
https://doi.org/10.1016/j.geoderma.2021.115512

Institional Repository GFZpublic: https://gfzpublic.gfz-potsdam.de/
Estimating heavy metal concentrations in Technosols with reflectance spectroscopy

Friederike Kästner¹, Magdalena Sut-Lohmann², Shaghayegh Ramezany², Thomas Raab², Hannes Feilhauer³,⁴, Sabine Chabrillat¹,⁵

¹Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Section Remote Sensing, Telegrafenberg, Potsdam 14473, Germany
²Brandenburg University of Technology Cottbus-Senftenberg, Chair of Geopedology and Landscape Development, Siemens-Halske-Ring 8, 03046 Cottbus, Germany
³Leipzig University, Remote Sensing Center for Earth System Research, Remote Sensing in Geo- and Ecosystem Research group, Talstr. 35, 04103 Leipzig, Germany.
⁴Department of Remote Sensing, Helmholtz-Centre for Environmental Research - UFZ, Permoserstr. 15, 04318 Leipzig, Germany.
⁵Leibniz University Hannover, Institute of Soil Science, Herrenhäuser Straße 2, 30419 Hannover, Germany.

Contact: friederike.kaestner@gfz-potsdam.de

Abstract

Reflectance spectroscopy in the visible-infrared and shortwave infrared (450-2500 nm) wavelength region is a rapid, cost-effective and non-destructive method that can be used to monitor heavy metal (PTE, potential toxic elements) contaminated areas. Due to the PTE pollution that has accumulated in the course of wastewater treatment, the existence of Technosols presents an environmental problem, a potential source for PTE uptake by vegetation, or even the release of PTEs into groundwater. In this study, multivariate procedures using Partial Least Squares Regression (PLSR) and Random Forest Regression (RFR) are applied to quantify relationships between soil heavy metal concentration (Cr, Cu, Ni, Zn) and reflectance data of highly contaminated Technosols from a former sewage farm near Berlin, Germany. Laboratory measurements of 110 soil samples in four different preparation steps
were acquired with HySpex hyperspectral cameras. The impact of the different preparation steps, namely “oven-dried”, “sieved”, “ground”, “LOI”, was evaluated for its potential to enhance the method performance or to reduce the time-consuming soil sample preparation. Furthermore, different spectral pre-processing methods were evaluated regarding improvements of spectral modelling performance and their ability to minimise noise and multiple scattering effects. Considering the optimal coefficient of determination ($R^2$), PLSR shows an improving performance and accuracy with increasing preparation steps such as ground or LOI for all metals of interest ($R^2_{\text{Cr}}$: 0.52-0.78; $R^2_{\text{Cu}}$: 0.36-0.73; $R^2_{\text{Ni}}$: 0.19-0.42 and $R^2_{\text{Zn}}$: 0.41-0.74). RFR shows a weaker estimation performance for all metals, even when using higher sample preparation levels ($R^2_{\text{Cr}}$: 0.36-0.62; $R^2_{\text{Cu}}$: 0.17-0.72; $R^2_{\text{Ni}}$: 0.20-0.35 and $R^2_{\text{Zn}}$: 0.26-0.67). The results show that an application of methods such as PLSR for the prediction of PTE concentration in Technosols is still a challenge but provides more robust estimations than the user-friendly RFR method. Additionally, this study shows that PTE estimation performance in heterogeneous soil samples can be improved by increased laboratory soil preparation steps and further spectral pre-processing steps.

Keywords
PTE, Reflectance Spectroscopy; Partial Least Squares (PLS) regression; Sewage farm, Random Forest regression, soil environment monitoring

1. Introduction

Natural or anthropogenic activities, such as sewage farming, mining, or human impact, can result in soil contamination and an increased presence of potential toxic elements (PTE). This causes long lasting disturbances of important soil functions for the abiotic and biotic environment, such as filtering, buffering, or plant growth (European Environment Agency, 2020). Specific remediation methods to monitor soil PTE contamination, including laboratory analysis with numerous soil samples, are time consuming and cost-intensive (Cheng et al., 2019; Kemper and Sommer, 2002; Schwartz et al., 2011). Thus, rapid, easy-to-use, and reasonably priced monitoring and controlling systems for soil remediation processes in PTE
contaminated Technosols are urgently needed in order to provide a wider monitoring and remediation application.

Reflectance spectroscopy is widely used and a well-known technique for the determination of soil properties (Ben-Dor et al., 2009; Stenberg and Rossel, 2010). Several pioneering studies have evaluated reflectance spectroscopy for the monitoring of soil PTEs being a fast, cost-effective, and non-destructive method with a minimum sample preparation, which can provide feasible real-time measurements (Cheng et al., 2019; Kemper and Sommer, 2002; Schwartz et al., 2011). As a result, many recent studies used visible and near infrared reflectance spectroscopy for soil PTE estimation in agricultural soils (Wang et al., 2014), suburban soils (Cheng et al., 2019; Wu et al., 2005), and soils influenced by mining activities (Gholizadeh et al., 2015b; Kemper and Sommer, 2002) and sewage sludges (Soriano-Disla et al., 2010). In addition, e.g., Soriano-Disla et al. (2010) and Reed et al. (2011) investigated sewage sludges based on stability parameters related to microbial respiration or anaerobic digestion. However, to date, applications of reflectance spectroscopy to soil PTE estimations in Technosols with strong heterogenic soils and high PTE concentration and variability are still missing.

