Computers and Electronics in Agriculture 191 (2021) 106539

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Soil profile analysis using interactive visualizations, machine learning, and deep learning

Vung Pham [a,*], David C. Weindorf [b], Tommy Dang [c]

[a] Department of Computer Science, Sam Houston State University, Huntsville, 77340 TX, USA
[b] Department of Earth and Atmospheric Sciences, Central Michigan University, Mount Pleasant, 48859 MI, USA
[c] Department of Computer Science, Texas Tech University, Lubbock, 79409 TX, USA

ARTICLE INFO

Keywords:
Chemical measurement data analysis
Intelligent visual analytics
pXRF Data visualization
Soil property predictions
Vis–NIR spectra
Machine learning and deep learning

ABSTRACT

Soil is an essential element of life, and soil properties are crucial in analyzing soil health. Recent developments of proximal sensor technologies, such as portable X-ray fluorescence (pXRF) spectroscopy or visible and near-infrared (Vis–NIR) spectroscopy, offer rapid and non-destructive alternatives for quantifying data from soil profiles. While the data collection time using these technologies decreases significantly, the subsequent analysis remains time-consuming, and current analysis solutions only provide basic visualizations. Furthermore, the use of collected data from proximal sensors to predict high-level soil properties has garnered worldwide attention in the past decade, owing to its convenience. Therefore, this paper discusses the objectives for software solutions in this area, consolidated from interviewing 102 stakeholders. Following these requirements, data visualizers work closely with soil scientists to propose a set of interactive visualizations for analyzing soil profiles using pXRF data. These interactive visualizations receive positive feedback from the domain experts. This project also explores various machine learning and deep learning approaches to predict soil properties from spectral data. This work then proposes a deep learning model called RDNet that achieves state-of-the-art results in predicting pH_{H_2O} and pH_{KCl} from Vis–NIR spectra acquired from a set of globally distributed soil samples.

1. Introduction

Soil properties are crucial in analyzing soil health (e.g., soil fertility) and environmental analyses. Measuring these complex soil properties requires complicated and time-consuming laboratory procedures. Using proximal sensors such as portable X-ray fluorescence (pXRF) or visible and near-infrared (Vis–NIR) spectroscopy to analyze soil is gaining favor because they offer a rapid means of quantifying elemental concentrations or other alternatives for data collection from the soil profiles. While data collection time is minimal, utilization of the collected data (i.e., analysis or high-level property predictions) remains time-consuming and error-prone.

There are two main trends in utilizing pXRF or Vis–NIR data in this domain. The first trend is to depict if the elements of interest (e.g., plant-essential elements, heavy metals) exist in the studying units and their spatial distributions over the units (e.g., Pham et al., 2019; Pham et al., 2020b). The second trend is using pXRF/Vis–NIR approaches as fast, cost-effective, and environmental-friendly alternatives to the conventional laboratory approach to quantify soil chemical and physical properties (e.g., Silva et al., 2020; Andrade et al., 2020; Adler et al., 2020).

Regarding the first trend, many of the existing works use tables or basic visualizations to report their findings (e.g., bar charts, line-graphs, heat-maps, and scatterplots). Additionally, most of the reviewed works used either conventional tools (e.g., Microsoft Excel), software applications (e.g., ArcGIS), or their combinations to analyze soil data and make graphics. Especially, several analysts also use *R* (e.g., Silva et al., 2020), Python (e.g., Adler et al., 2020), or JavaScript (e.g., Pham et al., 2019; Pham et al., 2020b) for these analysis tasks. Thus, it is beneficial to have a software solution that can provide appropriate interactive visualizations and allows non-professionals to create 2D and 3D charts of chemical measurements without special knowledge of programming languages.

Additionally, there are several machine learning (ML) methods and deep learning (DL) methods used in the literature to predict soil properties from pXRF or Vis–NIR data. The popular ML methods in this area include partial least squared regression (PLSR), random forest (RF), support vector regression (SVR), and multiple linear regression (MLR).

* Corresponding author.
E-mail address: vung.pham@shsu.edu (V. Pham).

https://doi.org/10.1016/j.compag.2021.106539
Received 8 April 2021; Received in revised form 22 October 2021; Accepted 31 October 2021
Available online 11 November 2021

Conversely, deep learning (DL) methods are gaining traction thanks to their state-of-the-art results in various domains. Soil property prediction from spectral data is no exception. The common DL methods in this area include the multiple perceptron neural networks (MLP) and Convolutional Neural Network (CNN).

Most of the current studies in this area have relatively small samples (ranging from 40 to 400). Concomitantly, there are large numbers of data features (e.g., Vis–NIR wavebands) available. Therefore, the conventional ML methods (e.g., PLSR and RF) are gaining more success in this area. Conversely, the DL techniques (e.g., MLP and CNN) applied to this area have several limitations. These limitations include DL approaches being too simple (few training weights) or too complicated (many training weights) compared to the available data samples and data features.

This project first elicited software requirements from proximal sensor end-users and other stakeholders. After consolidating the requirements, data visualizers and soil scientists worked together to develop a set of common interactive data visualizations to analyze soil profiles using elemental concentrations quantified by pXRF devices. Also, this work utilizes Vis–NIR spectra acquired from globally distributed soil samples for training ML/DL models to offer alternatives for predicting high-level soil properties. We hypothesize that requirement elicitation results help make the proposing solutions more practical. Furthermore, the interactive visualizations help reduce the pXRF data analysis time while the ML/DL solutions offer fast, cost-effective alternatives for measuring soil properties.

2. Materials and methods

2.1. Requirement elicitation

Interviewing 102 stakeholders during a National Science Foundation (NSF) I-Corps[1] program unveiled valuable insights about the requirements for software solutions in this area. This technology's stakeholders include the end-users of proximal sensors and related personnel. Specifically, out of 102 stakeholders interviewed, 44 of them are the proximal sensor end-users. These proximal sensor end-users are mainly soil scientists (21). The 23 others include anthropologists, geologists, and hydrologists. Also, many of them (21) are from United States government agencies, 17 are from universities, and six are from private laboratories/companies.

2.2. Interactive visualizations of pXRF data

Soil scientists perform the data collection step for the pXRF data used in this project. They go to the field of interest, excavate a pedon, and use strings to lay a physical wire-frame over the profile. They then scan each cell of the wire-frame with a pXRF device. In this particular case, the wire-frame has 130 cells, and each cell is $10 \times 10 cm^2$ in size (the number of cells and the type of evaluations may vary from study to study).

