.J.R.S.o.E., 1998. Quantifying Chlorophylls and Caroteniods at Leaf and Canopy Scales : An Evaluation of Some Hyperspectral Approaches. 66, 273-285.

Boardman, J., Vandaele, K., Evans, R., Foster, I.D., 2019. Off‐site impacts of soil erosion and runoff: Why connectivity is more important than erosion rates. Soil Use and Management 35, 245-256.

Brewer, R., Peard, J., Heskett, M., 2017. A Critical Review of Discrete Soil Sample Data Reliability: Part 1-Field Study Results. Soil & Sediment Contamination 26, 1-22.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132, 273-290.

Cao, Xiaoming, Wang, Juanle, Yang, Fei, Chen, Xi, Gao, Sciences, Z.J.E.E., 2013. Multiscale remote-sensing retrieval in the evapotranspiration of;Haloxylon ammodendron in the Gurbantunggut desert, China. 69, 1549-1558.

Chabrillat, S., Goetz, A.F.H., Krosley, L., Olsen, H.W., 2002. Use of hyperspectral images in the identification and mapping of expansive clay soils and the role of spatial resolution. Remote Sensing of Environment 82, 431-445.

Chakraborty, S., Weindorf, D.C., Deb, S., Li, B., Paul, S., Choudhury, A., Ray, D.P., 2017. Rapid assessment of regional soil arsenic pollution risk via diffuse reflectance spectroscopy. Geoderma 289, 72-81.

Chakraborty, S., Weindorf, D.C., Li, B., Aldabaa, A.A.A., Ghosh, R.K., Paul, S., Ali, M.N., 2015. Development of a hybrid proximal sensing method for rapid identification of petroleum contaminated soils. Science of the Total Environment 514, 399-408.

Chakraborty, S., Weindorf, D.C., Morgan, C.L., Ge, Y., Galbraith, J.M., Li, B., Kahlon, C.S., 2010. Rapid Identification of Oil-Contaminated Soils Using Visible Near-Infrared Diffuse Reflectance Spectroscopy. Journal of Environmental Quality 39, 1378-1387.

Chen, T., Chang, Q., Clevers, J.G.P.W., Kooistra, L., 2015. Rapid identification of soil cadmium pollution risk at regional scale based on visible and near-infrared spectroscopy. Environmental Pollution 206, 217-226.

Cheng, X., Drozdova, J., Danek, T., Huang, Q., Qi, W., Yang, S., Zou, L., Xiang, Y., Zhao, X., 2018. Pollution Assessment of Trace Elements in Agricultural Soils around Copper Mining Area. Sustainability 10.

Choe, E., Kim, K.W., Bang, S., Yoon, I.H., Lee, K.Y., 2009. Qualitative analysis and

mapping of heavy metals in an abandoned Au–Ag mine area using NIR spectroscopy. Environmental Geology 58, 477-482.

Choe, E., van der Meer, F., van Ruitenbeek, F., van der Werff, H., de Smeth, B., Kim, Y.-W., 2008. Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: A case study of the Rodalquilar mining area, SE Spain. Remote Sensing of Environment 112, 3222-3233.

Douglas, R.K., Nawar, S., Alamar, M.C., Mouazen, A.M., Coulon, F., 2018a. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. Science of the Total Environment 616, 147-155.

Douglas, R.K., Nawar, S., Cipullo, S., Alamar, M.C., Coulon, F., Mouazen, A.M., 2018b. Evaluation of vis-NIR reflectance spectroscopy sensitivity to weathering for enhanced assessment of oil contaminated soils. Science of the Total Environment 626, 1108-1120.

Dutkiewicz, A., Lewis, M., Ostendorf, B., 2009. Evaluation and comparison of hyperspectral imagery for mapping surface symptoms of dryland salinity. International Journal of Remote Sensing 30, 693-719.