Soil spectra show specific spectral fingerprints depending on soil compositions such as soil organic matter (SOM) and various PTEs (Gholizadeh et al., 2018a; Viscarra Rossel et al., 2006). However, the spectral features of PTEs are not directly observable as their spectral response is obscured by other features in the spectrum and PTEs form a complex with organic matter (OM) and other soil attributes, such as carbonates, hydroxides, oxides, or clay minerals (Gholizadeh et al., 2018a; Rathod et al., 2013; Schwartz et al., 2011).

A feasible method to assess PTE concentrations in soils is the use of prediction models, such as Partial Least Squares Regression (PLSR) (Pandit et al., 2010; Wang et al., 2014; Wu et al., 2005; Zhang et al., 2010) and Random Forest Regression (RFR) (Tan et al., 2019; Wei et al., 2019). The parametric PLSR is robust to data noise and missing values and is able to cope with multicollinearities between variables (Gholizadeh et al., 2013), whereas the non-parametric RFR is a more simple and user-friendly method needing fewer decisions on the
model parametrisation (Feilhauer et al., 2015). However, current applications of PLSR and RFR models merely focus on agriculture soils (Tan et al., 2019; Wang et al., 2014), suburban soils (Cheng et al., 2019; Wu et al., 2005), and mining areas (Gholizadeh et al., 2015b; Sun et al., 2019), but not on Technosols.

To the best of our knowledge, highly heterogeneous soil samples with a high PTE content variability in Technosols from former sewage farms have not yet been used to estimate PTE concentration using different preparation steps with reflectance spectroscopy. In our study, we propose a novel protocol for the analysis and estimation of Cr, Cu, Ni and Zn by evaluating the reflectance spectra of different soil preparation steps. Additionally, we combined spectral pre-processing steps and different prediction models for all sample preparations. Obtaining accurate PTE prediction in Technosols with a well-known imaging technique would support future mapping and monitoring of PTE concentration over highly contaminated and heterogeneous areas, which is still challenging.

2. Materials and methods

2.1 Study area and soil sampling

The former sewage farm “Deutsch Wusterhausen” was chosen as the study area. It is located in the South of Berlin and owned by Berliner Stadtgüter GmbH. The geological setting - Pleistocene till resulting in silt and sandy sediments - provided optimal conditions for sewage farms with its well-drained soils (Brandenburg, 2010). The long-term irrigation in the years between 1903 and 1994 influenced the chemical and physical soil properties especially in the former man-made sedimentation basin. Here, untreated sewage sediments and floating solids were deposited from industrial and domestic sources, resulting in Technosols with partially thick sewage sludge layers, high contamination with PTEs, and organic pollutants (LUA, 2003; Ritschel and Kratz, 2000).
Sampling was conducted at one of the former sedimentation basins (42 m × 50 m), as shown in Figure 1. A pre-analysis showed that the highest PTE concentrations are found between 15 cm and 20 cm depth, where the sewage sludge layer exists. Thus, in a 5 m × 5 m grid, disturbed soil samples were collected at a depth of 15-20 cm using a 250 ml stainless steel soil sample cylinder. These samples were placed and transported in plastic bags.

Figure 1: Sampling location and methodology: A) Spatial distribution map of the sample locations in Germany, Berlin. The underlying hillshade of the terrain surface display the sedimentation basins, and B) including the sample strategy.

2.2 Soil preparation and chemical analyses

All soil samples were oven-dried for two days at 40 °C, sieved to a fraction < 2 mm and ground to a fraction < 10 µm with a Mortar Grinder RM 200 for further laboratory analyses. Elemental concentrations were determined using microwave plasma atomic emission spectrometry.
(Agilent Technologies 4210 MP-AES) after digestion with nitric acid (65 %). Additionally, a sequential extraction was performed to determine the absorption affinity for each PTE into five fractions: exchangeable, bound to carbonates, bound to Fe/Mn oxides, bound to organic matter, and residual. For each fraction, a modified Tessier method was used, followed by AES analysis. The organic matter was calculated with LOI, following Sut et al., 2012.

2.3 Spectral Measurements

During preparation, a small amount of the sample of each step of the sample preparation were set aside. The preparation steps are defined as (stage 1) oven dried 40 °C, (stage 2) sieved < 2mm, (stage 3) ground < 10 μm, and (stage 4) after LOI treatment. The soil samples in the phases related to these four different preparation steps were analysed for their reflectance properties in the laboratory.