Fig. 1 gives an overview of the approach proposed by this work for analyzing soil profiles using interactive visualizations and pXRF data. First, soil scientists collect the data from the field. The data is then pre-processed and interpolated before generating 2D and 3D interactive visualizations for analysis purposes. Seamless integration among these steps offers an alternative for rapid characterization of soil profiles. Section 3 details these stages.

2.3. Predictions of soil properties from Vis–NIR data

Deep learning approaches are often data-intensive; they need a large amount of data to perform well. Therefore, this work uses the ICRAF-ISRIC world soil spectral library (Garrity and Bindraban, 2004) as the data source for ML/DL experiments. This library contains 4,437 samples from 785 soil profiles over 58 countries from Africa, the Americas, Asia, and Europe. The Vis–NIR spectra have wavelengths ranging from 350 nm to 2,500 nm in 216 wavebands.

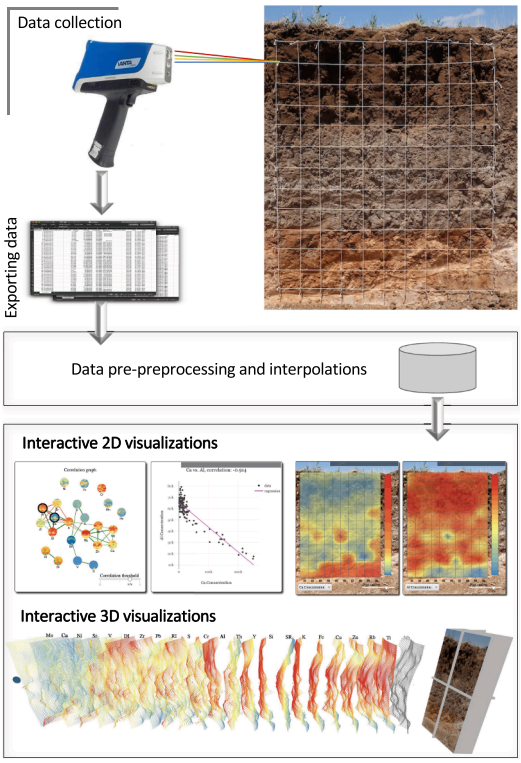

Fig. 1. Overview of the stages of the interactive 2D and 3D soil profile visualizations for elemental concentrations quantified using portable X-ray fluorescence spectroscopy.

This spectral library also has 23 soil properties that were measured using laboratory procedures. However, this work focuses on predicting pH_{H_2O} and pH_{KCl} from the available Vis–NIR spectra. The choice of these two soil properties in this project is inspired by a recent paper from Wang et al. (2019). In their work, they proposed to use laboratory procedures to measure pH_{H_2O}, then use machine learning methods to predict pH_{KCl} from the measured pH_{H_2O} to save time and cost for a massive measurement in the laboratory. This work goes one step further and proposes to predict both pH_{H_2O} and pH_{KCl} from Vis–NIR data, instead.

There are 3,838 samples after removing the missing data and duplicated records. These samples are divided into training (3,070), validation (384), and testing (384) sets. The validation set is used to find the best model trained on the training set. Also, the reported results are those evaluated on the independent test set. This approach is an attempt

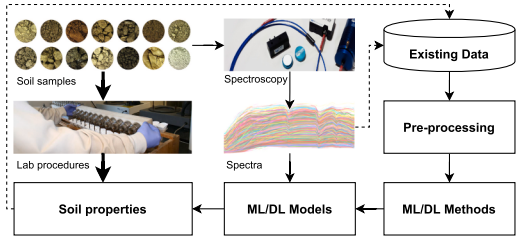

Fig. 2. Overview of the method that uses spectral data and machine learning (ML) or deep learning (DL) to predict soil properties (thin arrows) as an alternative to the conventional laboratory procedures (thick arrows) utilizing the existing data (dashed arrows).

[1] https://www.nsf.gov/news/special_reports/i-corps/.

to assure that the reported results generalize.

Fig. 2 depicts the main ideas of using relatively easy-to-collect spectral data for predicting soil properties. Conventionally, soil scientists need to go through all the time-consuming, complicated, and destructive (due to toxic chemicals) laboratory procedures to measure soil properties. Instead, this work proposes to use Vis–NIR spectra generated from soil samples to predict soil properties. Specifically, this solution utilizes existing data (Vis–NIR spectra and soil properties) to train ML and DL models to predict soil properties for new soil samples using Vis–NIR reflectances. Notably, scientists can always add to the existing data to train and improve future models.

3. Results

3.1. The requirement elicitation

Five objectives were discovered after the NSF I-Corps program. **Objective 1**, research and development of a set of typical visualizations for chemical measurement data, is essential because it helps to reduce data analysis time. Currently, there is no specific guidance in the analytics process of the proximal sensor data. Therefore, soil scientists must build new visualizations to suit particular analysis tasks. They also use conventional tools (e.g., Microsoft Excel) and software applications (e.g., ArcGIS) extensively for the analysis requirements. Still, these general-purpose applications are ill-suited for visualizing chemical measurement data.

Table 1 lists software types commonly used by the 44 interviewed proximal sensor end-users. There are two broad categories of use for these software applications: statistical analysis and chart generation. All of the interviewed end-users reported using Microsoft Excel for both of these tasks. Many of them (15 out of 44) often use custom codes (e.g., R, Python, or MatLab) for these two tasks. Also, many others use ArcGIS/ArcMap primarily for chart generations. Concomitantly, they also use SAS, SPSS, or JMP mainly for statistical analysis. There are also users (primarily from companies or government offices) who use custom-made software for their tasks.

Notably, besides using software to extract/copy/transfer data from proximal sensor devices to other platforms (e.g., computers or cloud data storage), only a few of the interviewed end-users use software from the hardware vendors for data analysis purposes. Interviews with original equipment manufacturers revealed various technical hurdles that encumber them from providing data visualization solutions for their equipment. Two main reasons are the lack of data visualization expertise and the different needs of individual customers. Thus, it is hard to develop data visualizations that suit the needs of all of their customers.