Egli, M., Fitze, P., Oswald, M., 1999. Changes in heavy metal contents in an acidic forest soil affected by depletion of soil organic matter within the time span 1969–93. Environmental Pollution 105, 367-379.

Ellis, K., Kerr, J., Godbole, S., Lanckriet, G., Wing, D., Marshall, S., 2014. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiological Measurement 35, 2191-2203.

Forrester, S.T., Janik, L.J., Mclaughlin, M.J., Soriano-Disla, J.M., Stewart, R., Dearman, B., 2013. Total Petroleum Hydrocarbon Concentration Prediction in Soils Using Diffuse Reflectance Infrared Spectroscopy. Soil Science Society of America Journal 77, p. 450-460.

Forzieri, G., Moser, G., Catani, F., 2012. Assessment of hyperspectral MIVIS sensor capability for heterogeneous landscape classification. Isprs Journal of Photogrammetry and Remote Sensing 74, 175-184.

Franke, J., Roberts, D.A., Halligan, K., Menz, G., 2009. Hierarchical Multiple Endmember Spectral Mixture Analysis (MESMA) of hyperspectral imagery for urban environments. Remote Sensing of Environment 113, 1712-1723.

Gholizadeh, A., Kopackova, V., 2019. Detecting vegetation stress as a soil contamination proxy: a review of optical proximal and remote sensing techniques. International Journal of Environmental Science and Technology 16, 2511-2524.

Gholizadeh, A., Saberioon, M., Bendor, E., Borůvka, L., 2018. Monitoring of selected soil contaminants using proximal and remote sensing techniques: Background, state-of-the-art and future perspectives. Critical Reviews in Environmental Science Technology 48, 243-278.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. J Geoderma 89, 1-45.

Götze, C., Beyer, F., Gläßer, C., 2016. Pioneer vegetation as an indicator of the geochemical parameters in abandoned mine sites using hyperspectral airborne data. Environmental Earth Sciences 75, 613.

Guan, Q., Zhao, R., Wang, F., Pan, N., Yang, L., Song, N., Xu, C., Lin, J., 2019. Prediction of heavy metals in soils of an arid area based on multi-spectral data. Journal of Environmental Management 243, 137-143.

Han, W., Gao, G., Geng, J., Li, Y., Wang, Y., 2018. Ecological and health risks assessment and spatial distribution of residual heavy metals in the soil of an e-waste circular economy park in Tianjin, China. Chemosphere 197, 325-335.

Hauser, A., Ali, F., Al-Dosari, B., Al-Sammar, H., Planning, 2013. Solvent-free determination of TPH in soil by near-infrared reflectance spectroscopy. International Journal of Sustainable Development 8, 413-421.

Hillnhuetter, C., Mahlein, A.K., Sikora, R.A., Oerke, E.C., 2011. Remote sensing to detect plant stress induced by Heterodera schachtii and Rhizoctonia solani in sugar beet fields. Field Crops Research 122, 70-77.

Hobley, E., Willgoose, G.R., Frisia, S., Jacobsen, G., 2014. Vertical distribution of charcoal in a sandy soil: evidence from DRIFT spectra and field emission scanning electron microscopy. European Journal of Soil Science

65, 751-762.

Hou, D., 2020. Sustainable Remediation of Contaminated Soil and Groundwater: Materials, Processes, and Assessment. Elsevier Inc.

Hou, D., Bolan, N.S., Tsang, D.C.W., Kirkham, M.B., O'Connor, D., 2020a. Sustainable soil use and management: an interdisciplinary and systematic approach. Science of the Total Environment.

Hou, D., O'Connor, D., Igalavithana, A.D., Alessi, D.S., Luo, J., Tsang, D.C.W., Sparks, D.L., Yamauchi, Y., Rinklebe, J., Ok, Y.S., 2020b. Metal contamination and bioremediation of agricultural soils for food safety and sustainability. Nature Reviews Earth & Environment 1, 366–381.