Imaging spectrometer data of the soil samples were acquired using the HySpex VNIR-1600 and HySpex SWIR 320m-e (HySpex, Norsk Elektro Optikk, 2015) hyperspectral sensors. The line scanners were mounted over a translational rack (nadir), allowing the line-wise movement of the samples for the measurement (Figure 2). The HySpex spectral range of both sensors covers 450 nm to 2500 nm with a spectral sampling interval of 3.7 nm for the VNIR sensor and 6.25 nm for the SWIR sensor. The spectral measurements were taken under spectral laboratory conditions in a dark room with an artificial light source (halogen bulb with diffusor) installed at 45° (zenith viewing angle) and both sensors were installed in nadir view above the samples. The arrangement of the four different preparation steps of the soil sample of one sampling position is shown in Figure 2B and Figure 2C. At-surface reflectance retrieval of the radiance data followed the GeoMAP-Trans approach (Rogass et al., 2017). The reflectance data were used to build a spectral library using one spectrum for each sample and each associated preparation step composed of an average of 10 × 10 pixels centred in the sample surface.
Figure 2: Laboratory spectral acquisition: A) Translation imaging spectroscopy stage, modified after Körting (2019); B) Photo of the samples during acquisition; and C) hyperspectral image in RGB true colour with a 10 × 10 pixel example using for the spectral library (red square).

2.4 Spectral data pre-processing and PTE modelling

The 110 spectra all passed the same pre-processing steps: (1) Savitzky-Golay (SG) smoothing with a second-order polynomial and a window size of 41 bands was applied for noise-reduction.
(Sun et al., 2019). (2) An outlier detection using Principle Component Analysis (PCA) and the Mahalanobis distance method was applied, and the outliers were removed from further analyses to increase prediction accuracy (Gholizadeh et al., 2015b; Liu et al., 2018). (3) The first order derivative (FD) using a Savitzky-Golay smoothing with a window size of 41 data points and a second-order polynomial fit was applied to remove baseline drifts, linear trends, and instrument noises (Gholizadeh et al., 2015b; Wang et al., 2014; Wu et al., 2005). (4) A multiplicative scatter correction (MSC), standard normal variate (SNV) and spectra normalisation by the mean (NME) calculation were applied to deal with multiple scattering effects due to different grain sizes (Ahmad, 2015; Wang et al., 2014; Wu et al., 2005). The major water absorption bands between 1333 - 1639 nm and 1837 - 2222 nm were removed for all pre-processing steps and for the FD pre-processing step (3) between 1303 - 1609 nm as well as between 1813 - 2053 nm for further analyses.

We applied two commonly used regression techniques in quantitative soil spectroscopy, PLSR and RFR, to establish a relationship between soil spectra and soil PTE concentration for each separated soil preparation step. Their respective performance can be compared through the cross-validation coefficient of determination ($R^2$) and the root-mean-square error (RMSE), which describe the goodness of model fitting and is used to assess the best model. Additionally, the ratio of the performance deviation (RPD) was calculated to consider the PTE variation (Cheng et al., 2019; Gholizadeh et al., 2015a; Tan et al., 2019). An ‘excellent’ prediction is considered for results with high $R^2$ (> 0.90), low RMSE and an RPD larger than 3.0. Indicative for a ‘good’ prediction are $R^2$ values between 0.82 and 0.90 and RPD values between 2.5 and 3.0. $R^2$ and RPD values from 0.66 to 0.81 and 2.0 to 2.5, respectively, reveal an ‘approximate quantitative’ prediction. Values between 0.50 and 0.65 ($R^2$) and between 1.5 and 2.0 (RPD) indicate that the model can distinguish between high and low values. ‘Unsuccessful’ prediction reveals values of $R^2$ and RPD lower than 0.5 and 1.5, respectively (Cheng et al., 2019; Gholizadeh et al., 2015a; Vohland et al., 2011).

The measured spectra with the associated PTE concentration were imported into a Python environment for further analysis, using modified PLSR and RFR codes based on the NIRPY
Research Blog module by Daniel Pelliccia were used for modelling (Pelliccia, 2020). The PLSR model was optimized with a 10-fold cross-validation approach. First, the optimal latent variables with the smallest RMSE was determined and, second, a user-defined threshold (0.60) band selection with the highest coefficients values was applied to identify important bands (Feilhauer et al., 2015). The maximum latent variable was set to 10 % of the sample size (LV = 6) to avoid overfitting of the model (Kooistra et al., 2001). With the optimal parameter selection, we used the second model to assess the prediction accuracy using another 10-fold cross validation (final \( R^2 \) and RMSE).

The RFR approach split the dataset into an independent and randomly selected subset (decision trees). The final result is the average of all the decision trees in a random forest to reduce over-fitting and to improve model accuracy (Wei et al., 2019). Because fewer parametrisation decisions on the model are required, just the number of trees in the forest \((ntree)\), maximum depth of the tree \((mdepth)\), and randomly chosen input variables at each tree at the root node \((nodes)\) needed to be specified here. The RFR is based on a two-run approach: The first run optimizes the hyper-parameters using a 10-fold cross validation approach on the test set. Here, the range of \(ntree\) (50, 60, 70, ..., 1000), of \(mdepth\) (3, 7, 50, 80, 90, 100) and \(nodes\) (3 - 8) were implemented in the model to find the optimal model parameter with the smallest RMSE (Tan et al., 2019). The second run assesses the final prediction accuracy using the optimal model parameter.