Therefore, **objective 2**, research and development of an intelligent visual recommendation component that provides personalized visualizations for chemical analysts, is a key differentiator. This component can learn and recommend appropriate visualizations to individual users based on their contexts. Personal recommendations have gained success in several fields, such as movie recommendations (e.g., Netflix), commercial product recommendations (e.g., Amazon), and advertisement recommendations (e.g., Google Ads). However, personalized visualization recommendations are still at the pure research stage and are not ready to produce deployable/marketable products. Most of the current visual recommendations (e.g., Microsoft Excel Recommended Charts) are based on fixed rules/heuristics and not continuous learnable components. The rules include using data types (e.g., categorical, numerical, ordinal) or graph-based visual features (e.g., Scagnostics Pham et al., 2020a).

Table 1
Statistics about common software used by proximal sensor end-users.

Excel	Custom code	ArcGIS/ ArcMap	SAS/ SPSS/ JMP	Custom software	Vendor's software
44	15	19	13	10	11

Quantitatively, interviews with proximal sensor end-users unveiled that they spend approximately 25% of their time building appropriate data analysis and visualization tools. Moreover, data visualizations are often stressful for users without visualization expertise. Therefore, deliverables of objectives 1 and 2 help reduce the data visualization time and user stress regarding analyzing and reporting their data.

Proximal sensor users also confer that approximately 5% of the sampling times they have errors in their collected data due to the lack of error indication while they are still on site. These errors are revealed during the analysis time at the office. Once there is an error in the scanning process, scientists must go back to the field and resample the soil profile. In many cases, going back to the area is time-consuming or even impossible. For instance, the National Aeronautics and Space Administration (NASA) is currently using related X-ray fluorescence and Vis–NIR spectroscopy sensors on its Mars rovers. NASA scientists with missions on Mars would not make another trip for their equipment to get back to Mars to resample the incorrectly sampled data. Consequently, **objective 3**, research and development of real-life error indications for proximal sensors, reduces the need for resampling or duplicated scanning efforts.

Different end-users scan for elemental concentrations from different materials. For instance, soil scientists work with soil, while anthropologists might be interested in a specific rock. Users need to purchase different calibration packages or (for a few expert/capable users) develop their calibration packages. Most current proximal sensor manufacturers use conventional, low-level spectrum analysis algorithms to calibrate the concentration data for different material types. By comparison, ML or DL are producing state-of-the-art results in various fields. Thus, our initial communications with the Chief Executive Officers and product development teams from proximal sensor manufacturers uncovered strong interest in using ML or DL to predict elemental concentrations.

Therefore, for **objective 4**, research and development of ML/DL components for device calibrations (e.g., spectrum to concentrations), this work proposes to use existing spectral and elemental concentration data, processed using conventional spectrum processing algorithms, to train models to predict the concentrations from the spectral data in the future. There are two purposes for this: accuracy improvement and reduced processing time.

In the same vein, **objective 5**, research and development of ML/DL components with an emphasis on soil property predictions, is equally important. Discussions with proximal sensor users suggest having ML/DL models to predict higher-level pedological characteristics from the data collected using proximal sensors. These models can replace the need to go through slow, costly, and destructive laboratory procedures to quantify these properties.

The scope of this project with these five objectives is enormous. Thus, this paper focuses on objectives 1 and 5 with some quick discussions about other objectives. Therefore, the following sections discuss implementation results related to a set of common interactive visualizations for soil profile analysis (objective 1) and ML/DL methods for predicting soil properties from spectral data (objective 5).

3.2. Soil profile visualizations using pXRF data

Processing of pXRF data: pXRF devices are good at detecting medium or heavy elements, but they have difficulties quantifying light elements or elements with low concentrations in the analyzing unit. In other words, pXRF devices have instrumental LODs (limit of detections). Therefore, one of the data-cleaning tasks is to remove the elements with LODs. Moreover, the software also statistically calculates outlying concentration values and alerts users if there are such abnormal concentrations.

Furthermore, to support pedological feature analysis, this solution

adds weathering indices and elemental ratios designed for these purposes (Stockmann et al., 2016) from the detected elemental data. Specifically, the software adds *Ruxton Weathering Index* (RI) defined as SiO_2/Al_2O_3, *Desilication Index* (DI) defined as $SiO_2/(Al_2O_3 + Fe_2O_3 + TiO_2)$, and *Stable Index* (SI) calculated as Ti/Zr. For simplicity, in this writing, "chemical element" also refers to these computed values. Similarly, any other computations of interest can be added at this stage.

Finally, concentration values are available in discrete points (130 in this case) in a soil pedon for each detected element. In other words, pXRF data are discretely distributed while its actual distribution is roughly continuous. Therefore, the data processing step interpolates the collected pXRF data using the Kriging algorithm (Van Beers, 2005). Specifically, this step uses a spherical Kriging model with $\sigma^2 = 0$ and $\alpha = 100$.

Data analysis tasks for pXRF data: Visualizations bring several benefits to exploring and analyzing complex environmental datasets and reporting the findings. Especially, collaborations between environmental experts and data visualizers bring great results. Specifically, the ecological experts bring up requirements, explain domain-specific tasks, and validate the findings, whereas the data visualizers can suggest appropriate interactive data visualizations for the domain requirements. Therefore, in this project, data visualizers collaborated with soil scientists to create interactive data visualization solutions that support the rapid study of pXRF data. Specifically, working with the soil scientists reveals the following analysis tasks (Pham et al., 2019) required while exploring pXRF soil horizon scanning data:

- **T1**: Provide an overview of all detected chemical elements and their derived values in the soil profile.
- **T2**: Show and quantify the correlation between any two selected elements.
- **T3**: Show and compare the spatial distributions of any two selected chemical elements over the pedon's surface.
- **T4**: Quantify the distributions of any two selected chemical elements over the pedon horizons.
- **T5**: Alert the potential existence of outlying data caused by in-field scanning errors.

As a solution for these discovered analysis tasks, this work adopts coordinated interactive views to visualize different chemical elemental distributions perspectives. These views include correlation graphs, scatter plots with linear regression lines, contour-maps/heat-maps, and box-plots.

Correlation graphs: It is crucial for the scientists to have an overview of all detected chemical elements and their relationships (Pham et al., 2019) (tasks **T1** and **T2**). Thus, this software calculates the correlations among the elements and generates a force-directed network graph, as shown in Fig. 3 (A). A vertex represents a chemical element and is overlaid by its contamination contour to guide users during the exploration process. A link indicates the correlation between two nodes: the thicker the link, the higher the correlation. The colors of the links (green vs. red) encode positive/negative correlations. Users can use the slider provided at the bottom right corner to refine the relationship network and focus on the strongly correlated chemical elements.