Hou, D., O'Connor, D., Nathanail, P., Tian, L., Ma, Y., 2017. Integrated GIS and multivariate statistical analysis for regional scale assessment of heavy metal soil contamination: A critical review. Environmental Pollution 231, 1188-1200.

Hou, D., Ok, Y.S., 2019. Speed up mapping of soil pollution. Nature 566, 455-455.

Hou, L., Li, X., Li, F., 2019. Hyperspectral-based Inversion of Heavy Metal Content in the Soil of Coal Mining Areas. Journal of Environmental Quality 48, 57-63.

Hu, B., Chen, S., Hu, J., Xia, F., Xu, J., Li, Y., Shi, Z., 2017a. Application of portable XRF and VNIR sensors for rapid assessment of soil heavy metal pollution. Plos One 12.

Hu, B., Wang, J., Jin, B., Li, Y., Shi, Z., 2017b. Assessment of the potential health risks of heavy metals in soils in a coastal industrial region of the Yangtze River Delta. Environmental Science and Pollution Research 24, 19816-19826.

Huang, Y., Hu, Y., Liu, Y.J.A.E.S., 2009. Heavy metal accumulation in iron plaque and growth of rice plants upon exposure to single and combined contamination by copper, cadmium and lead. 29, 320-326.

Ibrahim, M., Hameed, A.J., Jalbout, A., 2008. Molecular spectroscopic study of River Nile sediment in the greater Cairo region. Applied Spectroscopy 62, 306-311.

Jia, H., Hou, D., O’Connor, D., Pan, S., Zhu, J., Bolan, N.S., Mulder, J., 2020. Exogenous phosphorus treatment facilitates chelation-mediated cadmium detoxification in perennial ryegrass (Lolium perenne L.). Journal of hazardous materials 389, 121849.

Jiang, J., Wua, L., Luo, Y., Liu, L., Zhao, Q., Zhang, L., Christie, P.J.E.J.o.S.B., 2010. Effects of multiple heavy metal contamination and repeated phytoextraction by Sedum plumbizincicola on soil microbial properties. 46, 18-26.

Kemper, T., Sommer, S., 2002. Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. Environmental Science & Technology 36, 2742-2747.

Kooistra, L., Wehrens, R., Leuven, R., Buydens, L.M.C., 2001. Possibilities of visible-near-infrared spectroscopy for the assessment of soil contamination in river

floodplains. Analytica Chimica Acta 446, 97-105.

Kumpiene, J., Lagerkvist, A., Maurice, C., 2007. Stabilization of Pb- and Cu-contaminated soil using coal fly ash and peat. Environmental Pollution 145, 365-373.

Laberge, C., Cluis, D., Mercier, G.J.W.R., 2000. Metal bioleaching prediction in continuous processing of municipal sewage with Thiobacillus ferrooxidans using neural networks. 34, 1145-1156.

Lassalle, G., Credoz, A., Hedacq, R., Fabre, S., Dubucq, D., Elger, A., 2018. Assessing Soil Contamination Due to Oil and Gas Production Using Vegetation Hyperspectral Reflectance. Environmental science & technology 52, 1756-1764.

Liao, K.W., Guo, J.J., Fan, J.C., Lee, C.C., Huang, C.L., 2019. Evaluation of rainfall kinetic energy and erosivity in northern Taiwan using kriging with climate characteristics. Soil Use and Management 35, 630-642.

Liu, F., Liu, X., Zhao, L., Ding, C., Jiang, J., Wu, L., 2015. The Dynamic Assessment Model for Monitoring Cadmium Stress Levels in Rice Based on the Assimilation of Remote Sensing and the WOFOST Model. Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8, 1330-1338.

Liu, K., Zhao, D., Fang, J.-y., Zhang, X., Zhang, Q.-y., Li, X.-k., 2017. Estimation of Heavy-Metal Contamination in Soil Using Remote Sensing Spectroscopy and a Statistical Approach. Journal of the Indian Society of Remote Sensing 45, 805-813.