3. Results

3.1 Soil Chemical Analyses

Descriptive statistics of all PTE concentration and selected soil properties (SOM, pH) are summarized in Table 1. Here, from the 110 samples, one to five samples were removed due to possible chemical measurement errors and unusual spectral behaviour. The analysis show that the mean concentration of all PTE and even the minimum concentration of Zn, Cu, and Cr are exceeding the German preventative values (GPV) of the Regulation of the Federal Soil
Protection and Contamination Ordinance for sandy soils (§ 8 Abs. 2 Nr. 1 BBodSchV),

classifying these soils as highly contaminated. The pH values range from 6.5 to 7.2 and

indicate slightly acidic to neutral soils. The mean soil organic matter content over all samples

is 34 % and ranges between 7 % and 49 % with a strong negative-skewed distribution into

higher SOM content. The PTE concentration shows a high variation in the data with a strong

right-skewed distribution towards higher concentration.

Table 1: Descriptive statistics of soil heavy metal concentration (mg/kg) for nickel (Ni), zinc
( Zn), copper (Cu) and chromium (Cr) and soil parameters (soil organic matter (SOM), pH).

Abbreviations: SD (Standard Deviation), GPV (German Preventative Values).

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Ni [mg/kg]</th>
<th>Zn [mg/kg]</th>
<th>Cu [mg/kg]</th>
<th>Cr [mg/kg]</th>
<th>SOM [%]</th>
<th>pH</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNT</td>
<td>97</td>
<td>95</td>
<td>96</td>
<td>95</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>MEAN</td>
<td>132.5</td>
<td>4703.4</td>
<td>1831.8</td>
<td>688.1</td>
<td>34</td>
<td>6.9</td>
</tr>
<tr>
<td>SD</td>
<td>48.0</td>
<td>1262.9</td>
<td>518.6</td>
<td>205.4</td>
<td>10.7</td>
<td>0.2</td>
</tr>
<tr>
<td>MAX</td>
<td>226</td>
<td>8713</td>
<td>2983</td>
<td>966</td>
<td>49</td>
<td>7.2</td>
</tr>
<tr>
<td>MIN</td>
<td>10</td>
<td>1144</td>
<td>425</td>
<td>75</td>
<td>7</td>
<td>6.5</td>
</tr>
<tr>
<td>SKEWNESS</td>
<td>-0.9</td>
<td>-0.7</td>
<td>-0.7</td>
<td>-1.2</td>
<td>-1.0</td>
<td>-0.7</td>
</tr>
<tr>
<td>GPV</td>
<td>15</td>
<td>60</td>
<td>20</td>
<td>30</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2 Spectral Measurements

Reflectance soil sample spectra from all four preparation steps with the corresponding mean

spectra are illustrated in Figure 3. All soil samples for each stage show similar reflectance

spectra. The first three stages show similar spectral shape and features with typical OH-,

H₂O-, and OM- absorption features at 1400 nm, 1900 nm, 2306 nm, and 2348 nm. The

reflectance spectra of stage 4 had weak absorption peaks around 900 nm. This region shows

the presence of iron oxides, bounded on with other metal cations or hydroxyl groups

(Gholizadeh et al., 2015a). The stages 1, 2, and 4 show a higher spectral variability, while

the individual spectra in stage 3 are distributed close around the mean spectra. The water

absorption features at 1400 nm and 1900 nm decrease with increasing preparation steps and

loss of water, that is, being most pronounced in stage 1 and least pronounced in stage 4.

The shape and slope of the spectra from stage 1 to 3 did not show any noticeable

differences compared to stage 4. Spectra based on the last preparation step demonstrate

higher reflectance values between 0.2 and 0.6 relative to the other preparation steps with a
maximum of 0.4 and 0.5.

Figure 3: All soil sample spectra with the associated average spectra of the four stages (orange) for stage 1 (oven-dried 40°C), stage 2 (sieved < 2 mm), stage 3 (ground < 10 µm) and stage 4 (LOI).