A scatterplot is generated on demand to verify the correlation of any two chemical elements (task **T2**). In Fig. 3 (B), each data point is a pXRF scan on the 2D soil profile grid. There is also a linear regression line to serve as a reference for the estimated correlation. Besides this qualitative reference, on top of the scatterplot, there is the *Pearson* correlation score to quantify the relationship of the two selected chemical elements.

Contour-maps/heat-maps: Soil scientists use physical strings to divide the pedon's surface into cells and scan each of them while collecting data (Pham et al., 2019). As a natural solution to this type of discrete data, this software creates a heat-map to depict the generated data from these individual cells, as shown in Fig. 4 (B). Though the scanned cells are discrete, the elemental distributions are continuous on the pedon's surface. Therefore, this software uses the Kriging algorithm to interpolate the elemental distributions over the surface and generates a contour-map to mimic the actual spatial distributions of the elemental concentrations over the profile (task **T3**). (see Fig. 5).

Box-plots: Soil scientists use box-plots to visualize the statistical distributions of the elemental concentrations across the soil horizons (Pham et al., 2019) (task **T4**). Specifically, the box-plot displays statistical data distribution over each soil horizon. It also shows the outliers (task **T5**) computed using the box-plot rule (Pham et al., 2020a). Moreover, many works in the soil research field recommend removing outlying data before applying techniques to improve the soil profile analysis accuracy. Thus, indicators of these outlying points are important while analyzing soil profiles.

Step-wise area-charts: Soil scientists often want to slide through vertical and horizontal slices of the analyzing unit and view how corresponding elemental distributions change (Pham et al., 2020b). Therefore, this software offers another visualization to depict elemental distributions at the horizontal or vertical slices of interest (task **T3**). Fig. 6 is a snapshot of the step-wise area-chart to depict the distributions of the detected plant essential elements and the computed pedological features (DI, RI, and SI) in the underlying soil profile. Furthermore, this software normalizes the elemental concentrations into the range of [0 to 1] in this visualization. Finally, the color scale goes from blue to red for values from 0 to 1 normalized value range. Notably, there are gray handlebars that allow scientists to slide through the soil profile and view the distribution changes accordingly.

In this case, soil scientists usually use the line-graphs to represent the elemental distributions. However, when there is a large number of elements, line-graphs suffer from visual cluttering issues. The typical approach to overcome this visual cluttering issue is to use one chart for every element. However, the space limitation of the small screen sizes

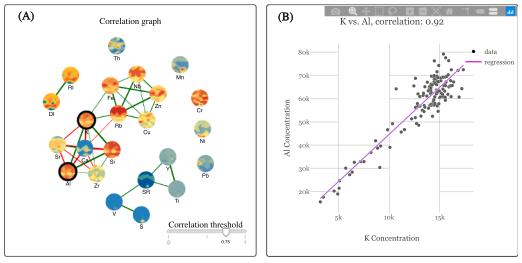

Fig. 3. Force-directed correlation graph (A) and Scatter plot (B) visualize and quantify the correlation among the detected elements. The correlation graph (A): contours over the detected elements give overviews of how the elements are distributed over the profile (red/blue for high/low concentration), link thicknesses depict correlation, and red/green link colors designate negative/positive correlations.

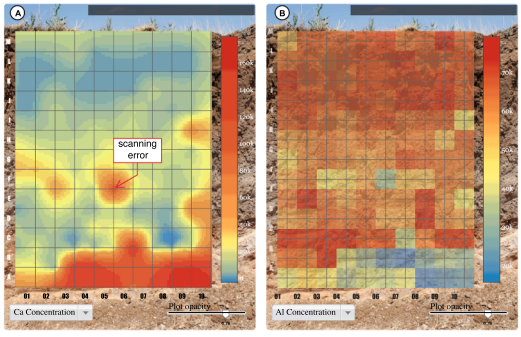

Fig. 4. Visualizations of the spatial distributions of the elemental concentrations. Contour-map (A) uses interpolation to mimic natural elemental distributions, while heat-map (B) depicts the actual elemental values from discrete, scanned points.

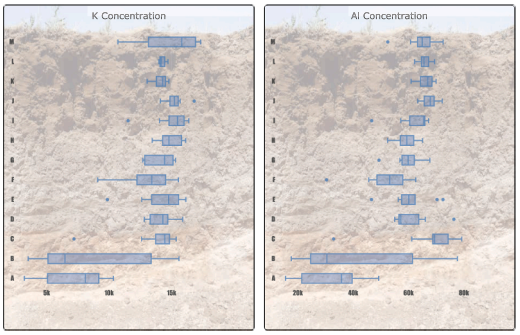

Fig. 5. Box-plots to summarize statistical distributions of elements per horizon for 13 horizons (from 'A' to 'M') in the profile.

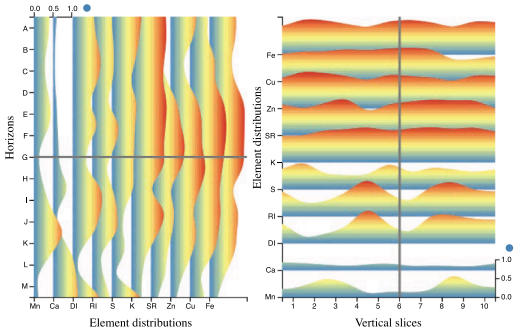

Fig. 6. Elemental distributions for selected elements over horizontal and vertical slices. Blue to red colors depicts low to high relative concentrations, respectively. Gray handlebars allow users to slide through the profile.

makes this approach impractical. Therefore, elemental step-wise charts are more appropriate in this case. In this approach, elements are shifted for a specified space rather than separated into different charts. However, using line-graphs with the shifting strategy makes it difficult to observe an individual element's base (Pham et al., 2020b). The solution to this issue is to use area-charts in place of line-graphs because area-charts have solid-filled areas that help indicate where the elements' bases are (Pham et al., 2020b). The filled area-charts still have a disadvantage as they may hide one another. Therefore, this visualization provides interactive options to order the area-charts for individual elements so that the area-charts in the front have lower heights than those in the back.