Liu, M., Liu, X., Wu, M., Li, L., Xiu, L., 2011. Integrating spectral indices with environmental parameters for estimating heavy metal concentrations in rice using a dynamic fuzzy neural-network model. Computers & Geosciences 37, 1642-1652.

Liu, P., Liu, Z., Hu, Y., Shi, Z., Pan, Y., Wang, L., Wang, G., 2019a. Integrating a Hybrid Back Propagation Neural Network and Particle Swarm Optimization for Estimating Soil Heavy Metal Contents Using Hyperspectral Data. Sustainability 11.

Liu, Z., Lu, Y., Peng, Y., Zhao, L., Wang, G., Hu, Y., 2019b. Estimation of Soil Heavy Metal Content Using Hyperspectral Data. Remote Sensing 11.

Martinez-Carvajal, G.D., Oxarango, L., Adrien, J., Molle, P., Forquet, N., 2019. Assessment of X-ray Computed Tomography to characterize filtering media from Vertical Flow Treatment Wetlands at the pore scale. Science of the Total Environment 658, 178-188.

MEE, 2014. the Report on the national general survey of soil contamination. Beijing.

MEE, 2017. Technical guideline on sampling scheme of detailed investigation on soil

pollution in agricultural land.

Ng, W., Malone, B.P., Minasny, B., 2017. Rapid assessment of petroleum-contaminated soils with infrared spectroscopy. Geoderma 289, 150-160.

O'Connor, D., Hou, D., Ok, Y.S., Lanphear, B.P., 2020. The effects of iniquitous lead exposure on health. Nature Sustainability 3, 77-79.

Okparanma, R.N., Coulon, F., Mayr, T., Mouazen, A.M., 2014a. Mapping polycyclic aromatic hydrocarbon and total toxicity equivalent soil concentrations by visible and near-infrared spectroscopy. Environmental Pollution 192, 162-170.

Okparanma, R.N., Coulon, F., Mouazen, A.M.J.E.P., 2014b. Analysis of petroleum-contaminated soils by diffuse reflectance spectroscopy and sequential ultrasonic solvent extraction–gas chromatography. 184, 298-305.

Okparanma, R.N., Mouazen, A.M., 2013. Determination of Total Petroleum Hydrocarbon (TPH) and Polycyclic Aromatic Hydrocarbon (PAH) in Soils: A Review of Spectroscopic and Nonspectroscopic Techniques. Applied Spectroscopy Reviews 48, 458-486.

Okparanma, R.N., Mouazen, A.M.J.W.A., Pollution, S., 2013. Combined Effects of Oil Concentration, Clay and Moisture Contents on Diffuse Reflectance Spectra of Diesel-Contaminated Soils. 224, 1539.

Ourcival, J.M., Joffre, R., Rambal, S.J.N.P., 2010. Exploring the relationships between reflectance and anatomical and biochemical properties in Quercus ilex leaves. 143, 351-364.

Park, J., Li, D., Murphey, Y.L., Kristinsson, J., McGee, R., Kuang, M., Phillips, T., Ieee, 2011. Real Time Vehicle Speed Prediction using a Neural Network Traffic Model.

Pascucci, S., Belviso, C., Cavalli, R.M., Laneve, G., Misurovic, A., Perrino, C., Pignatti, S., 2009. Red mud soil contamination near an urban settlement analyzed by airborne hyperspectral remote sensing.

Patriche, C.V., 2019. Quantitative assessment of rill and interrill soil erosion in Romania. Soil Use and Management 35, 257-272.

Pelta, R., Ben-Dor, E., 2019. Assessing the detection limit of petroleum hydrocarbon in soils using hyperspectral remote-sensing. Remote Sensing of Environment 224, 145-153.

Pelta, R., Carmon, N., Ben-Dor, E., 2019. A machine learning approach to detect

crude oil contamination in a real scenario using hyperspectral remote sensing. International Journal of Applied Earth Observation and Geoinformation 82.