3.3 Spectral data pre-processing and PTE modelling

PLSR and RFR models were generated based on the different data pre-processing methods for each stage. PLSR and RFR modelling was first performed on the original, smoothed dataset and subsequently on the FD dataset with SNV, MSC and NME correction in each case. The results of both models with the optimal spectral pre-processing, the model parameters, and discarded wavelength bands are summarized in Table 2 and Table 3 (regression plots are summarized in Figures S 3 and S 4 of the supplementary materials). The pre-processing methods MSC and SNV in connection with the FD show the most accurate estimation
performance for all PTEs, according to the minimal RMSE and optimal $R^2$ values for both
regression techniques. However, for PLSR predictions of Zn (stage 1 and 3), Cr (stage 3), and
Ni (stage 4), a pre-processing based on SG smoothing resulted in a more accurate fit. Along
the different preparation steps (stage 1 to 4), there is no specific pre-processing pattern, which
would always lead to the best model performance. The PLSR prediction of Ni and Zn (stage
1), as well as Ni for stage 3 and Cr for stage 4 have to be interpreted with caution, because
the relationship between the bias of predicted and observed concentration with the observed
PTE concentration shows no homoscedasticity, therefore indicates a high multicollinearity in
the data (see Figures S 1 and S 2 of the supplementary materials).

Table 2: PLSR results with the optimal pre-processing (PP) for heavy metal concentration
(mg/kg). Abbreviations: PTE (potential toxic elements), $R^2$ (coefficient of determination),
RMSE (root-mean-square error), RPD (ratio of performance deviation), LV (latent variable),
DWB (discarded wavebands).

<table>
<thead>
<tr>
<th>Stage</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cr</td>
<td>Cu</td>
<td>Ni</td>
<td>Zn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.52</td>
<td>0.66</td>
<td>0.78</td>
<td>0.67</td>
<td>0.36</td>
<td>0.46</td>
<td>0.66</td>
<td>0.73</td>
<td>0.19</td>
<td>0.37</td>
<td>0.21</td>
<td>0.42</td>
<td>0.41</td>
<td>0.54</td>
<td>0.59</td>
<td>0.74</td>
</tr>
<tr>
<td>RMSE [%]</td>
<td>13.4</td>
<td>12.4</td>
<td>9.3</td>
<td>13.3</td>
<td>12.7</td>
<td>12.9</td>
<td>9.5</td>
<td>10.1</td>
<td>18</td>
<td>17.3</td>
<td>17.6</td>
<td>19.5</td>
<td>10</td>
<td>9.2</td>
<td>8.6</td>
<td>7.9</td>
</tr>
<tr>
<td>RPD</td>
<td>1.4</td>
<td>1.7</td>
<td>2.1</td>
<td>1.7</td>
<td>1.3</td>
<td>1.4</td>
<td>1.7</td>
<td>1.9</td>
<td>1.1</td>
<td>1.3</td>
<td>1.1</td>
<td>1.3</td>
<td>1.3</td>
<td>1.5</td>
<td>1.6</td>
<td>2.0</td>
</tr>
<tr>
<td>LV</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>DWB</td>
<td>216</td>
<td>269</td>
<td>229</td>
<td>219</td>
<td>226</td>
<td>272</td>
<td>190</td>
<td>171</td>
<td>226</td>
<td>75</td>
<td>283</td>
<td>203</td>
<td>249</td>
<td>182</td>
<td>106</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: RFR results with the optimal pre-processing (PP) for heavy metal concentration
(mg/kg). Abbreviations: PTE (potential toxic elements), $R^2$ (coefficient of determination),
RMSE (root-mean-square error), RPD (ratio of performance deviation), ntree (number of
trees in the forest), mDepth (maximum depth of the tree).

<table>
<thead>
<tr>
<th>Stages</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cr</td>
<td>Cu</td>
<td>Ni</td>
<td>Zn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.36</td>
<td>0.62</td>
<td>0.58</td>
<td>0.37</td>
<td>0.17</td>
<td>0.58</td>
<td>0.59</td>
<td>0.72</td>
<td>0.20</td>
<td>0.29</td>
<td>0.20</td>
<td>0.35</td>
<td>0.26</td>
<td>0.52</td>
<td>0.44</td>
<td>0.67</td>
</tr>
<tr>
<td>RMSE [%]</td>
<td>15.5</td>
<td>12.9</td>
<td>12.8</td>
<td>15.3</td>
<td>14.7</td>
<td>11.9</td>
<td>10.4</td>
<td>11.5</td>
<td>17.7</td>
<td>18.2</td>
<td>17.2</td>
<td>20.9</td>
<td>11.2</td>
<td>10</td>
<td>10.3</td>
<td>9.9</td>
</tr>
<tr>
<td>RPD</td>
<td>1.2</td>
<td>1.7</td>
<td>1.5</td>
<td>1.3</td>
<td>1.1</td>
<td>1.5</td>
<td>1.6</td>
<td>1.7</td>
<td>1.3</td>
<td>1.2</td>
<td>1.5</td>
<td>1.2</td>
<td>1.2</td>
<td>1.4</td>
<td>1.3</td>
<td>1.6</td>
</tr>
<tr>
<td>ntree</td>
<td>50</td>
<td>155</td>
<td>366</td>
<td>261</td>
<td>261</td>
<td>50</td>
<td>155</td>
<td>155</td>
<td>50</td>
<td>50</td>
<td>155</td>
<td>155</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>mDepth</td>
<td>7</td>
<td>7</td>
<td>80</td>
<td>3</td>
<td>80</td>
<td>3</td>
<td>90</td>
<td>3</td>
<td>80</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>80</td>
<td>90</td>
<td>50</td>
</tr>
<tr>
<td>nodes</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>PP</td>
<td>MSC</td>
<td>FD &amp; MSC</td>
<td>FD</td>
<td>MSC</td>
<td>FD &amp; SNV</td>
<td>FD</td>
<td>MSC</td>
<td>FD &amp; SNV</td>
<td>FD</td>
<td>SNV</td>
<td>FD</td>
<td>SNV</td>
<td>FD</td>
<td>SNV</td>
<td>FD</td>
<td>SNV</td>
</tr>
</tbody>
</table>
4. Discussion