3D visualizations for pXRF data: The medical scanning techniques (e.g., CT/Ultrasound scanning) inspire this 3D approach to soil profiling. Precisely, doctors can continuously navigate different slices of the brain or body. Similarly, soil scientists often want to scan through the soil profile and observe the changes in the elemental distributions. Additionally, 3D visualizations of these elements help differentiate the elemental spatial distributions quantitatively and perceptively rather than merely perceptive cognition as using only colors to represent values in the 2D heat-map approach (Pham et al., 2020b).

Fig. 7 is a snapshot of an interactive 3D visualization for this purpose.

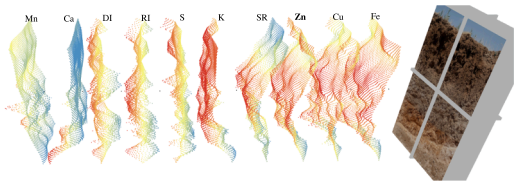

Fig. 7. Overview of the elemental distributions for selected elements in 3D visualization. Blue to red colors depicts low to high relative concentrations, respectively. Gray handlebars allow users to slide through the profile.

It provides an overview of the elemental distributions over the profile's spatial space (task **T1**) for the plant essential elements and the computed pedological features (or other selected elements). In other words, it is the 3D version of the 2D slice views (Fig. 6). The elemental distribution values have been normalized (to the [0, 1] range) and color-coded (from blue to red) by the cell values. Ranking and ordering the chemical elements enable grouping chemicals with similar concentration distributions together. Additionally, there is a picture of the soil profile that helps correlate the elemental distributions to the underlying soil profile. Users can drag the gray vertical or horizontal handlebars in the profile object to scan through the vertical or horizontal slices correspondingly.

3D scatter views: Soil scientists often like to investigate and correlate spatial distributions of individual elements or pairs of elements (Pham et al., 2020b) (task **T3**). In that case, users can select the element (s) from the previous 3D view (Fig. 7)) to display 3D scatter view(s). Fig. 8 shows the detailed 3D scatter view of the computed SI values for the underlying soil profile. The x-axis and y-axis represent the horizons and vertical slices, respectively, while the z-axis designates the soil property values. While the coded colors help indicate the elemental distributions qualitatively, the z-values quantitatively represent these distributions.

3.3. Soil property predictions using Vis–NIR data

Vis–NIR spectrum pre-processing: Generating errors in data while recording the spectra by the hardware devices (e.g., Vis–NIR spectroscopy) is inevitable. Thus, scientists often apply data transformation to soil spectra to generate more meaningful data. Different spectral pre-processing methods resulted in different pre-processing effects. However, many works in this area (e.g., Ba et al., 2020; Liu et al., 2020) use Savitzky-Golay (SG) for pre-processing of the signals. The SG smoothing method deletes the effects of high-frequency random noise produced by ground interposition, improves the signal-to-noise ratio, and reforms the reflectance spectra (Savitzky and Golay, 1964). Therefore, this work also experimented with SG for transforming the data before applying ML/DL techniques.

Specifically, this work applies an SG transformation that uses the first derivative, window length of 11, and polynomial order of 5. Fig. 9 shows the resultant SG transformation on the original reflectance data. Notably, SG results capture the changes in reflectance values between every consecutive waveband. These characteristics of SG transformation provide better information to train ML/DL models. Precisely, except for PLSR, all other experimented methods (RF and DL methods), the models created using SG-transformed data outperform those trained on the raw reflectance data.

Performance evaluation metrics: This work evaluates different models' performances using mean squared error (MSE), R^2 correlation, and residual prediction deviation (RPD). Using MSE as a loss function for regression problems is a prominent practice. However, in this area, soil scientists often report R^2 and RPD scores too. R^2 reflects the correlation between the actual and the predicted values. Moreover, RPD gives a

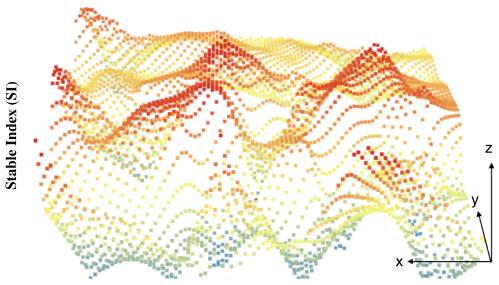

Fig. 8. 3D scatter view of the Stable Index (SI) for analyzing pedological properties. Blue to red colors depicts low to high relative concentrations, respectively.

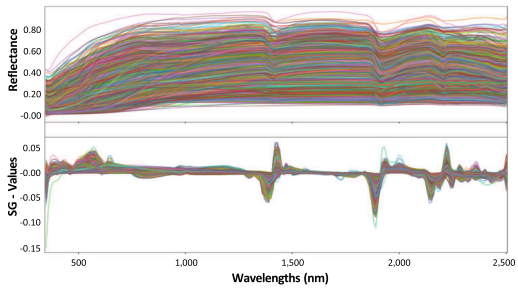

Fig. 9. Visible and near-infrared (Vis–NIR) reflectances and corresponding Savitzky-Golay (SG) values of the dataset used in this project.

relative indication to compare the predictive powers of models validated on different data samples. Table 2 shows the RPD values and their corresponding predictive abilities (Chang et al., 2001).

Common ML methods for soil property prediction: The common ML methods in this area are PLSR and RF due to the small number of samples and many data features. Therefore, this work also experimented with these two ML methods. For PLSR, the number of components is one important hyperparameter to tune. Thus, this work searched for the best number of PLSR components on validation data. Consequently, a PLSR model is trained with the best number of components (69). Specifically, this PLSR model has MSE, R^2, and RPD of 0.72, 0.63, and 1.68 for pH_{H_2O} and 0.72, 0.57, and 1.68 for pH_{KCl}, respectively on the test set. These results imply that though PLSR is appropriate for cases with a small number of samples and the output data tends to be normal, it does not perform well on a larger number of heterogeneous samples.