Peng, Y., Kheir, R.B., Adhikari, K., Malinowski, R., Greve, M.B., Knadel, M., Greve, M.H., 2016. Digital Mapping of Toxic Metals in Qatari Soils Using Remote Sensing and Ancillary Data. Remote Sensing 8.

Rathod, P.H., Rossiter, D.G., Noomen, M.F., Fd, V.D.M., 2013. Proximal spectral sensing to monitor phytoremediation of metal-contaminated soils. International Journal of Phytoremediation 15, 405-426.

Ren, H.-Y., Zhuang, D.-F., Singh, A.N., Pan, J.-J., Qiu, D.-S., Shi, R.-H., 2009. Estimation of As and Cu Contamination in Agricultural Soils Around a Mining Area by Reflectance Spectroscopy: A Case Study. Pedosphere 19, 719-726.

Rossel, R.A.V., Adamchuk, V.I., Sudduth, K.A., McKenzie, N.J., Lobsey, C., 2011. PROXIMAL SOIL SENSING: AN EFFECTIVE APPROACH FOR SOIL MEASUREMENTS IN SPACE AND TIME, in: Sparks, D.L. (Ed.), Advances in Agronomy, Vol 113, pp. 237-282.

Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46-54.

Rossel, R.A.V., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. European Journal of Soil Science 63, 848-860.

Rossel, R.V., Behrens, T., Ben-Dor, E., Brown, D., Demattê, J., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V.J.E.-S.R., 2016. A global spectral library to characterize the world's soil. 155, 198-230.

Rosso, P.H., Pushnik, J.C., Mui, L., Ustin, S.L., %J Environmental Pollution, 2005. Reflectance properties and physiological responses of Salicornia virginica to heavy metal and petroleum contamination. 137, 241-252.

Roy, D.P., Wulder, M.A., Loveland, T.R., Woodcock, C.E., Allen, R.G., Anderson, M.C., Helder, D., Irons, J.R., Johnson, D.M., Kennedy, R., 2014. Landsat-8: Science and product vision for terrestrial global change research. Remote Sensing of Environment 145, 154-172.

Sanches, I.D., Filho, C.R.S., Magalhães, L.A., Quitério, G.C.M., Alves, M.N., Oliveira, W.J.J.I.J.o.P., Sensing, R., 2013. Assessing the impact of hydrocarbon leakages on vegetation using reflectance spectroscopy. 78, 85-101.

SC, 2016. Soil pollution control action plan.http://www.gov.cn/zhengce/content/2016-05/31/content_5078377.htm.

Shan, J., Zhao, J., Liu, L., Zhang, Y., Wang, X., Wu, F., 2018. A novel way to rapidly monitor microplastics in soil by hyperspectral imaging technology and chemometrics. Environmental Pollution 238, 121-129.

Shen, Q., Xia, K., Zhang, S., Kong, C., Hu, Q., Yang, S., 2019. Hyperspectral indirect inversion of heavy-metal copper in reclaimed soil of iron ore area. Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy 222.

Shi, T., Chen, Y., Liu, Y., Wu, G., 2014a. Visible and near-infrared reflectance spectroscopy-An alternative for monitoring soil contamination by heavy metals. Journal of Hazardous Materials 265, 166-176.

Shi, T., Guo, L., Chen, Y., Wang, W., Shi, Z., Li, Q., Wu, G., 2018. Proximal and remote sensing techniques for mapping of soil contamination with heavy metals. Applied Spectroscopy Reviews 53, 1-23.

Shi, T., Liu, H., Chen, Y., Wang, J., Wu, G.J.J.o.H.M., 2016. Estimation of arsenic in agricultural soils using hyperspectral vegetation indices of rice. 308, 243-252.