4.1 Spectral responses and important bands in PTE modelling

The absorption features at 1400 nm and 1900 nm are associated to bound hydroxyl groups and water (Gholizadeh et al., 2015a). Weak absorption features between 2000 nm and 2250 nm are related to bound C-H groups related to soil organic matter (SOM), carbonates, lignin and humic acid (Ahmad, 2015; Kooistra et al., 2001; Rathod et al., 2013). Although the high SOM content in the samples are caused by the high amount of organic components in the wastewater (Chen et al., 2015; LUA, 2003). The stage 4 spectra (Figure 3) show weak absorption peaks around 900 nm that are associated with the spectral activity of iron oxide (e.g., the electronic transition of iron in the mineral goethite and hematite) (Gholizadeh et al., 2015a; Wu et al., 2005).

The spectra of the first three stages for all PTEs (Figure 3) indicate important mineralogical spectral bands that can be associated with iron oxides, soil organic matter, secondary clay minerals, and carbonate absorptions (Gholizadeh et al., 2015a). Additionally, stages 1 and 2 indicate important bands between 1010 nm and 1215 nm, as well as between 1650 nm and 1850 nm. The prediction model for stage 4 for the PTEs focuses more on the VNIR region, except for Ni and Zn. Here, there are few bands identified as important associated with SOM, secondary clay minerals, or carbonates. However, SOM is removed from the samples in stage 4, the important bands are linked to the absorption features associated with carbonates (near 2300 nm) and with secondary clay minerals (near 2200 nm) (Chabrillat et al., 2002; Kemper and Sommer, 2002; Shin et al., 2019). Previous studies investigated similar important bands for Cu, Zn, and Cr (Cheng et al., 2019). Comparing the results in addition to correlations between soil PTE concentration and reflectance values, all important wavelengths indicate that spectral active soil features play an crucial role in the prediction of PTE with reflectance spectroscopy (Cheng et al., 2019; Kooistra et al., 2001; Shin et al., 2019; Wu et al., 2007). Wu et al. (2007) investigate high correlations between the siderophile elements Cr and Ni and the
iron oxides (480 nm and 510 nm). The relationship between Cu and Zn and spectral bands shows a negative correlation, which is linked to the ferrous absorption feature around 700 nm and 1400 nm (Kemper and Sommer, 2002; Shin et al., 2019) and other soil mineral compositions, such as AlOH-Hydroxide at 2100 nm (Shin et al., 2019).

4.2 Prediction modelling of individual PTEs

Chromium

The prediction accuracies for Cr increase with increasing preparation steps ($R^2 = 0.52$ for stage 1 to $R^2 = 0.78$ for stage 3, Table 2) but decreases for stage 4 with a $R^2 = 0.67$. The overall best performance of the model was obtained with stage 3. Here, almost 80 % of the variance in the Cr concentration is explained by the spectra and a differentiation of high and low concentrations. Moreover, an “approximate quantitative” prediction can be made. Considering the RPD = 2.1, a quantitative prediction can be made with stage 3. Many previous studies reported $R^2$ values between 0.78 and 0.86 for different pre-processing methods and also a $R^2$ value of 0.89 for PLSR models (Cheng et al., 2019; Tan et al., 2019; Wu et al., 2007). In comparison, the estimation performance in our study was much better for a comparable preparation step. The $R^2$ values of the previous studies based on soil samples from agricultural or suburban soils rather than from inhomogeneous soil samples of former sewage farms. The lower performance of stage 4 is probably caused by the above mentioned heteroscedasticity of the data (Figure S 2). The measured concentration and the bias show almost a linear trend at higher Cr concentration. Another reason could be the removal of SOM from the samples. A comparison of the scatter plots for stage 3 and 4 (Figure S 3) reveals a stronger variation between the measured and predicted concentrations in stage 4. Two of the measured concentrations show much higher predicted concentrations, which is caused by the brighter soil sample after LOI.

For the RFR method, the prediction accuracy for stage 2 ($R^2 = 0.62$, RPD = 1.7) and stage 3 ($R^2 = 0.58$, RPD = 1.5) showed the best performance. The high performance of stage 2 implies that the FD in combination with MSC is the most suitable pre-processing method, by removing
baseline drifts and spectral noises. Studies from Tan et al. (2019) and Cipullo et al. (2019) report fits of $R^2 = 0.90$ and 0.76 for predictions of Cr contents. Contents, which are comparable with stage 3 and stage 2 in our study, respectively. The prediction accuracy of RFR is much higher for the results in Tan et al. (2019), whereas the prediction performance in Cipullo et al. (2019) is close to the performance observed in our study with $R^2 = 0.63$ with stage 2. The similar performance at stage 2 for RFR is most likely a result of RFR being a non-parametric method, which deals better with heterogeneous data, thus, with non-linear relationships compared to PLSR ($R^2 = 0.67$ at stage 4) (Wang et al., 2020).

Copper

For Cu, advanced preparation steps showed increasing $R^2$ values and an enhancement in prediction accuracy for PLSR. An RPD = 1.9 indicates the possibility to “distinguish between high and low” PTE values and is very close to a “good prediction” ($R^2$= 0.73) (Gholizadeh et al. (2015a)). Studies from Cheng et al. (2019), Gholizadeh et al. (2015b), and Wang et al. (2014) report $R^2$ values of 0.26, 0.50, and 0.58 - 0.69, respectively. Their preparation steps are similar to stage 2 and stage 3. Sun et al. (2019) report a similar $R^2$ value of 0.78, comparable to stage 4, by using just spectra associated with iron oxides. Further spectral pre-processing techniques connected with higher-level preparation steps improve the overall estimation performance for Cu in the PLSR approach. The reason for this could be the combination of removing the spectral baseline offset and artificial noise and removal of the interfering soil attributes, such as SOM. The influence of SOM agrees with the results of the sequential analysis. Cu is strongly bound to SOM and, thus, organic components may influence the behaviour of the metal (Adriano, 2001). By removing SOM in stage 4, this may lead to a spectral enhancement of Cu, and the remaining Cu is presumably bound to soil minerals including Fe/Mn oxides. In contrast to the dominant wavelengths for stage 4, the results reveal a strong bond to iron oxides, which agree with the findings of Sun et al. (2019). A possible reason could be that the Cu adsorption maxima override the SOM complexation, if clay
minerals and Fe oxides are dominant in the soil (Adriano, 2001). This also affects the performance of the RFR model. The RFR showed an improvement in estimation performance with increasing preparation steps with an $R^2 = 0.72$ and an RPD = 1.7. In this case, the same preparation steps but additional spectral pre-processing are required compared to PLSR. Cipullo et al. (2019) report an $R^2 = 0.60$ using RFR for Cu prediction for a comparable preparation step to stage 2 and a SG-FD spectral pre-processing. Enhancing the preparation steps for Cu estimation leads to a higher model performance and, thus, an approximation of a good prediction.

Nickel

The prediction models for Ni resulted in relatively low $R^2$ values from 0.19 to 0.42 and RPD values from 1.1 to 1.3 for all stages. The PLSR model for Ni prediction was 'unsuccessful' and cannot discriminate between high and low Ni concentrations (Gholizadeh et al., 2015a; Vohland et al., 2011; Wang et al., 2014). However, previous studies indicate much higher $R^2$ values of 0.93 (comparable to stage 3 preparation; Wu et al. (2005)) and 0.81 (comparable to stage 3 preparation; Wu et al. (2007)). In our study, Ni is bound to Fe/Mn oxides and carbonates, which might explain the better estimation performance in stage 4 for PLSR. Compared to the above-mentioned studies, the overall poor performance of Ni estimation with PLSR is possibly caused by the different mineralogy of the analysed soil samples. Comparing with the RFR method, stage 2 and stage 4 reveals a similar performance behaviour but a weaker prediction accuracy of 0.29 and 0.35, respectively. All stages show heteroscedasticity and must be interpreted with caution. Cipullo et al. (2019) report a $R^2 = 0.77$ for stage 2 and a much stronger performance of estimating Ni than the present study observed. Comparing the trend analysis of the measured to predicted concentration (PLSR: S 3 and RFR: S 4), our results show a high variance. Low Ni concentrations are overestimated in all stages for the PLSR model. RFR also tends to overestimate towards lower Ni concentrations. It is
therefore likely that the low concentration of Ni with 132.53 mg/kg and a low standard derivation of 48 mg/kg combined with the heterogeneous dataset lead to the low performance of PLSR and RFR. This agrees with the low RPD values, which consider the standard variation of the PTE content (Vohland et al., 2011). Brokamp et al. (2017) also show the lowest model performance with low Ni concentration and lowest standard deviation.

Zinc

The prediction accuracy of Zn increases with increasing preparation steps. The $R^2$ values increase from a weak to a moderate prediction accuracy of 0.74 and an $RPD = 2.0$. Comparable studies report $R^2$ values for Zn retrieval between 0.58 and 0.64 (Wang et al., 2014), 0.77 (Gholizadeh et al., 2015a), and 0.86 (Tan et al., 2019) for preparation steps similar to stage 2. The sequential extraction indicates a strong connection to Fe-/Mn oxides, followed by carbonates and organic matter. This result coincides with the increasing accuracy when removing the interfering soil attributes (e.g., SOM) and multiple scattering effects due to different grain sizes by using the spectral pre-processing FD-SNV (Gholizadeh et al., 2015a).