For the RF method, there are many hyperparameters to tune. However, the number of trees (*n_estimators*) and the maximum depth of each tree (*max_depth*) are the two most important ones. Therefore, this work implemented a grid search function to look for the best RF set of hyperparameters using the validation data. Specifically, the searched hyperparameters include $max\_depth = [10, 20, 40, 80, 150, 200, 400, 500, 600, 700, 1000, 2000, 3000]$ and $n\_estimators = [50, 100, 200, 300, 1000]$. These hyperparameters' search patterns follow a common practice that starts with small values and small skips, subsequently increasing these values exponentially. The best configuration has $max\_depth = 20$ and $n\_estimators = 1000$. This RF model has MSE, R^2, and RPD of 0.46, 0.77, and 2.11 for pH_{H_2O} and 0.46, 0.72, and 2.1 for pH_{KCl} respectively. Better RF results compared to PLSR's suggest the potential non-linear relationship between the input Vis–NIR spectra and the pH_{H_2O}/pH_{KCl} outputs.

Common DL methods for soil property prediction: DL approaches tend to have excellent capability to work with non-linearity in the input/output relationships. Therefore, there is a potential to explore DL approaches for this scenario. In the literature, MLP and CNN are the two commonly used DL methods for extracting soil properties from Vis–NIR spectra. Consequently, this work experimented with these methods in this scenario.

The best experimented MLP neural network's architecture and hyperparameters include five fully connected layers with 512, 256, 128, 64, and 32 hidden units, respectively. There is also one dropout layer (dropout rate of 0.1) after every fully connected layer to account for overfitting. This MLP model provides MSE, R^2, and RPD of 0.36, 0.83, and 2.46 for pH_{H_2O} and 0.34, 0.81, and 2.53 for pH_{KCl} respectively. Notably, these results outperform the experimented PLSR and RF models. These initial improvements in results indicate a great potential to apply DL methods in this area.

In a recent work (Xu et al., 2019), Xu et al. applied DenseNet (a specific type of CNN) to predict predicting soil organic matter (SOM) from Vis–NIR spectra. Therefore, the same DenseNet architecture is used in this scenario with the assumption that it can extract useful salient features from Vis–NIR spectral data. Interested readers can refer to the original work for the description details of this architecture. This model provides MSE, R^2, and RPD of 0.51, 0.75, and 2.06 for pH_{H_2O} and 0.47, 0.74, and 2.15 for pH_{KCl} respectively.

Though DenseNet results are discouraging in this scenario, it is common to think of the spectral data's consecutive wavebands as a data sequence. Also, CNN has the power to extract information from 1D sequential data. Therefore, this work also explored another CNN architecture using Visual Geometry Group (VGG)[2] blocks. This architecture has four VGG blocks (each contains two CNN layers and one pooling layer), with 32, 64, 64, and 64 hidden units. As an intermediate 'buffer,' a fully connected layer with 128 hidden units is added before the final output layer. Additionally, there is a dropout layer (rate of 0.1) after every VGG/fully connected layer to reduce overfitting issues. This model provides MSE, R^2, and RPD of 0.35, 0.83, and 2.5 for pH_{H_2O} and 0.36, 0.8, and 2.46 for pH_{KCl} respectively. Though these results are better than DenseNet, they are only comparable to those produced by the MLP model.

Notably, considering the spectral data as consecutive wavebands as a data sequence, this work also applied recurrent neural network (RNN) and long-short term memory neural network (LSTM) to this problem. However, they do not work due to the very long sequence length (216) in this case. For the sake of space and clarity, this paper does not report the results of using these types of neural networks in tackling the current problem.

WaveNet and RDNet for soil property prediction: CNN architecture has gained success in computer vision tasks. One of the main reasons for its success is the large number of training images available (e.g., ImageNet Deng et al., 2009 contains 1,281,167 images for training). Though this case has a relatively larger amount of soil samples than other work in the same area, this number is still small for training an efficient CNN model. Therefore, this work experimented with WaveNet (Oord et al., 2016) for this scenario. Fig. 10 depicts the difference between CNN architecture and its WaveNet counterpart. Specifically, with dilation, WaveNet needs fewer model weights while still having the

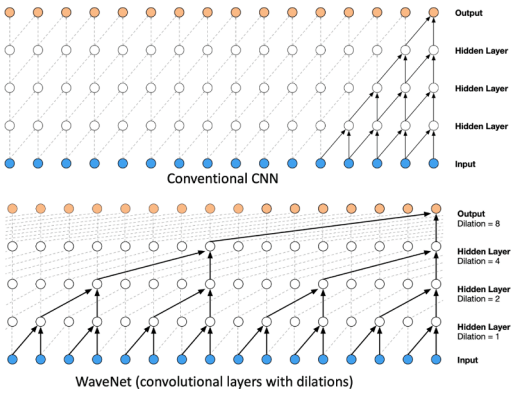

Fig. 10. Convolutional Neural Network (CNN) vs. WaveNet (adopted from Oord et al., 2016). WaveNet can gather information from the complete sequence with fewer parameters (solid arrows) to be trained, thanks to dilation rates.

Table 2
Residual prediction deviation (RPD) value ranges and corresponding predictive abilities.

Ability	bad	rough	moderate	good	excellent
RPD	<1.5	1.5–2.0	2.0–2.5	2.5–3.0	>3.0

[2] http://www.robots.ox.ac.uk/vgg/.

power to gather information from the whole data sequence. Having fewer weights means requiring a smaller number of training samples to fit the model.

Specifically, after experimenting with different configurations, this work defined a WaveNet block with six convolutional layers (32 filters each) and dilation rates of 1, 2, 4, 8, 16, and 32 correspondingly. As usual, there is one dropout layer (rate of 0.2) before the output layer to tackle the overfitting issue. This WaveNet architecture gives MSE, R^2, and RPD of 0.32, 0.85, and 2.6 for pH_{H_2O} and 0.29, 0.84, and 2.75 for pH_{KCl} respectively.

Though the WaveNet results outperform all other experimented models described so far, it has a limitation that it only contains one WaveNet block. The model's performance degrades when more such blocks are stacked to the network. Thus, this work adopted the skip connection idea from ResNet (He et al., 2016) to overcome this performance degradation issue. As shown in Fig. 11, this work defines a residual dilated block (RD block) as the WaveNet block with a skip connection. This skip connection adds the input to the output before the last activation layer. This modification allows stacking up to 10 RD blocks to create a neural network, called Residual Dilated Neural Network (RDNet), for this scenario. Fig. 12 reports RDNet results evaluated on the test set. Specifically, MSE, R^2, and RPD are 0.28, 0.86, and 2.76 for pH_{H_2O} and 0.25, 0.86, and 2.93 for pH_{KCl} respectively. These results indicate that RDNet outperforms other experimented ML/DL models for this case.