Shi, T., Liu, H., Wang, J., Chen, Y., Fei, T., Wu, G., 2014b. Monitoring Arsenic Contamination in Agricultural Soils with Reflectance Spectroscopy of Rice Plants. Environmental science & technology 48, 6264-6272.

Shuman, L.M., 1982. SEPARATING SOIL IRON-OXIDE AND MANGANESE-OXIDE FRACTIONS FOR MICRO-ELEMENT ANALYSIS. Soil Science Society of America Journal 46, 1099-1102.

Somsubhra, C., Weindorf, D.C., Bin, L., Md Nasim, A., Majumdar, K., ., Ray, D.P., 2014. Analysis of petroleum contaminated soils by spectral modeling and pure response profile recovery of n-hexane. Environmental Pollution 190, 10-18.

Song, Y., Li, F., Yang, Z., Ayoko, G.A., Frost, R.L., Ji, J., 2012. Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang River Delta, China. Applied Clay Science 64, 75-83.

Soriano-Disla, J.M., Janik, L.J., Rossel, R.A.V., 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, 139-186.

Sridhar, B.B.M., Vincent, R.K., Roberts, S.J., Czajkowski, K., 2011. Remote sensing

of soybean stress as an indicator of chemical concentration of biosolid amended surface soils. International Journal of Applied Earth Observation and Geoinformation 13, 676-681.

Stazi, S.R., Antonucci, F., Pallottino, F., Costa, C., Marabottini, R., Petruccioli, M., Menesatti, P., 2014. Hyperspectral Visible-Near Infrared Determination of Arsenic Concentration in Soil. Communications in Soil Science and Plant Analysis 45, 2911-2920.

Sun, W., Zhang, X., 2017. Estimating soil zinc concentrations using reflectance spectroscopy. International Journal of Applied Earth Observation and Geoinformation 58, 126-133.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: A classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences 43, 1947-1958.

Tao, H., Pan, W.L., Carter, P., Wang, K., 2019. Addition of lignin to lime materials for expedited pH increase and improved vertical mobility of lime in no-till soils. Soil Use and Management 35, 314-322.

Tayebi, M., Naderi, M., Mohammadi, J., Tayebi, M.H., 2017. Comparing different statistical models for assessing Fe-contaminated soils based on VNIR/SWIR spectral data. Environmental Earth Sciences 76.

Tian, S., Wang, S., Bai, X., Zhou, D., Luo, G., Wang, J., Wang, M., Lu, Q., Yang, Y., Hu, Z., Li, C., Deng, Y., 2019. Hyperspectral Prediction Model of Metal Content in Soil Based on the Genetic Ant Colony Algorithm. Sustainability 11.

Todorova, M., Mouazen, A.M., Lange, H., Atanassova, S., 2014. Potential of Near-Infrared Spectroscopy for Measurement of Heavy Metals in Soil as Affected by Calibration Set Size. Water Air and Soil Pollution 225.

UNEA, 2018. UNEP/EA.3/L.14 Managing soil pollution to achieve Sustainable Development. United Nations Environment Assembly.

Wang, J., Cui, L., Gao, W., Shi, T., Chen, Y., Gao, Y., 2014. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. Geoderma 216, 1-9.

Wang, L., Li, X., Tsang, D.C., Jin, F., Hou, D., 2020a. Green remediation of Cd and Hg contaminated soil using humic acid modified montmorillonite: immobilization

performance under accelerated ageing conditions. Journal of hazardous materials, 122005.

Wang, L., Ok, Y.S., Tsang, D.C., Alessi, D.S., Rinklebe, J., Wang, H., Mašek, O., Hou, R., O'Connor, D., Hou, D., 2020b. New Trends in Biochar Pyrolysis and Modification Strategies: Feedstock, Pyrolysis Conditions, Sustainability Concerns and Implications for Soil Amendment. Soil Use and Management.

Wang, L., Wu, W.-M., Bolan, N.S., Tsang, D.C.W., Li, Y., Qin, M., Hou, D., 2020c. Environmental fate, toxicity and risk management strategies of nanoplastics in the environment: Current status and future perspectives. Journal of hazardous materials.