The RFR model shows good prediction results for stage 4 with $R^2 = 0.67$. However, if the $RPD = 1.6$ is considered, the model can just distinguish between high and low values. Compared to Tan et al. (2019) with $R^2 = 0.90$, our study shows a much weaker model accuracy.

4.2 Method performance

The performance of the PLSR and RFR model is almost very similar for Cu and Ni estimations and varies little for Cr and Zn estimations. The poor performance of the PLSR, compared to the previously discussed studies, shows that the preparation steps need to be more enhanced to achieve similar prediction qualities. Possible reasons why these studies were obtaining a better model performance while using the same approach could be due to the inhomogeneity of the soil samples (e.g., sludge layer, mixed grain size, plastic contaminants) and the high
PTE concentration variability within the study area (Cipullo et al., 2019; Wang et al., 2020).

Different soil colours have an impact on the soil reflectance spectra and may additionally influence the performance of a good estimation model for PTEs (Gholizadeh et al., 2018b).

RFR shows a high variability in model performance for different preparation steps, but a generally weaker performance with increasing spectral pre-processing (FD-MSC or FD-SNV) and increasing preparation steps, compared to PLSR. Except for the performance for Cr estimation, the RFR model shows a $R^2 = 0.63$ with further pre-processing steps for stage 2, which is similar to the PLSR model with stage 4 and an $R^2 = 0.67$. Cheng et al. (2019) conclude that RFR are better in handling inhomogeneous soil samples, skewed distributions and a high variability of the PTE concentration than PLSR. This does not, however, hold true for the here investigated Technosol samples. The unknown overall impact of a single value in the RFR analysis, using between 50 and 366 individual regression trees, may influence the RFR performance (Wang et al., 2020). The study by Sun et al. (2019) reveals an improvement of the PLSR performance by correcting the skewed concentration data. However, the transformation of the skewed concentration data did not feature a Gaussian distribution, which means that no improved PLSR performance was achieved. In addition, LOI preparation was not yet considered in other studies. Our study therefore shows one additional possibility to increase the performance of PTE modelling by applying spectroscopy on samples after LOI.

5. Conclusion

The present study investigates Technosols from a sewage farm and different sample laboratory preparation and the relationship of pre-processing with different regression techniques to estimate soil PTE concentration using spectral reflectance from hyperspectral cameras in the laboratory. The results indicate that laboratory hyperspectral imagery provides good quality data for estimating PTE concentration even for highly heterogeneous urban soil samples. PTE content can be successfully predicted for Cu, Cr, and Zn with variable performances based on different pre-processing approaches. With increasing preparation
steps, the performance of PLSR can be improved to a moderate accuracy for Cr, Cu and Zn estimation, whereas RFR shows a weaker prediction accuracy with increasing preparation steps. The best models were obtained with the following pre-processing techniques: SNV, FD SNV and FD MSC. The poor performance of RFR, compared to recent studies, shows a low quantitative prediction within the heterogeneous soil samples, which is typical for Technosols. Increasing the sampling rate and grouping the soil samples by spectral responses, may increase the RFR performance. PLSR showed to be the more promising approach for estimating PTEs at more extensive preparation steps. More research is needed for the consolidation of PTE prediction based on soil spectroscopy for this technique to be operationally applicable to variable environments and with differing soil properties.

6. Acknowledgments

The authors acknowledge the financial support of the “Central Innovation Programme for small and medium-sized enterprises (SMEs)” of the Federal Ministry for Economic Affairs and Energy. The assistance of Constantin Hildebrand, Martina Heinrich, Dr. Florian Hirsch and Gbenga Fawehinmi for helping in the field and doing the chemical analysis, as well as Dr. Robert Milewski for checking the code and results during the model analysis. We thank the anonymous reviewers for their constructive comments that helped to improve this manuscript.

References


https://doi.org/10.1016/j.rse.2015.03.033

https://doi.org/10.17221/113/2015-SWR


https://doi.org/10.1080/10643389.2018.1447717


https://doi.org/10.1021/es015747j

Körting, F., 2019. Development of a 360 ° hyperspectral drill core scanner Test of technical conditions and validation of high-resolution near-field analysis of crystalline basement rocks using COSC-1 core samples 123. https://doi.org/http://doi.org/10.2312/GFZ.b103-19071


Rogass, C., Koerting, F.M., Mielke, C., Brell, M., Boesche, N.K., Bade, M., Hohmann, C.,


Appendix
S 1: Scedasticity Plots for PLSR
S 2: Scedasticity Plots for RFR
S 3: Scatter plots of the measured vs. predicted concentration of Cu, Cr, Zn and Ni for all four stages, using the PLSR model.
S 4: Scatter plots of the measured vs. predicted concentration of Cu, Cr, Zn and Ni for all four stages, using the RFR model.