ML/DL model explainability: ML/DL approaches have their predictive power thanks to their ability to learn salient features from the input data. This statement is especially true for DL approaches. The proposed RDNet model has 10 RD blocks. This deep network has the potential to memorize trivial characteristics of the data instead of learning from useful features. Therefore, this work explores techniques to explain what the experimented ML/DL models pay attention to when making their predictions. In other words, it is beneficial to know what Vis–NIR wavebands have high contributions to the final predictions of pH_{H_2O} and pH_{KCl} values.

For PLSR, it is common to use its coefficients to denote the importance of the input features. By comparison, RF has its built-in function to provide feature importance values. This built-in function uses SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) to calculate feature importance. SHAP is a unified approach to explain the output of any ML model by game theory and local explanations. Similarly, a SHAP Kernel Explainer (Lundberg and Lee, 2017) is used to generate feature importance for the RDNet. This paper only presents RDNet's feature importance for this section's brevity because it is the deepest among the experimented DL models. In other words, it has the highest potential to learn salient features that might not be explainable.

As shown in Fig. 13, PLSR learns from a wide range of bands, while RF only focuses on the last two high-entropy wavebands (especially the last one). Contrariwise, RDNet leverages the information from all these high-entropy wavebands to make its predictions. The RDNet's feature importance values in appropriate useful wavebands indicate that RDNet learns valuable information from the underlying data rather than memorizing trivial features from the training inputs. One evidence to support this statement is that the reported results, evaluated on the test set, are generalized and comparable to that on the validation set, in this case.

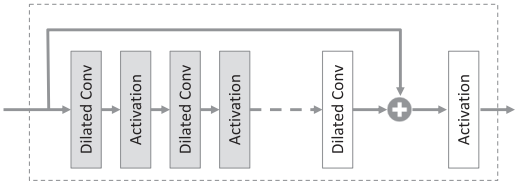

Fig. 11. Architecture of a residual dilated block. It utilizes dilation to reduce the number of training parameters. The skip connection allows stacking more of this block together to train a deeper network.

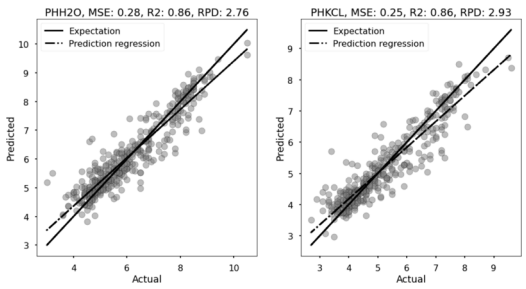

Fig. 12. Residual Dilated Neural Network (RDNet) results on the test set. The solid line is the expectation (when predicted values equal actual values), and the dashed line depicts the prediction regression.

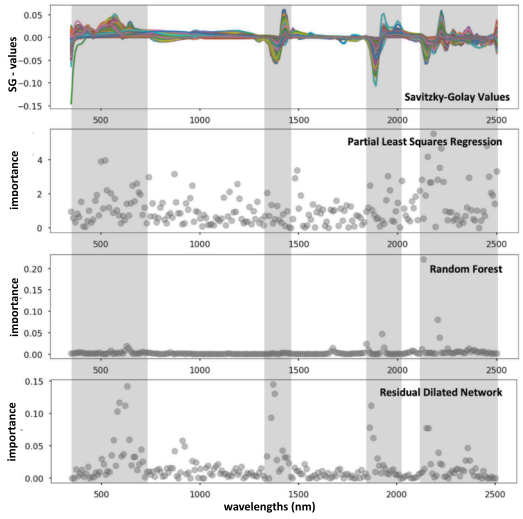

Fig. 13. Savitzky-Golay (SG) values in wavebands with more entropy (shaded) and corresponding feature importances for the experimented models.

4. Discussion

4.1. Interactive visualizations for pXRF data

The soil scientists provided positive feedback for the proposed interactive data visualizations. They also suggest that tangential adaptations of the aforementioned interactive visualization approach can be made to analyze the data from cores extracted from a larger geographic area. Cores are easier to extract, elemental quantification using pXRF devices is rapid, and this proposed solution can generate visualizations in seconds. These together offer alternatives to analyze many soil profiles quickly.

The soil scientists reported good use-cases regarding analyzing elemental data. For instance, the visualization helps highlight the extremely high value of *Ca* concentration in the cell *F6* visually in the contour map as shown in Fig. 4 (A). This outlying point was not observable if the scientists were simply looking into every individual data item in the tabular representation. The soil scientists explained that the pXRF equipment was hitting directly toward some unwanted material in this cell during the scanning time.

Though these interactive visualizations are essential for analyzing soil profiles using pXRF data, they do not mean to be an extensive list of common visualizations for chemical analysis (**objective 1**). In the future, there should be more work with proximal sensor end-users in various fields to complete this objective. Furthermore, though these

visualizations have some indications regarding data collection errors (**objective 3**), they are only some initial indicators, and there should be more future research regarding this objective.

4.2. Soil property predictions from Vis–NIR data

Table 3 summarizes the results for the experimented ML/DL methods evaluated on the test set. RDNet outperforms all other experimented methods thanks to its ability to extract useful features from the wavebands with high entropy (as shown in Fig. 13). Notably, visualization of the results for the RDNet model (Fig. 12) shows the tendency of predicting higher values for low pH_{H_2O}/pH_{KCl} outputs and lower values for the high ones. This fact can be utilized to tune this RDNet model and gain better prediction accuracy. Moreover, besides pH_{H_2O} and pH_{KCl}, more ML/DL models should be explored to predict other soil properties in the future (**objective 5**).

Furthermore, proximal sensor device calibration (**objective 4**) continues to be an open problem for both device manufacturers and end-users. There should be more collaborations with device manufacturers to have X-ray fluorescence data, and corresponding quantified elemental concentrations in the future. Using these data, similar strategies described in this paper can be used to create ML/DL models for device calibration purposes. Finally, **objective 2** discovered in Section 3.1 is crucial to bring this innovation to the broad range of proximal end-users. Therefore, there should also be work on reinforcement learning methods to have an artificial agent that can recommend appropriate visualizations to the users based on their contexts.

4.3. Implementations and reproducibility

This work implements the proposed visualizations as a web application for portability. The data, source codes, and web prototypes of the proposed interactive 2D and 3D visualizations are available on Github pages[3]. Additionally, the experimented ML/DL models are developed in Python using TensorFlow. Notably, to assure reproducibility, the random state for the execution environment (Numpy and Tensorflow) is fixed to a default value (42 in this case). Data, source codes, and result visualizations of the experimented ML/DL models are also available on Github[4].

5. Conclusions

This work presents the findings of an NSF I-Corps program in which 102 proximal sensor end-users and other stakeholders in the field of soil profile analysis were interviewed. These interviews unveiled valuable insights regarding a set of objectives for software solutions that support this specific domain. These objectives include 1) having a set of typical visualizations for chemical measurement data, 2) having an intelligent visual recommendation component that provides personalized visualizations for chemical analysts, 3) having real-life error indications for proximal sensors, 4) having machine learning components for device calibrations, and 5) having machine learning components with an emphasis on soil property predictions.

This paper focuses on objectives 1 and 5 and provides quick discussions regarding the other three objectives. Specifically, in this work, data visualizers explored the current visualizations in this area and collaborated closely with soil scientists to devise interactive visualization solutions to analyze soil profiles using pXRF data. Furthermore, current machine learning and deep learning approaches for soil property predictions were also explored. This work then proposed a neural network, called RDNet, that achieves state-of-the-art results in predicting pH_{H_2O} and pH_{KCl} values from Vis–NIR spectra for a globally distributed set of soil samples. Though receiving positive feedback from the soil scientists and having good predictive results for pH_{H_2O} and pH_{KCl}, these results are only initial work toward completing the discovered objectives. In the future, there should be more work with the domain experts and proximal sensor manufacturers to achieve these five discovered objectives.

Table 3
Summary of the prediction results for the experimented models.

	MSE		R^2		RPD	
pH	H_2O	KCl	H_2O	KCl	H_2O	KCl
PLSR	0.72	0.72	0.63	0.57	1.68	1.68
RF	0.46	0.46	0.77	0.72	2.11	2.11
MLP	0.36	0.34	0.83	0.81	2.46	2.53
DenseNet	0.51	0.47	0.75	0.74	2.06	2.16
VGG	0.35	0.36	0.83	0.80	2.50	2.46
WaveNet	0.32	0.29	0.85	0.84	2.60	2.75
RDNet	**0.28**	**0.25**	**0.86**	**0.86**	**2.76**	**2.93**

CRediT authorship contribution statement

Vung Pham: Methodology, Investigation, Software, Visualization, Writing – original draft. **David C. Weindorf:** Data curation, Validation, Writing – review & editing. **Tommy Dang:** Project administration, Conceptualization, Supervision, Resources.

Declaration of Competing Interest

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.

Acknowledgments

This work was supported by the National Science Foundation under the I-Corps award number 2017018. The authors gratefully acknowledge the contributions of NRCS, NASA, and Olympus representatives.

References

Adler, K., Piikki, K., Söderström, M., Eriksson, J., Alshihabi, O., 2020. Predictions of Cu, Zn, and Cd Concentrations in Soil Using Portable X-Ray Fluorescence Measurements. Sensors 20 (2), 474.

Andrade, R., Faria, W.M., Silva, S.H.G., Chakraborty, S., Weindorf, D.C., Mesquita, L.F., Guilherme, L.R.G., Curi, N., 2020. Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains. Geoderma 357, 113960.

Ba, Y., Liu, J., Han, J., Zhang, X., 2020. Application of Vis-NIR spectroscopy for determination the content of organic matter in saline-alkali soils. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 229, 117863.

Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65 (2), 480–490.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09.

Garrity, D., Bindraban, P., 2004. A globally distributed soil spectral library visible near infrared diffuse reflectance spectra. ICRAF (World Agroforestry Centre)/ISRIC (World Soil Information) Spectral Library: Nairobi, Kenya.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Liu, J., Han, J., Xie, J., Wang, H., Tong, W., Ba, Y., 2020. Assessing heavy metal concentrations in earth-cumulic-orthic-anthrosols soils using Vis-NIR spectroscopy transform coupled with chemometrics. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 226, 117639.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp. 4765–4774.

Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., 2016. Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499.

Pham, V., Dang, T., 2019. SOAViz: Visualization for Portable X-ray Fluorescence Soil Profiles. In: Workshop on Visualisation in Environmental Sciences (EnvirVis), The

[3] http://idatavisualizationlab.github.io/Soil.

[4] http://idatavisualizationlab.github.io/V/globalsoilspectral.

Eurographics Association, Porto, Portugal. https://doi.org/10.2312/envirvis.20191102.
Pham, V., Dang, T., 2020a. ScagnosticsJS: Extended Scatterplot Visual Features for the Web. In: Wilkie, A., Banterle, F. (Eds.), Eurographics 2020 - Short Papers. The Eurographics Association. https://doi.org/10.2312/egs.20201022.
Pham, V., Weindorf, D., Dang, T., 2020. SoilScanner: 3D Visualization for Soil Profiling using Portable X-ray Fluorescence. In: Workshop on Visualisation in Environmental Sciences (EnvirVis), The Eurographics Association, https://doi.org/10.2312/envirvis.20201094.
Savitzky, A., Golay, M.J., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36 (8), 1627–1639.
Silva, S.H.G., Weindorf, D.C., Pinto, L.C., Faria, W.M., Junior, F.W.A., Gomide, L.R., de Mello, J.M., de Pádua Junior, A.L., de Souza, I.A., dos Santos Teixeira, A.F., et al., 2020. Soil texture prediction in tropical soils: A portable X-ray fluorescence spectrometry approach. Geoderma 362, 114136.
Stockmann, U., Cattle, S., Minasny, B., McBratney, A.B., 2016. Utilizing portable X-ray fluorescence spectrometry for in-field investigation of pedogenesis. Catena 139, 220–231.
Van Beers, W., 2005. Kriging metamodeling in discrete-event simulation: an overview, in. In: Proceedings of the 37th conference on Winter simulation, Winter Simulation Conference, pp. 202–208.
Wang, A., Li, D., Huang, B., Lu, Y., et al., 2019. A brief study on using pHH2O to predict pHKCl for acid soils. Agric. Sci. 10 (02), 142.
Xu, Z., Zhao, X., Guo, X., Guo, J., 2019. Deep learning application for predicting soil organic matter content by VIS-NIR spectroscopy. Comput. Intell. Neurosci. 2019.