Webster, G.T., Soriano-Disla, J.M., Kirk, J., Janik, L.J., Forrester, S.T., McLaughlin, M.J., Stewart, R.J., 2016. Rapid prediction of total petroleum hydrocarbons in soil using a hand-held mid-infrared field instrument. Talanta 160, 410-416.

Wei, L., Yuan, Z., Zhong, Y., Yang, L., Hu, X., Zhang, Y., 2019. An Improved Gradient Boosting Regression Tree Estimation Model for Soil Heavy Metal (Arsenic) Pollution Monitoring Using Hyperspectral Remote Sensing. Applied Sciences-Basel 9.

Workman, J., Workman, J., 2007. Practical guide to interpretive near-infrared spectroscopy.

Wu, Y., Zhang, X., Liao, Q., Ji, J., 2011. Can Contaminant Elements in Soils Be Assessed by Remote Sensing Technology: A Case Study With Simulated Data. Soil Science 176, 196-205.

Wu, Y.Z., Chen, J., Ji, J.F., Gong, P., Liao, Q.L., Tian, Q.J., Ma, H.R., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Science Society of America Journal 71, 918-926.

Wu, Y.Z., Chen, J., Ji, J.F., Tian, Q.J., Wu, X.M., 2005. Feasibility of reflectance spectroscopy for the assessment of soil mercury contamination. Environmental science & technology 39, 873-878.

Xia, X.Q., Mao, Y.Q., Ji, J., Ma, H.R., Chen, J., Liao, Q.L., 2007. Reflectance spectroscopy study of Cd contamination in the sediments of the Changjiang River, China. Environmental Science & Technology 41, 3449-3454.

Xu, D.Y., Chen, S.C., Rossel, R.A.V., Biswas, A., Li, S., Zhou, Y., Shi, Z., 2019. X-ray fluorescence and visible near infrared sensor fusion for predicting soil chromium content. Geoderma 352, 61-69.

Zhang, N., Xu, F., Zhuang, S., He, C., 2016. Monitoring heavy metal Cr in soil based on hyperspectral data using regression analysis, in: Liu, W., Wang, J. (Eds.), Hyperspectral Remote Sensing Applications and Environmental Monitoring and Safety Testing Technology.

Zhang, Y., O'Connor, D., Xu, W., Hou, D., 2020. Blood lead levels among Chinese children: The shifting influence of industry, traffic, and e-waste over three decades. Environment International 135, 105379.

Zhao, L., Hu, Y.-M., Zhou, W., Liu, Z.-H.v., Pan, Y.-C., Shi, Z., Wang, L., Wang, G.-X., 2018. Estimation Methods for Soil Mercury Content Using Hyperspectral Remote Sensing. Sustainability 10.

Zhu, X., Tsang, D.C., Wang, L., Su, Z., Hou, D., Li, L., Shang, J., 2020a. Machine learning exploration of the critical factors for CO2 adsorption capacity on porous carbon materials at different pressures. Journal of Cleaner Production 273, 122915.

Zhu, X., Wan, Z., Tsang, D.C., He, M., Hou, D., Su, Z., Shang, J., 2020b. Machine learning for the selection of carbon-based materials for tetracycline and sulfamethoxazole adsorption. Chemical engineering journal, 126782.

Zuzolo, D., Cicchella, D., Albanese, S., Lima, A., Zuo, R., De Vivo, B., 2018. Exploring uni-element geochemical data under a compositional perspective. Applied Geochemistry 91, 174-184.

Highlight

VIRS based detection serves as a sustainable way in mapping soil pollution.

The combination of machine learning enables VIRS to provide an accurate result.

Heavy metals and organic pollutants in soil can be monitored this way.

Fe-oxide, clay minerals and soil organic matters are influential factors.

Field-based study is requisite to improve this method.

Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: