Multi-source data integration for soil mapping using deep learning

Alexandre M.J-C. Wadoux [1], José Padarian [2], and Budiman Minasny [2]

[1]Soil Geography and Landscape group, Wageningen University & Research
[2]Sydney Institute of Agriculture, The University of Sydney

Correspondence: Alexandre Wadoux (alexandre.wadoux@wur.nl)

Abstract. With the advances of new proximal soil sensing technologies, soil properties can be inferred by a variety of sensors, each having its distinct level of accuracy. This measurement error affects subsequent modelling and therefore must be integrated when calibrating a spatial prediction model. This paper introduces a deep learning model for contextual Digital Soil Mapping (DSM) using uncertain measurements of the soil property. The deep learning model, called Convolutional Neural Network (CNN), has the advantage that it uses as input a local representation of environmental covariates to leverage the spatial information contained in the vicinity of a location. Spatial non-linear relationships between measured soil properties and [c1]neighboring covariate pixel values are found by optimizing an objective function, which can be weighted with respect to a measurement error of soil observations. In addition, a single model can be trained to predict a soil property at different soil depths. This method is tested in mapping top- and subsoil organic carbon using laboratory analyzed and spectroscopically inferred measurements. Results show that CNN significantly increased prediction accuracy as indicated by the coefficient of determination and concordance correlation coefficient, when compared to a conventional DSM technique. Deeper soil layer prediction error decreased, while preserving the interrelation between soil property and depths. The tests conducted [c2]suggest that CNN benefits from using local contextual information up to 260 to 360 metres. We conclude that CNN is a flexible, effective and promising model to predict soil properties at multiple depths while accounting for contextual covariates information and measurement error.

[c1] *Text added.*

[c2] ~~using different window size of input covariates matrix to predict organic carbon~~

1 Introduction

Digital Soil Mapping (DSM) techniques are [c3] commonly used to predict a soil property at unsampled locations using measurements at a finite number of spatial locations. Prediction is routinely done by exploiting the relationship between a soil property and one or several environmental covariates, which are assumed to represent soil forming factors. Examples of covariates are Digital Elevation Model (DEM) or its derivatives (Moore et al., 1993). Demattê et al. (2018) used multi-temporal and multispectral remote sensing images to map soil spectral reflectance while Nussbaum et al. (2018) investigated the use of a large set of covariates for mapping eight soil properties at four soil depths. The choice of covariates is governed either by their availability, pre-selected using *a priori* pedological expertise, or based on the pedological concepts whereby covariates must portray the factors of soil formation such as climate, organisms, relief, parent material and time (McBratney et al., 2018)[c4].

[c3] ~~now~~

[c4] reference added

In most cases, the relation between soil property and the chosen covariates is modelled by a regression [c5]model which relates either linearly (Wadoux et al., 2018) or non-linearly (Grimm et al., 2008) sampled (point) soil properties and a vector [c6]of[c7] covariates value extracted at same point location.

[c5] *Text added.*
[c6] *Text added.*
[c7] ~~a~~

Several authors have shown that this is not satisfactory (e.g., Moran and Bui, 2002). Pedogenesis and thus soil properties spatial variation is governed by complex relationships with soil forming factors and landscape characteristics, materialized at a local, regional or supra-regional scale (Behrens et al., 2014). Point information of the covariates can only describe approximately the soil property because a large part of the spatial contextual information is missing. For example, soils on a gentle slope might have a great accumulation of soil organic matter, accumulation which varies according to the surrounding slope gradients. Several studies have shown that incorporating covariates contextual information improves prediction accuracy (Behrens et al., 2014; Grinand et al., 2008; Gallant and Dowling, 2003). Smith et al. (2006) tested different neighbouring size in computing terrain attributes for use in a soil survey. The authors showed that the amount of contextual information supplied to the model significantly impacts the output of the survey. In spite of these conclusions, contextual information surrounding a sampling location is usually disregarded in DSM studies.

Several attempts have been made to incorporate the spatial domain of the covariates into the analysis. Behrens et al. (2010a) developed ConMAP which computes the elevation difference from the centre pixel to each pixel in a neighborhood and ConStat (Behrens et al., 2014) which derives statistical measures of elevation within a growing radius of the centre pixel. This generates a very large number of hyper-covariates, abstract representation of the context, which can be used as predictor in subsequent regression models. Another approach uses spatial transform such as wavelet to represent the covariate as a function of various local spatial support (e.g., Lark and Webster, 2001). Alternatively, one may account for contextual information by simply using covariates aggregated at larger support than their original resolution (Miller et al., 2015). This technique, referred to as the multi-scale approach, provides surprisingly large increase of prediction accuracy. It is now acknowledged that using covariates with coarse spatial resolution can provide satisfactory prediction (Samuel-Rosa et al., 2015).

However, while [c1]these[c2] approaches enable to contextualize the spatial information supplied to the regression, they rely either on heavy covariates pre-processing (Behrens et al., 2014), subjective decisions based on the resolution to which covariates must be treated as input to the model (Miller et al., 2015) or modeller's choice regarding neighbouring size (Behrens et al., 2010b). In light of these drawbacks, we propose to use Convolutional Neural Network (CNN) as a[c3]n alternative tool for mapping while explicitly accounting for local contextual information contained in covariates. Recently, Padarian et al. (2018) have shown that is is possible to use CNN for soil mapping while accounting for contextual covariate information while Behrens et al. (2018) compared deep neural network to random forest for mapping and found that the former model provides more accurate predictions. [c4]In Behrens et al. (2018)[c5], the covariates must still be pre-processed. The authors used a deep learning architecture which uses a vector as input. The CNN proposed here has the advantage that it relies on the local representation of covariates so as to leverage the spatial information contained in the vicinity of a sampled point. [c6]CNN uses an image as input, does not require any pre-processing of the input covariates and performs a multi-scaling analysis directly on the image (Padarian et al., 2018). As for other regression methods, CNN is trained using measured soil properties at point location.

[c1] *Text added.*
[c2] ~~this~~
[c3] *Text added.*
[c4] *Text added.*
[c5] *Text added.*
[c6] *Text added.*

Measured soil properties are never error-free. Soil measurements can be at best performed under controlled conditions in the laboratory. In the latter case, the error of those measurements is small and their impact on prediction is safely ignored. With the advent of new technology, soil measurements are often inferred using sensors such as spectrometers. The result is the creation of databases of soil properties measured or inferred using several sensors which predicted soil properties with different accuracy levels. Recently, Ramirez-Lopez et al. (2019) and Somarathna et al. (2018) have shown that measurement error may have a significant impact in subsequent spatial analysis. For example, Ramirez-Lopez et al. (2019) estimated a measurement error of about 50% for top- and subsoil Ca^{++} inferred using Near-Infrared (NIR) spectroscopy. In most cases, measurement error can be quantified and must therefore be accounted for when calibrating a spatial model using uncertain measurements. While Padarian et al. (2018) demonstrated the use of CNN at a country extent mapping, we further advanced this concept for mapping soil properties at a landscape scale which consider measurement error of soil measurements.

The objectives of this study are to (i) develop the framework of Convolutional Neural Network for contextual spatial modelling [c1]at a regional scale, (ii) [c2]develop a methodology for multi-source data integration by account[c3]ing for the soil property measurement error in the CNN model calibration and (iii) demonstrate the usefulness of CNN to map top- and subsoil organic carbon in a potential application scenario.

[c1] *Text added.*
[c2] *Text added.*
[c3] *Text added.*

2 Methodology

2.1 [c4]Artificial Neural Network[c5]

[c4] *Text added.*
[c5] ~~Model definition~~

We first describe the principle of [c6]Artificial Neural Network ([c7]ANN), basis of CNN. A measured soil property of interest z_{s_i} at location $s_i (i = 1, \ldots, n; s_i \in \mathcal{A})$ in the study area $\mathcal{A}$ is modelled by a [c8][c9]regression model:

[c6] *Text added.*
[c7] *Text added.*
[c8] ~~Artificial Neural Network (ANN)~~
[c9] *Text added.*

$$z_{s_i} = f(\mathbf{X}_{s_i}; \boldsymbol{\theta}) + \varepsilon_{s_i}, \tag{1}$$

where $\mathbf{X}$ is either a $c \times (w \times h)$ 2-D matrix or a 3-D input matrix of size $c \times w \times h$ which contains c environmental covariates of size $w \times h$ centred at spatial location s_i. The vector $\boldsymbol{\theta}$ are model parameters used by the neural network regression model f to map non-linearly $\mathbf{X} \rightarrow z$ and leave room for a zero mean random error ε. Note that unlike [c10]geostatistics, measurements of the soil property are assumed [c11]spatially independent and identically distributed.

[c10] ~~classical~~
[c11] *Text added.*

An ANN model is formed of several layers, or "computation steps". The input layer provides the raw information to the network [c12](see h^0 Fig. 1b), which is connected to at least one hidden layer [c13](h^k in Fig. 1b), which in turn is connected to an output layer[c14](h^L in Fig. 1b), which outputs the predictions of interest z_{s_i}. Each layer contains units, [c15][c16]called nodes for the input layer (as no computation occurs at their level) and neurons for the hidden and output layers (Fig. 1a). The behaviour of the neurons depends on the activity of the previous layer neurons and the [c17]connection weights between the previous and [c18]next[c19] layer neurons (LeCun et al., 2015). The parameters of the models defined by Eq. 1 are thus the connection weights to the neuron j, $\mathbf{w}_j = (w_{j,1}, \ldots, w_{j,l})$ and a bias component per neuron b_j. They are associated to an activation function ϕ which

[c12] *Text added.*
[c13] *Text added.*
[c14] *Text added.*
[c15] ~~called neurons~~
[c16] *Text added.*
[c17] *Text added.*
[c18] *Text added.*
[c19] ~~new~~

gives output z_j by [c20](Fig. 1a):

$$z_j = \phi(\langle \mathbf{w}_j, \mathbf{x} \rangle + b_j), \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the dot product and $\mathbf{x}$ is a vector of inputs from previous layer neurons output. A graphical representation is provided in Fig. 1a. In this study we use the [c1]Rectified Linear Unit (ReLU)[c2] activation function $\phi(x) = \max(0, x)$ while the output layer for regression uses the linear activation function $\phi'(x) = x$. [c3]Eq. 2 shows that the activation function determines the output of a neuron by applying a linear or non-linear transform on the input..

[c20] *Text added.*
[c1] *Text added.*
[c2] ~~ReLU~~
[c3] *Text added.*

Our neural network will contain more than one single hidden layer. For $k = 1, \ldots, L$ hidden layers [c4]h and an input layer h^0 where no processing occurs (Fig. 1b):

[c4] *Text added.*

$$h^0(x) = x, \tag{3}$$

$$h^k = \mathrm{ReLU}(\mathbf{W}^k h^{k-1}(x)), \quad \text{for } k = 1, \ldots, L-1 \tag{4}$$

$$h^L = \mathbf{W}^L h^{L-1}(x). \tag{5}$$

For each layer, $\mathbf{W}$ is a matrix of size $J^k \times J^{k-1}$, i.e. the number of neurons in the current layer by the number of neurons in the previous layer. Therefore the model parameters $\boldsymbol{\theta} = (\mathbf{W}^1, b^1, \ldots, \mathbf{W}^L, b^L)$. [c5]

[c5] Figure 1 caption has been updated

2.2 Convolutional Neural Network

In this paper we use the vicinity information of the measured soil property [c6]by including the covariate pixel values surrounding a sampling location. In this case, an ANN is not well adapted because [c7]it uses[c8] vectors as input data, while (correlated) spatial information is better represented as images. In convolutional neural network, at least one layer is a convolution (Goodfellow et al., 2016)[c9], i.e. an element-wise product and sum between two matrices. Let there be an input image matrix $\mathbf{X}$, e.g. a Digital Elevation Model (DEM) cropped for size $w \times h$ pixels surrounding a measured soil property at location s_i. We apply a 2D convolution using the filter F of size $m \times m'$ to the input image $\mathbf{X}$ such that:

[c6] *Text added.*
[c7] *Text added.*
[c8] ~~they deal with~~
[c9] *Text added.*

$$(F * \mathbf{X})_{(w,h)} = \sum_{m,m'} F_{(m,m')} \mathbf{X}_{(w+m', h+m)}, \tag{6}$$

which can be rewritten with little modification to include the case where we have $c = 3$ environmental covariates [c10](Goodfellow et al., 2016, Eq. 9.4). Eq. 6 shows that each element in $(F * \mathbf{X})$ is calculated as the sum of the products of one element in $\mathbf{X}$ and one element in F. In other words, the elements of $(F * \mathbf{X})$ are the the sum of the element-wise multiplication of F by $\mathbf{X}$. [c11]The size of the output image from a convolution is thus smaller than that of its input image (see Fig. 1c).

[c10] Here reference added
[c11] *Text added.*

Filters detect features [c12](e.g. edges) related in the vicinity of a sampling location and leverage the spatial structure of the covariates. In practice, the original covariate image go through several filters, each exploiting an abstract representation of the image features. Similar to ANN, CNN has a number of hidden layers, called convolutional layers. The convolutions are combined with an activation function at the end of each neuron, obtained by:

[c12] *Text added.*

$$z_j(\mathbf{X}) = \phi(F^L * h^{L-1} + b_j). \tag{7}$$

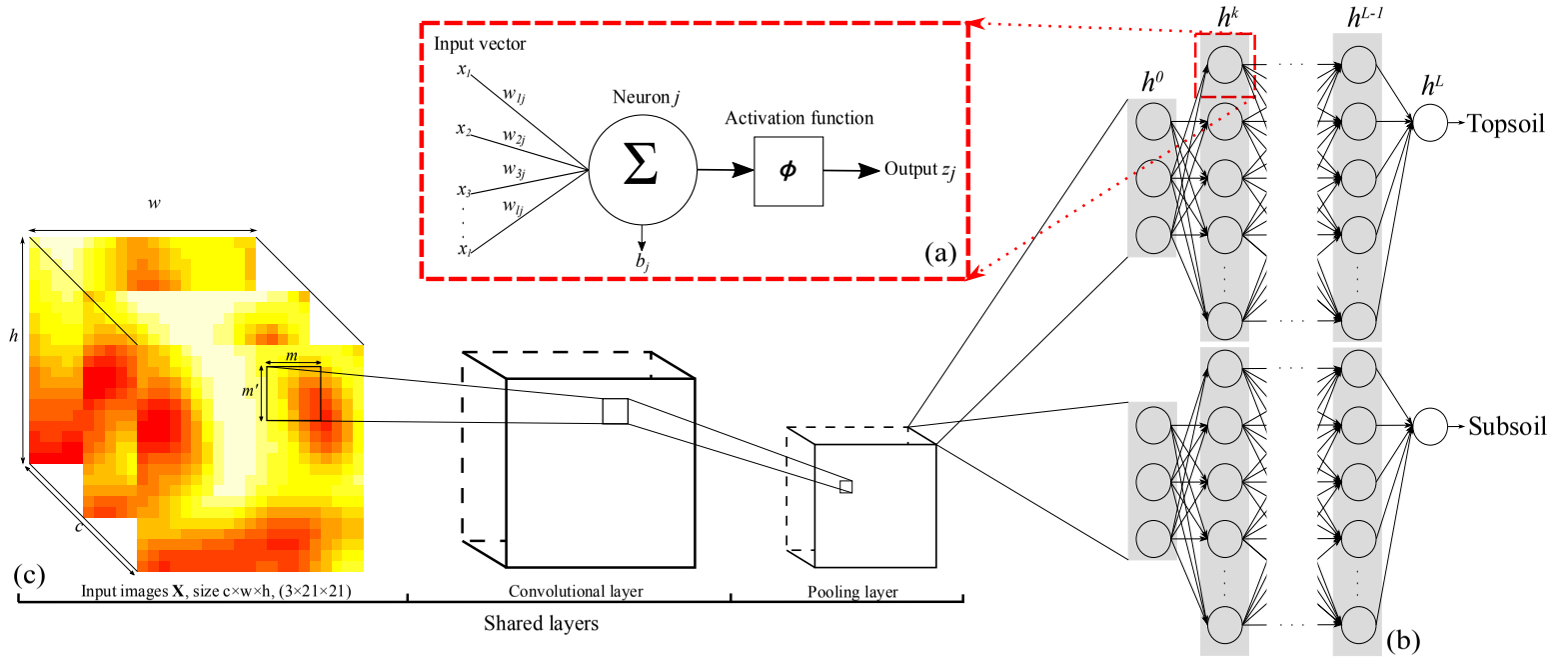

Figure 1. Representation of the CNN architecture developed in this study for (a) a neuron, (b) the ANN architecture or fully connected layer and (c) the convolutional and pooling layers. Note that $\sum = \langle \mathbf{w}_j, \mathbf{x} \rangle$. The size of the output from a convolutional or pooling layers (c) is smaller than that of its input. Part (c) is the shared structure to extract common features between the two depths which is then separated into two branches (one per soil depth) in part (b).

In addition to the convolutional layers, another set of operation consists of pooling layers. Pooling reduce the spatial size of the images by down-sampling along the spatial dimension. [c13]Several types of pooling operations exist, such as minimum, maximum or average pooling. The most common is the max-pooling operation, which consists in selecting the maximum value in the convoluted image using a given filter size. Each convolution accepts one input image of a given size and number of channels [c14]i.e. covariates[c15], and returns another image of possibly a different size and number of channels. Usually, one want to reduce the size of each image at each convolution, while augmenting the number of channels. Then, the last convolution returns an image of size 1×1 and with a number of channels. [c16]This operation is named "flattening", as it converts the matrices into a vector which can pass to a fully connected layer. A fully connected layer is an ANN layer, as noted in Fig. 1b. The large number of connections between neurons generate a high number of parameters which can provoke overfitting. This can be restrained by introducing dropout layers which randomly disconnect a number of neurons. Usually, the dropout rate is no greater than 0.5.[c17]

[c13] *Text added.*
[c14] *Text added.*
[c15] ~~(aka depth)~~
[c16] *Text added.*
[c17] ~~This is a vector that we can pass to a fully connected layer.~~

2.3 Parameter estimation

The CNN model is trained on dataset $\mathcal{D} = \{(\mathbf{X}_{s_i}, z_{s_i}) \dots (\mathbf{X}_{s_n}, z_{s_n})\}$, that is, a 4-D matrix of size $n \times c \times w \times h$. The dataset $\mathcal{D}$ is used to derive an optimized value of the parameters $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ by minimizing the mean squared error (MSE) as objective function, given by:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \delta_i (z_{s_i} - \hat{f}(\mathbf{X}_{s_i}; \hat{\boldsymbol{\theta}}))^2, \tag{8}$$

where $\boldsymbol{\delta}$ are the [c1]measurement error weights, which are all 1 if the soil property is measured without error. In this study, we used the Adam optimizer (Kingma and Ba, 2014) to minimize Eq. 8. Adam uses the derivative of the objective function with respect to each model parameter to update its value. This process is called backpropagation (LeCun et al., 1989). The optimization process runs for a number of epochs. An epoch describes the number of times the network sees the entire input dataset. During each epoch, the entire dataset is shown to the network in small subsets shuffled at random, called "minibatches". The number of epochs must be chosen, as well as the batch size. In addition, one must chose the learning rate of the optimizer, i.e. how fast the optimizer moves the [c2]connection weights in the opposite direction of the gradient after each update. A too small learning rate increases the computation time to find the optimum of the objective function because the steps are small. If the learning rate is too large training may not converge because the [c3]connection weights oscillate. [c4]The learning rate can be tuned as other model architecture hyperparameters. This is explained in the next sections.

[c1] *Text added.*
[c2] *Text added.*
[c3] *Text added.*
[c4] *Text added.*

2.4 Multi-source data integration

The values of the soil property z_{s_i} used to train the CNN model might be uncertain. For example, they are derived using an infrared spectroscopy model. This uncertainty must be accounted for when calibrating the CNN model. A solution is to assign a [c5]measurement error weight to each value of the soil property, depending of its relative error compared to a [c6c7]"true"

[c5] *Text added.*
[c6] ~~true~~
[c7] *Text added.*

measurement of the soil property at the same location. [c8]We refer "true measurement" of a soil property as a measurement made using a standard laboratory technique which has a small error. Meanwhile spectroscopy inferred property is the value predicted using measurement from an infrared spectrometer. The prediction is based on a calibration model that relates observations of spectra and their true measurement values. For a vector of a soil property inferred using a spectroscopic model $\mathbf{z}_{IR}$, one can assign a [c9]measurement error weight by comparing the variance of the predicted soil property by the infrared model to the variance of the measurements used for the infrared model calibration. The [c10]measurement error weight δ for is given by:

[c8] *Text added.*
[c9] *Text added.*
[c10] *Text added.*

$$\delta = 1 - \frac{\text{var}(\mathbf{z}_{IR})}{\text{var}(\mathbf{z})}, \tag{9}$$

which can then be applied to weight importance of the values of the soil property inferred by the infrared model, at locations where the true soil property is unknown. [c1]Note that while all measurements have a measurement error weight, the same value of measurement error is used for a given spectroscopic model (NIR or MIR) and soil depth. In this work, a true value of a soil property is measured in the laboratory, with assigned [c2]measurement error weight of 1. The [c3]measurement error weights for each observation are used in model calibration, by updating objective function to minimize in Eq. 8.

[c1] *Text added.*
[c2] *Text added.*
[c3] *Text added.*

2.5 Quality of predictions

Once model parameter vector $\boldsymbol{\theta}$ have been estimated, they are used to predict at new, unobserved location s_0 by:

$$\hat{z}_{s_0} = \hat{f}(\mathbf{X}_{s_0}; \hat{\boldsymbol{\theta}}), \tag{10}$$

which is used to evaluate the prediction accuracy on an independent test set. Let there be $N - n$ independent test locations $s_i, i = (N-n), \ldots, N)$ where N is the total number of measurements and n is the set of samples used for calibration, generally 80% of the measured values. We quantify the quality of predictions by the Root Mean Squared Error (RMSE)[c4] and the R^2. The bias is assessed by the Mean Error (ME):

[c4] Formulas RMSE and R2 removed

$$\text{ME} = \frac{\sum_{i=1}^{N-n}(z_{s_i} - \hat{z}_{s_i})}{N-n}, \tag{11}$$

and the agreement of the predictions to the measurements with respect to the 1:1 line is assessed by the Concordance Correlation Coefficient (ρ) (Lawrence and Lin, 1989):

$$\rho = \frac{2\rho'\sigma_{\mathbf{z}}\sigma_{\hat{\mathbf{z}}}}{\sigma_{\mathbf{z}}^2 + \sigma_{\hat{\mathbf{z}}}^2 + (\mu_{\mathbf{z}} - \mu_{\hat{\mathbf{z}}})^2}, \tag{12}$$

where μ and $\sigma^2$are mean and variance for either the vector of true measurements $\mathbf{z}$ or the vector of predicted values $\hat{\mathbf{z}}$. The value ρ' represents the correlation between $\mu_{\mathbf{z}}$ and $\mu_{\hat{\mathbf{z}}}$.

3 Case study

3.1 Study area and data

We tested the methodology in a 220 km^2 area located in the lower Hunter valley area, Australia. Elevation ranges from 27 to 322 m above sea level with a pronounced slope ascending South-West. Measurements of the Total Soil Carbon (TC) expressed in g/100g^{-1} are available for topsoil (0-10 cm) and subsoil (40-50 cm). The lower Hunter area measurements were surveyed along several years, which yielded the use of three TC measurement methods, denoted CNS, NIR and MIR hereafter:

- Laboratory analysis (CNS). Soil samples were analyzed for TC using the dry combustion method, i.e. by determining the loss on ignition at 400°C under controlled conditions. This was done by an ElementarVario Max CNS analyser (Elementar Analysesysteme GmbH, Hanau, Germany). The standard deviation of the TC values inferred by the latter device is small (less than 0.004 g/100g^{-1}).
- Inferred using Near-Infrared (NIR) spectroscopy. Soil samples were scanned in the NIR range using an Agrispec portable spectrophotometer with a contact probe attachment (Analytical SpectralDevices, Boulder, Colorado). TC values were inferred using a spectroscopic model calibrated by the cubist regression tree method, using the spectral library of 316 soil samples from Geeves et al. (1995).
- Inferred using Mid-Infrared (MIR) spectroscopy. Soil samples were scanned in the MIR region using a Bruker TENSOR 37 Fourier transform spectrometer. TC values were inferred using the MIR calibration model defined by Minasny et al. (2008).

A large number of location contains more than one single measurement of TC. This is particularly visible in the western part of the area, where many samples have been analyzed using the two or three methods, and with a replication (Fig. 2). In total, [c1][c2]2,388 measurements of TC are available for the first depth, among which [c3][c4]645 are from the CNS methods, [c5][c6]923 for the NIR and [c7][c8]820 from the MIR method. In the second depth, there are [c9][c10]2,058 measurements of the TC: [c11][c12]187 using the CNS method, [c13][c14]999 using the NIR method and [c15][c16]872 using the MIR method. They are shown in Fig. 2.

In addition to the TC measurements, three covariates from the study of Somarathna et al. (2018) at 25 × 25 m resolution were used:

- A Digital Elevation Model (DEM) from the SRTM (Shuttle Radar Topography Mission)(Fig. 3a), see Farr et al. (2007).
- A map of the Landsat 5 ETM band 5 (Fig. 3b), which corresponds to the Shortwave Infrared (SWIR) band for the wavelength 1.55-1.75 μm.
- A map of the normalized difference vegetation index (NDVI) (Fig. 3c), derived from the NIR (band 4) and red (band 3) of the Landsat 5 ETM sensor.

[c1] ~~2,962~~
[c2] *Text added.*
[c3] ~~722~~
[c4] *Text added.*
[c5] ~~1,146~~
[c6] *Text added.*
[c7] ~~1,094~~
[c8] *Text added.*
[c9] ~~2,500~~
[c10] *Text added.*
[c11] ~~304~~
[c12] *Text added.*
[c13] ~~1,165~~
[c14] *Text added.*
[c15] ~~1,031~~
[c16] *Text added.*

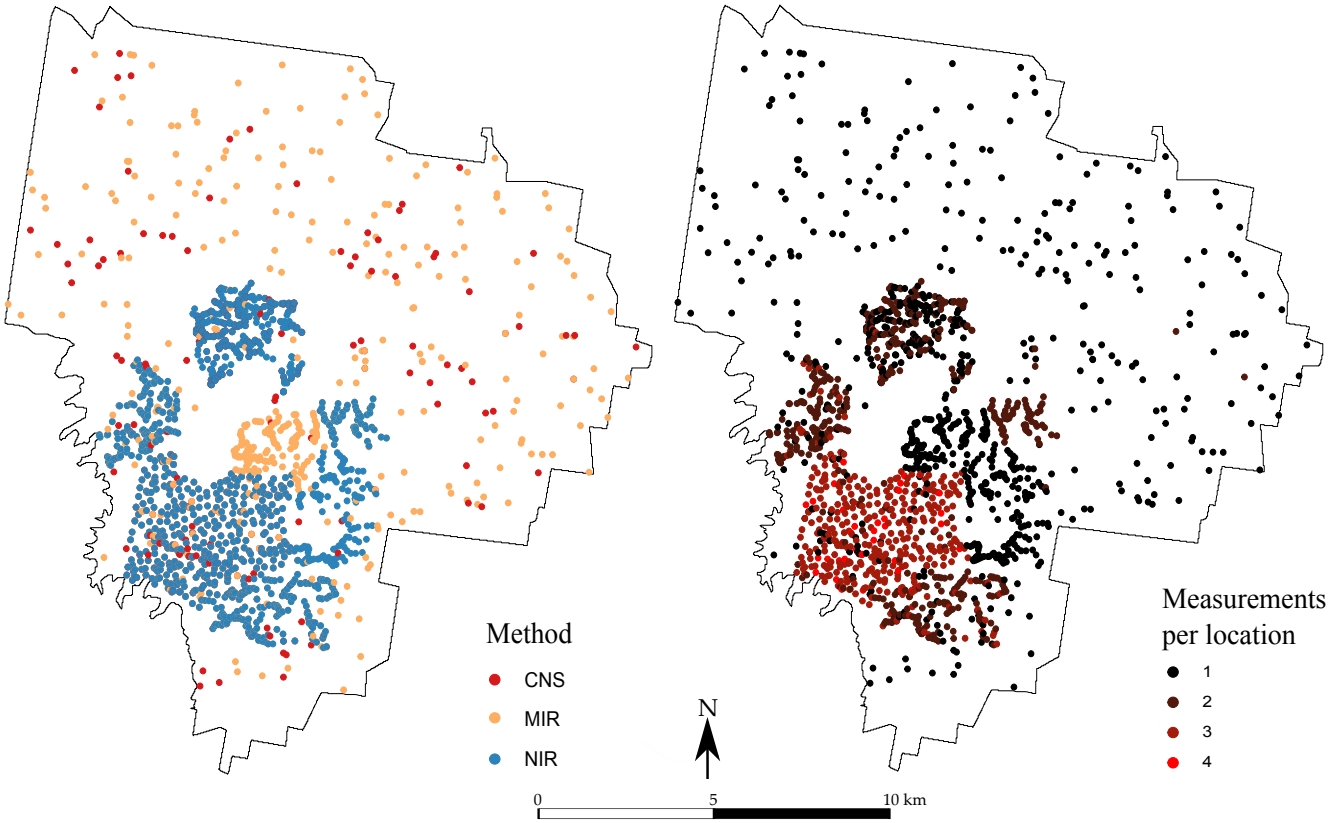

Figure 2. Spatial distribution of the observations for each measurement type (left) for the 0-10 cm topsoil and (right) number of measurements recorded per sampling location for the 0-10 cm topsoil. The subsoil (40-50 cm) map is not shown but closely resembles the topsoil in both number of sampling locations and number of measurements per location.

3.2 Practical implementation

3.2.1 Model definition

The dataset of TC measurements was randomly splitted between test (20%) and calibration (80%) sets. Both topsoil and subsoil measurements were jointly selected for either test or calibration. All soil measurements were normalized between 0 and 1 using the minimum and maximum values of the calibration set. In addition, all covariates were centred on 0 and scaled to a standard deviation of 1 (see Fig. 3). Next, two 4-D matrices of dimension $n \times c \times w \times h$ and $(N-n) \times c \times w \times h$ were created (test and calibration), where n is the number of [c1] calibration TC measurements, $N-n$ is the number of test measurements, c is the number of covariates and $w = h$ is the vicinity size (the square matrix) surrounding the TC measurements. We have [c2c3] $n = 3,557$ for calibration and [c4c5] $(N-n) = 889$ for test, $c = 3$ and $w = h$ of different sizes. When a square of size $w \times h$ is created in the vicinity of a soil property at the border of the area, several missing values are reported. Since CNN can not

[c1] *Text added.*

[c2] ~~$n = 3,5$~~

[c3] *Text added.*

[c4] ~~$(N-n)$~~

[c5] *Text added.*

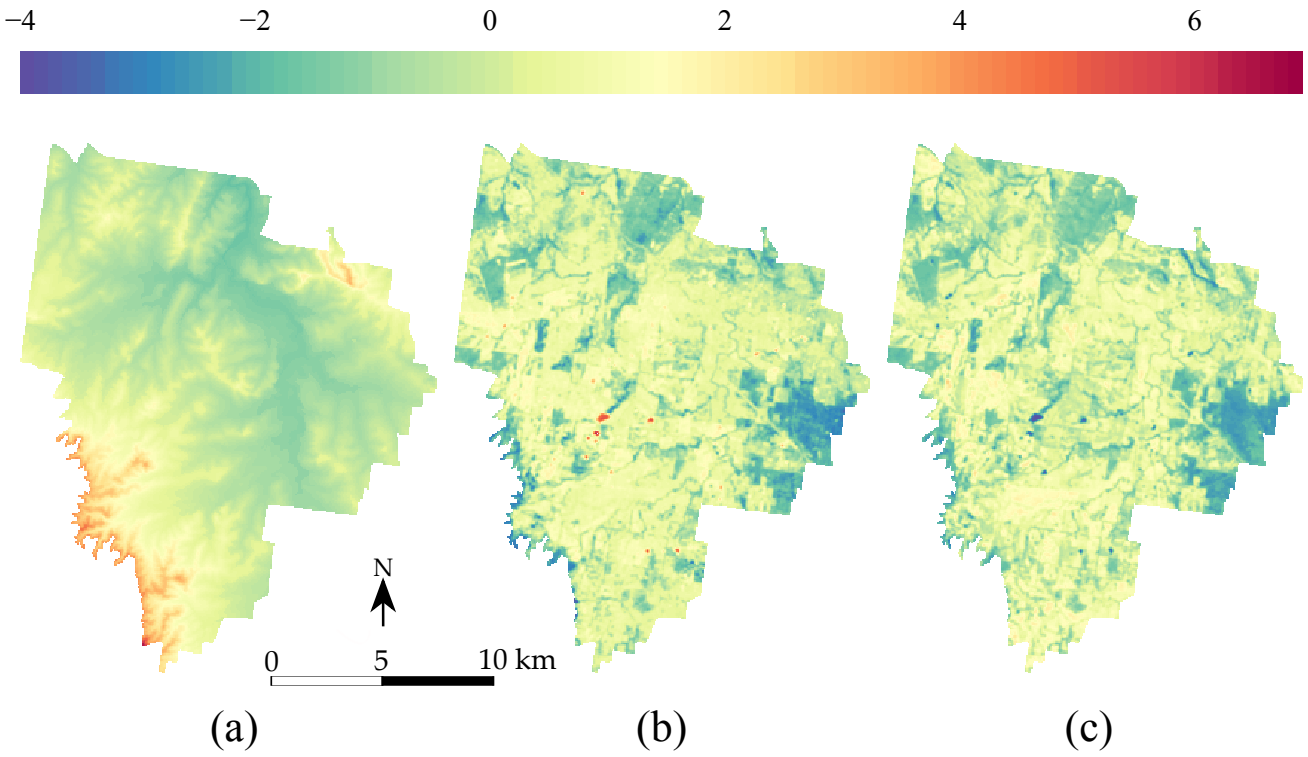

Figure 3. Standardized covariates used in model calibration (a) DEM, (b) Landsat 5 ETM band 5 and (c) NDVI.

handle this type of input we assigned to the missing values the number -1. This practical problem is discussed more extensively in the Discussion section.

A sequential multi-task (top- and subsoil) CNN model was built. The CNN is composed of a common architecture for the two soil depths (shared layers) followed by two separate sets of fully connected layers, one for each soil depth. An illustration of the model is provided in Fig. 1b-c. The model specifications are reported in Table. 1. Note that for the convolutional layers, zero padding is always applied to the original input image before the dot product with the filters. This operation keeps the original size of the input image and preserve its information at an early stage of the model. [c1]We use the max-pooling operator, i.e. we select the maximum value in the convoluted image using a given filter size, in our case a filter of size 2×2 pixels.

In order to compare the CNN prediction to a reference method, we also calibrated [c2]two[c3] Random forest (RF) model[c4]s, one per soil depth. Random forest is a non-linear machine learning method which have been widely used for soil mapping. For more information, the reader is redirected to Hengl et al. (2018). For a fair comparison between the CNN and RF models, we used the same calibration/test sets, with normalized TC measurements and standardized covariates as input. A RF forest model is calibrated for each soil depth, using [c5]depth-specific fine-tuned parameter values.[c6].

[c1] *Text added.*

[c2] *Text added.*

[c3] ~~an univariate~~

[c4] *Text added.*

[c5] *Text added.*

[c6] ~~1000 trees~~

Table 1. Layers used in the sequential model built for topsoil and subsoil TC prediction.

Order	Layer type	Shared	Filter size	Number of filters/neurons	Activation
1	Convolutional	yes	3×3	64	ReLU
2	Max-pooling	yes	2×2	-	-
3	Convolutional	yes	2×2	32	ReLU
4	Dropout (0.3)	yes	-	-	-
5	Flatten	yes	-	-	-
6	Fully-connected	no	-	40	ReLU
7	Dropout (0.3)	no	-	-	-
8	Fully-connected	no	-	50	ReLU
9	Dropout (0.2)	no	-	-	-
10	Fully-connected	no	-	1	Linear

3.2.2 Parameter estimation

Once the sequential CNN model was specified, the parameters were estimated by minimizing the MSE as objective function (Eq. 8), for which we used the Adam optimizer. Overfitting was carefully checked by (i) modifying the dropout rate in the model and (ii) ensuring that the model does not provide a considerably larger objective function on an independent dataset. Since the test set was used solely to validate the predictions it could not be used to this purpose. We therefore randomly splitted the calibration set into two sets, denoted calibration and validation sets hereafter. The calibration set (90%, c1c2 3,201 measurements) was used to calibrate the model and the validation set (10%, c3c4 356 measurements) was used to tune the parameters and prevent overfitting.

The model was trained using different window size of the input images. We compared a model with window size ($w \times h$) of 3, 5, 9, 15, 21, 29 and 35 c5 covariate pixels for the input. The comparisons were made based on the c6c7 average between the two depths RMSE of the predictions, made on the validation set. Optimization of the parameter of a single model (Table. 1) using 3 covariates, an input window size of 21 $\times$ 21 and c8c9 3,557 TC measurements took approximately 1 hour in parallel on a standard 4-cores laptop. All processing was done in R 3.5.1 (R Core Team, 2018), using the keras package (Allaire and Chollet, 2018) and tensoflow (Abadi et al., 2016) backend.

Once the windows size selected, the hyperparameters of the model architecture were optimized using Bayesian optimization (Snoek et al., 2012). Note that this is different from the optimization of the objective function using Adam. In Bayesian optimization, the objective function is treated as a random function characterized by a prior probability distribution. Each function evaluation is treated as data which enables updating the objective function posterior distribution. The latter is used to determine where to evaluate next. The process is repeated until reaching a stopping criteria. Bayesian optimization enables to find optimized values of machine learning hyperparameters with commonly less iteration than when using a random search. In this work, we optimized the filters number, the neurons number, the batch size and the learning rate using 50 iterations.

c1 ~~3,167~~
c2 *Text added.*
c3 ~~352~~
c4 *Text added.*
c5 *Text added.*
c6 ~~depth~~ ~~averaged~~
c7 *Text added.*
c8 ~~3,519~~
c9 *Text added.*

Table 2. Weights given to the measurements.

	CNS	NIR	MIR
Topsoil	1	0.43	0.62
Subsoil	1	0.52	0.61

4 Results

Based on the procedure detailed in Section 2.4, measurements of TC inferred from NIR spectra were assigned a [c10]measurement error weight of 0.43 and 0.52 for topsoil and subsoil, respectively. The MIR inferred TC measurements had a [c11]measurement error weight of 0.62 for topsoil and 0.61 for subsoil. This suggests that the MIR range of the spectra is more accurate in predicting TC. Recall that all CNS inferred measurements had a [c12]measurement error weight of 1, as explained in the previous section.

[c10] *Text added.*

[c11] *Text added.*

[c12] *Text added.*

Figure 4 shows the RMSE of topsoil and subsoil TC for different vicinity size of the input images. Contextual information is accounted for by representing the input data as images of a square format surrounding a soil measurement. Each pixel has a resolution of 25 $\times$ 25 m so that a window of size 3 $\times$ 3 includes contextual information up to 3/2 $\times$ 25 = 37.5 m. For both soil depths, Fig. 4 shows a similar pattern with increasing size of the window. The RMSE becomes significantly smaller when using a larger window of size 5 $\times$ 5. The lowest averaged (topsoil and subsoil) RMSE is found for a window size of 21 $\times$ 21 (radius of about 262 meters). It seems that model calibration does not benefit from using a larger window size as the RMSE increases for a window size of 29 $\times$ 29 and 35 $\times$ 35. From now on, all the results presented come from using an input window size of 21 $\times$ 21 pixels.

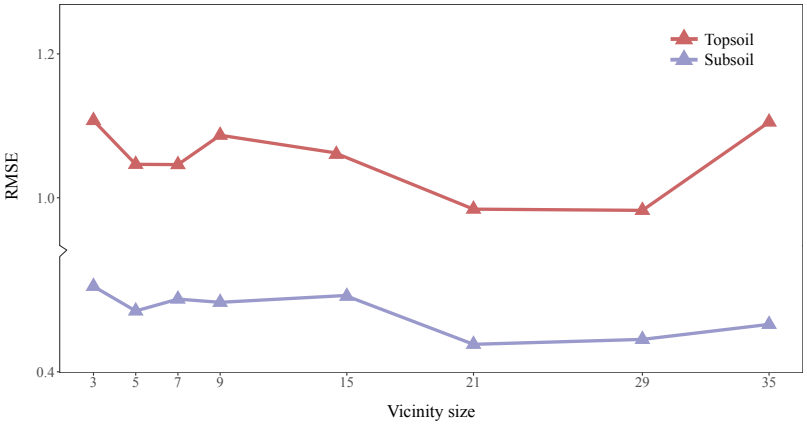

Figure 4. Effect of the vicinity size of input images. The RMSE corresponds to the error between the predictions and measured values in the test set.

The scatterplots of the measured against predicted TC values are presented in Fig. 5 for both the CNN and RF models. For the CNN model, the agreement between measured and predicted TC was found to be satisfactory for both soil depths. Topsoil

predicted TC tends to be underestimated for large measured values of TC. This is also the case for subsoil where large measured values of TC (e.g. 7.5 and 8 g/100g^{-1}) have smaller predicted values at around 4 g/100g^{-1}. High density of predicted values are close to the 1:1 line. In contrast to the CNN model, the predictions using the RF model are more dispersed, with several over-predicted values for low-range of measured TC values. Visual inspection of Fig. 5 suggests that the CNN model predicts more accurately than the RF model.

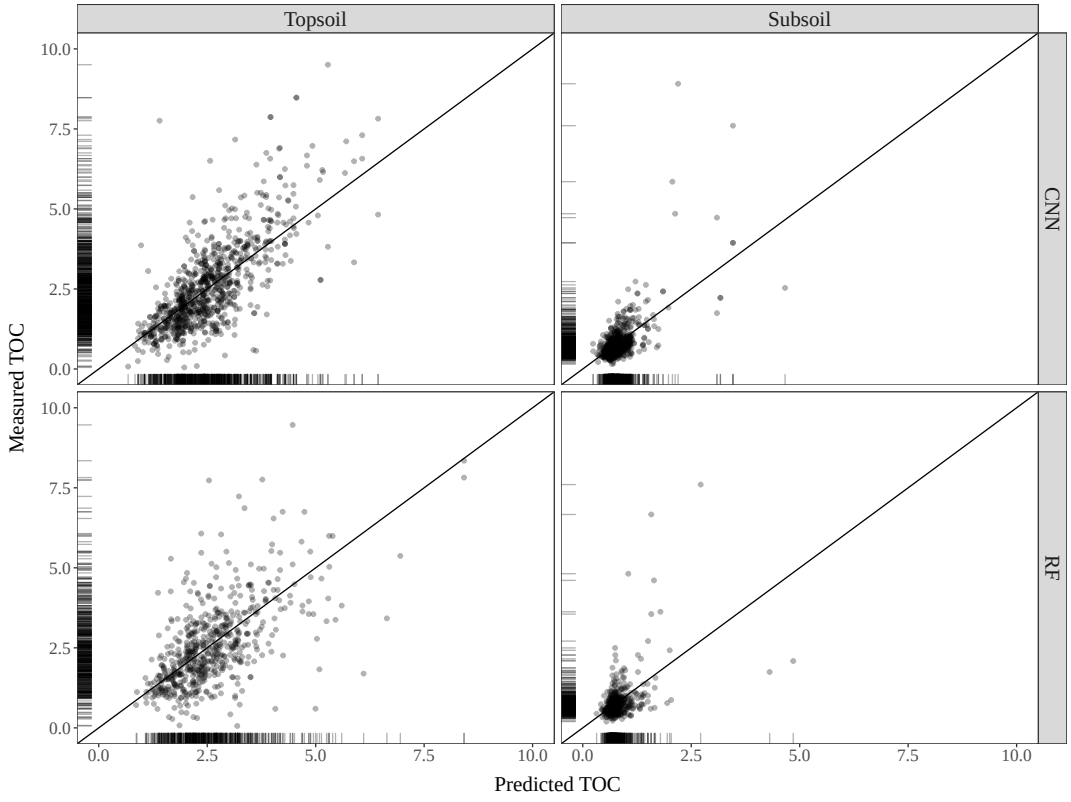

Figure 5. Scatterplot of the measured against predicted topsoil and subsoil TC for the CNN and RF models, along with the 1:1 line. Values are expressed in g/100g^{-1}.

This is confirmed by the quantitative assessments of the predictions shown in Table 3. Correlation between predicted and measured TC, as measured by the R^2, is stronger for CNN (R^2 = 0.55 for topsoil and R^2 = 0.46 for subsoil) than for RF (R^2 = 0.35 for topsoil and R^2 = 0.21 for subsoil). The ME shows that predictions for both models are relatively unbiased (ME close to zero in all cases). CNN model provides a significantly smaller accuracy measure (topsoil RMSE of 0.93 against [c1][c2]1.08 for the RF model) while providing as well larger degree of prediction falling on the 45° line through the origin (about 15% higher for both topsoil and subsoil), as already noticed visually in Fig. 5.

The maps produced using CNN are shown in Fig. 6 for topsoil (left) and subsoil (right) soil organic carbon. Both maps have a relatively smooth pattern. The topsoil map of TC shows the highest concentrations in the South-East border of the area (>

[c1] ~~1.07~~
[c2] *Text added.*
[c3] Table 3 values have been updated

Table 3. Evaluation of prediction accuracy on the independent test set.[c3]

	R^2	ME	RMSE	ρ
Convolutional Neural Network				
Topsoil	0.55	0.04	0.93	0.68
Subsoil	0.46	−0.02	0.43	0.59
Random Forest				
Topsoil	0.35	−0.01	1.08	0.56
Subsoil	0.21	−0.01	0.54	0.40

8 g/100g^{-1}), with relatively large concentration in the centre of the area (5 g/100g^{-1}). There seems the be more TC in areas where the NDVI have large values, but this pattern is not obvious. The subsoil maps of TC have a very different pattern than the topsoil map. There seems to be an uniform distribution of TC around 1 g/100g^{-1} for most of the area. High concentration of TC (> 5 g/100g^{-1}) are seen in a patch in the centre of the catchment and in a large area in the south. [c4]

[c4] In Figure 3, depths have been added.

5 Discussion

The proposed modelling approach explicitly accounts for the TC measurement error in the model calibration. The measurement error of NIR inferred TC was larger than that of the MIR inferred TC. This is an expected result reported in many previous studies (e.g., Rossel et al., 2006). The reason is that fundamental molecular vibration of bands associated to soil organic constituents occurs in the MIR region, while overtones and combinations appear in the NIR. Accounting for measurement error in spatial modelling of soil property using spectroscopically inferred soil data received recently much attention. Using the same case study, Somarathna et al. (2018) showed that acknowledging for measurement error almost halved prediction uncertainty. Similarly, Ramirez-Lopez et al. (2019) emphasized the importance of estimating and accounting for measurement error of spectroscopically inferred soil properties, as those can be larger than the sampling error. To the best of our knowledge, our study is [c1]one of the first to account for measurement error for mapping using machine learning. [c2]We note the contribution of Hengl et al. (2018) [c3]which use the measurement error to change to probability of a measurement to be selected in the bootstrap sample to calibrate a RF model.

[c1] *Text added.*
[c2] *Text added.*

The window size of the input images had a significant impact on model's accuracy measure, as tested on an independent test set. This is because the size of the input image is closely related to the amount of contextual information we supply to our model. CNN integrates spatial context by using the pixels of covariates surrounding a sampling location. In our regional scale case study of TC mapping, a window size of 21 × 21 and 29 × 29 provided the lowest prediction error, but larger window size worsened the prediction accuracy. In a similar context, this confirms the results found by Behrens et al. (2010b)[c4]. The authors showed that prediction accuracy of topsoil silt content increased remarkably by using larger neighbourhood size. However, our results also clearly indicated that including larger scale contextual information (larger input images window size) is not always better. This is similar to the results of Smith et al. (2006) who noted that the windows size greatly varies between landscapes and concluded that the appropriate size is case-dependent.

[c3] *Text added.*

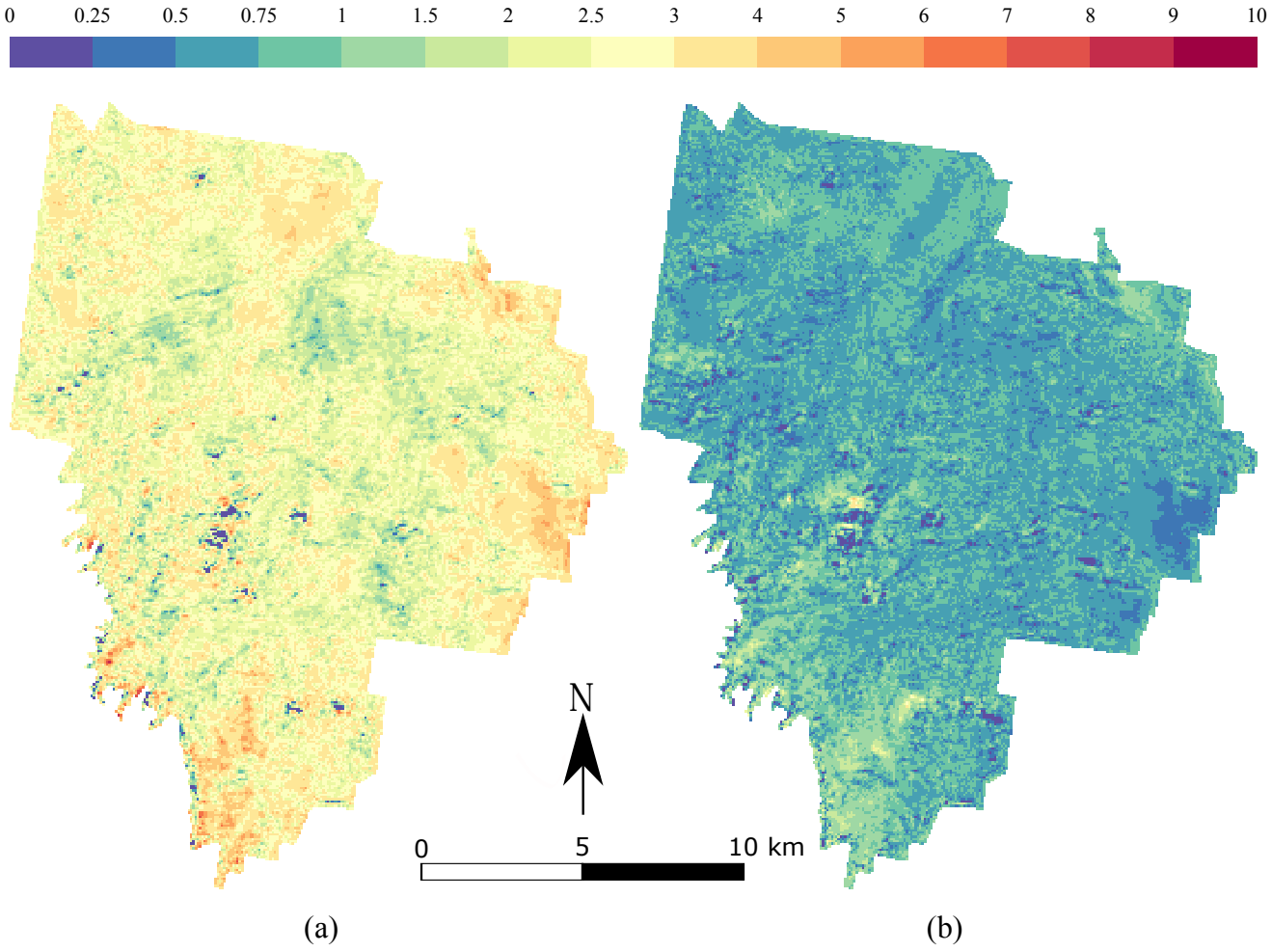

Figure 6. Maps of the prediction of organic carbon for (a) topsoil (0-10 cm) and (b) subsoil (40-50 cm). The values are expressed in g/100g^{-1}.

Using a window size of 21 × 21 to 29 × 29 is equivalent to including spatial information in a radius from the sampling location of about 262 to 362 m. Thus, it can be assumed that the window size relates to the range of spatial auto-correlation of TC. Several authors provided equivalent values of spatial correlation range. Kumhálová et al. (2011) reported values of organic matter spatial correlation range between 240 and 270 m using data of an experimental field in the Czech Republic. Similarly, Jian-Bing et al. (2006) found a spatial correlation range of 309 m for a small watershed in northeast China. For our case study, we verified this assumption by fitting a spherical variogram to the experimental variogram of TC. The fitted value of range was 329 m for the topsoil and 275 m for the subsoil. This is close to the actual radius of the window size that we found optimal. This is however delicate to draw conclusion. The actual correlation between auto-correlation range of a soil property and window size deserves further investigation so as to generate rules.

Our approach predict TC for both topsoil and subsoil using a single model. The predictions benefit from using a common architecture for the two soil depths. When compared to predicting each depth separately using random forest, our method reduced the mean squared error by 15 and 25% for topsoil and subsoil, respectively. Other studies reported similar results than those produced by the random forest model. For example Kempen et al. (2011) reported a R^2 of 0.23 for the 30-60 cm depth range. Brus et al. (2016) also noticed a significant increase of the error for predicting organic carbon at deeper soil layers. Our results confirm the recent study of Padarian et al. (2018) who showed a substantial decrease of the error associated to the prediction of deeper soil layer using CNN. The authors showed that CNN generates a representation of the vertical distribution of the soil profile, which reproduced closely the observed vertical distribution. Following Angelini et al. (2017) we also tested this by assessing the interrelation [c1][c2]between topsoil and subsoil for the measured, predicted by CNN or predicted using RF TC (Table. 4). CNN maintains much better the correlation between depths than RF, as shown by the value of the [c3][c4]Pearson's r correlation coefficient. This is an important finding which needs to be confirmed in further studies. Soil properties are often predicted depth by depth, which can result in predicting physically unrealistic soil profiles. In this study we showed that the deterministic behaviour of a depth function can be partly reproduced by CNN.

[c4] ref changed
[c1] ~~of~~ among
[c2] *Text added.*
[c3] ~~Person~~

[c1]We used three covariates to calibrate the CNN model. This might be a surprisingly small number compared to other studies on mapping with machine learning techniques (e.g. Nussbaum et al. (2018))[c2]. This is however a similar number compared to Padarian et al. (2018) [c3]who used four covariates for mapping organic carbon at large scale. Further research will test the effect of the number of covariates into the CNN model calibration. We foresee a large increase of the computational load when using more covariates. This is because in most studies on deep learning, only three colour channels are used. In this study we showed that only a small number of covariates was sufficient to provide satisfactory prediction accuracy. [c4]

[c4] *Text added.*
[c1] *Text added.*
[c2] *Text added.*
[c3] *Text added.*

Adapting CNN for soil mapping poses some practical problems. We mention two of them along with our solution for future research:

[c4] Upadated values in Table 4

- Input images containing missing values are disregarded by CNN during calibration and prediction. This means that (i) sampling locations close to the border of the area will be discarded from the analysis because their corresponding covariate images contain missing values and (ii) prediction will suffer from an edge effect, i.e. pixels at the edge of the area will not be predicted. This is a common problem when using a moving windows operation in GIS. In our study, we padded the rows and columns of the covariates with -1, so as to avoid providing missing values to the model. The CNN is capable of predicting TC while learning that values containing -1 are missing values so that the subsequent prediction do not acknowledge an edge effect. We note that padding a value of -1 is an arbitrary choice which might no be right in another case study.

- A CNN model takes more time to train and predict than a RF one. In our case study, it took 5 seconds to fit RF and about 30 seconds to predict at about 600,000 centre of grid cells, using a standard 4-cores laptop. CNN took 15 minutes to fit and 30 seconds to predict but requires preparing a 4-D matrix of size $n \times c \times w \times h$ for the training (n is the number of sampling locations) and for predicting (n is the number of prediction locations). This is tractable when the study

Table 4. Pearson's r correlation coefficient between the topsoil and subsoil TC for the original measurements of the test set, the predicted TC by CNN and the predicted TC by RF.

	Original	CNN	RF
Pearson's r	0.20	0.27	0.39

area is small or for predicting on a coarse grid, but becomes quickly computationally cumbersome for large scale or high-resolution mapping. This is new problem arising from using multiple covariates when image recognition problems commonly use three [c5c6]channels (colour channels R-B-G). A [c7c8]straightforward solution is to increase the number of cores available. Most available software implementation of deep learning models already support the use of parallel computing solutions.

Finally, we note that in spite of its predictive power, CNN has the major disadvantage of being a "black box" machine learning model, where results provide little knowledge, if any, into soil processes. In fact, many authors have noted that machine learning models are difficult to interpret. Recent publications (e.g., Angelini et al., 2017) have made a step forward "conscious" digital soil mapping where cause-effect relationships are adjusted with pedological knowledge. Solutions to interpret CNN or more common ANN models exist but they have been unexplored in digital soil mapping, for example automated sensitivity analysis (Tickle et al., 1998) which consists in keeping track of the error computed during back propagation to measure the degree to which each covariate contributes to the prediction error. The larger the contribution, the larger the influence of the covariate. Another solution is to extract set of rules (Andrews et al., 1995) for each hidden layer based on the [c1]connection weight vector and associated bias of each [c2c3]neuron. Taking these methods into account would certainly make a valuable extension to future CNN studies.

[c5] ~~layers~~
[c6] *Text added.*
[c7] ~~solution is to move to could computing or~~
[c8] *Text added.*
[c1] *Text added.*
[c2] ~~unit~~

6 Conclusions

We have shown how to train a deep learning model to predict total organic carbon at two soil depths using uncertain measurement of the soil property. The results and discussion bring us to the following conclusions:

- The uncertainty of the organic carbon values inferred by NIR spectroscopy was larger than those inferred by MIR. The uncertainty of the NIR inferred soil carbon measurement was large. Ignoring the latter uncertainty during model calibration results in a substantial part of the uncertainty being ignored, which can potentially lead to biased parameter estimates.
- A known measurement error can easily be accounted for when calibrating a CNN model, by weighting the objective function to optimize.
- CNN can be used for soil mapping using contextual covariates information. However the amount of contextual information we supply to the model, as represented by the window size of the input covariates, must be chosen with attention. In

our case study a radius of 262 to 360 m provided the [c4]best results. This is closely related to the range of the soil organic carbon spatial auto-correlation. Future studies may show whether this is a consistent finding or case dependent.

[c3] *Text added.*

- In our case study, CNN outperforms RF as assessed by several prediction accuracy measures.
- A single CNN model can be used to predict multiple outputs. In our case study, we predicted simultaneously at two soil depths. Deeper depth was much better predicted by CNN than RF. In addition, the reported predictions preserve the interrelation between depths. CNN is more suited for predicting correlated outputs. This also needs to be further investigated so as to generate rules.
- More research is needed to (i) identify solutions for fast CNN soil data (pre-)processing for large scale or high resolution soil mapping, (ii) develop methods to interpret CNN models and extract pedological knowledge from the neural network and (iii) derive uncertainty bounds of the predictions made by CNN.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgements. Alexandre Wadoux received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607000. We thank David Lopez-Paz, Facebook AI Research, for valuable comments.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning., in: OSDI, vol. 16, pp. 265–283, 2016.

Allaire, J. and Chollet, F.: keras: R Interface to 'Keras', https://CRAN.R-project.org/package=keras, R package version 2.2.0, 2018.

Andrews, R., Diederich, J., and Tickle, A. B.: Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-based systems, 8, 373–389, 1995.

Angelini, M. E., Heuvelink, G., and Kempen, B.: Multivariate mapping of soil with structural equation modelling, European Journal of Soil Science, 68, 575–591, 2017.

Behrens, T., Schmidt, K., Zhu, A.-X., and Scholten, T.: The ConMap approach for terrain-based digital soil mapping, European Journal of Soil Science, 61, 133–143, 2010a.

Behrens, T., Zhu, A.-X., Schmidt, K., and Scholten, T.: Multi-scale digital terrain analysis and feature selection for digital soil mapping, Geoderma, 155, 175–185, 2010b.

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.-X., and Scholten, T.: Hyper-scale digital soil mapping and soil formation analysis, Geoderma, 213, 578–588, 2014.

Behrens, T., Schmidt, K., MacMillan, R. A., and Viscarra Rossel, R. A.: Multi-scale digital soil mapping with deep learning, Scientific reports, 8, 15 244–15 244, 2018.

Brus, D. J., Yang, R.-M., and Zhang, G.-L.: Three-dimensional geostatistical modeling of soil organic carbon: A case study in the Qilian Mountains, China, Catena, 141, 46–55, 2016.

Demattê, J. A. M., Fongaro, C. T., Rizzo, R., and Safanelli, J. L.: Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images, Remote Sensing of Environment, 212, 161–175, 2018.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., et al.: The shuttle radar topography mission, Reviews of geophysics, 45, 2007.

Gallant, J. C. and Dowling, T. I.: A multiresolution index of valley bottom flatness for mapping depositional areas, Water resources research, 39, 4.1–4.13, 2003.

Geeves, G., Cresswell, H., Murphy, B., Gessler, P., Chartres, C., Little, I., and Bowman, G.: The physical, chemical and morphological properties of soils in the wheat-belt of southern New South wales and northern Victoria. CSIRO Aust. Division of Soils Occasional Report., 1995.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y.: Deep learning, vol. 1, MIT press Cambridge, 2016.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H.: Soil organic carbon concentrations and stocks on Barro Colorado Island-Digital soil mapping using Random Forests analysis, Geoderma, 146, 102–113, 2008.

Grinand, C., Arrouays, D., Laroche, B., and Martin, M. P.: Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context, Geoderma, 143, 180–190, 2008.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., and Gräler, B.: Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, PeerJ, 6, e5518, 2018.

Jian-Bing, W., Du-Ning, X., Xing-Yi, Z., Xiu-Zhen, L., and Xiao-Yu, L.: Spatial variability of soil organic carbon in relation to environmental factors of a typical small watershed in the black soil region, northeast China, Environmental monitoring and assessment, 121, 597–613, 2006.

Kempen, B., Brus, D., and Stoorvogel, J.: Three-dimensional mapping of soil organic matter content using soil type–specific depth functions, Geoderma, 162, 107–123, 2011.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Kumhálová, J., Kumhála, F., Kroulík, M., and Matějková, Š.: The impact of topography on soil properties and yield and the effects of weather conditions, Precision Agriculture, 12, 813–830, 2011.

Lark, R. and Webster, R.: Changes in variance and correlation of soil properties with scale and location: analysis using an adapted maximal overlap discrete wavelet transform, European journal of soil science, 52, 547–562, 2001.

Lawrence, I. and Lin, K.: A concordance correlation coefficient to evaluate reproducibility, Biometrics, pp. 255–268, 1989.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D.: Backpropagation applied to handwritten zip code recognition, Neural computation, 1, 541–551, 1989.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, nature, 521, 436, 2015.

McBratney, A. B., Minasny, B., Stockmann, U., et al.: Pedometrics, Springer, 2018.

Miller, B. A., Koszinski, S., Wehrhan, M., and Sommer, M.: Impact of multi-scale predictor selection for modeling soil properties, Geoderma, 239, 97–106, 2015.

Minasny, B., McBratney, A. B., and Salvador-Blanes, S.: Quantitative models for pedogenesis - A review, Geoderma, 144, 140–157, 2008.

Moore, I. D., Gessler, P., Nielsen, G., and Peterson, G.: Soil attribute prediction using terrain analysis, Soil Science Society of America Journal, 57, 443–452, 1993.

Moran, C. J. and Bui, E. N.: Spatial data mining for enhanced soil map modelling, International Journal of Geographical Information Science, 16, 533–549, 2002.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, Soil, 4, 1–22, 2018.

Padarian, J., Minasny, B., and McBratney, A. B.: Using deep learning for Digital Soil Mapping, SOIL Discussions, 2018, 1–17, 2018.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2018.

Ramirez-Lopez, L., Wadoux, A. M. J.-C., Franceschini, M., Terra, F., Marques, K., Sayao, V. M., and Demattê, J.: Robust soil mapping at farm scale with vis-NIR spectroscopy, European Journal of Soil Science, In Press, 2019.

Rossel, R. V., Walvoort, D., McBratney, A., Janik, L. J., and Skjemstad, J.: Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties, Geoderma, 131, 59–75, 2006.

Samuel-Rosa, A., Heuvelink, G., Vasques, G., and Anjos, L.: Do more detailed environmental covariates deliver more accurate soil maps?, Geoderma, 243, 214–227, 2015.
Smith, M. P., Zhu, A.-X., Burt, J. E., and Stiles, C.: The effects of DEM resolution and neighborhood size on digital soil survey, Geoderma, 137, 58–69, 2006.
Snoek, J., Larochelle, H., and Adams, R. P.: Practical bayesian optimization of machine learning algorithms, in: Advances in neural information processing systems, pp. 2951–2959, 2012.
Somarathna, P., Minasny, B., Malone, B. P., Stockmann, U., and McBratney, A. B.: Accounting for the measurement error of spectroscopically inferred soil carbon data for improved precision of spatial predictions, Science of the Total Environment, 631, 377–389, 2018.
Tickle, A. B., Andrews, R., Golea, M., and Diederich, J.: The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks, IEEE Transactions on Neural Networks, 9, 1057–1068, 1998.
Wadoux, A. M. J.-C., Brus, D. J., and Heuvelink, G. B.: Accounting for non-stationary variance in geostatistical mapping of soil properties, Geoderma, 324, 138–147, 2018.

We thank Dr. Amelie Beucher for the constructive comments on our paper. You can find below our answer to your suggestions.

Concerning the ANN section:

Each layer contains units: nodes for the input layer (as no computation occurs at their level) and neurons for the hidden and output layers;

We agree with this comment.

You should avoid confusion on the term "model parameters" right in this ANN section. Further in the case study part (3.2.2 Parameter estimation), you make the difference between model parameters and "the hyperparameters of the model architecture": these should be clearly defined before. I would first consider that the model parameters are the numbers of hidden layers, hidden neurons and iterations (i.e. parameters of the model architecture), not the connection weights and bias. You need to clarify this.

We call the connection weights and the bias as 'model parameters' as opposed to with the 'hyperparameters of the model architecture', i.e. number of layer, number of neurons per layer, etc... The authors would like to keep the names as it is because the model parameters have to be estimated by minimization of the objective function, while the hyperparameters of the model architecture can simply be 'tuned'. We will emphasize the differences between the two earlier in the manuscript to avoid confusion.

You use the abbreviation ReLU (for Rectified Linear Unit): you should clarify it and shortly explain it (giving the equation is not enough).

We agree, we will explain the ReLU function in the revised manuscript.

Equations 3 to 5 are not self-sufficient; you need to introduce them more. Particularly, you start by writing "For k = 1, : : :, L hidden layers,", but h0 is for the input layer and hL for the output layer. You have to modify Figure 1 accordingly as well.

Thank you for spotting this small mistake, we will modify Fig. 1 accordingly.

Concerning the CNN section:

You should refer more to your figure 1 and list shortly the different steps or layers of a CNN before describing them one by one.

We will modify the text to refer more to Fig. 1. This comment has also been suggested by other reviewers.

Define the word convolution straight ahead (i.e. element-wise product and sum between two matrices), this is a new concept for most in the DSM community.

We will include this sentence in the revised manuscript.

You use the Max-pooling operator and should define it simply (i.e. selecting the maximum value in the convoluted image using a window size). For now, we only read the term in table 1.

In Table 1, we mentioned max-pooling and described its meaning and operation in the text . In the revised manuscript we will add a sentence to explain the max-pooling operation.

You should clarify what you mean by "the number of channels (aka depth)": Which depth are you referring to?

We will modify the text accordingly to your comments.

You do not define the flatten operation as such, only writing that ": : :the last convolution returns an image of size 1 x 1 and with a number of channels. This is a vector that we can pass to a fully connected layer." Again, we only read the term in table 1.

The flatten operation will be better described in the revised manuscript.

You should state clearly that a fully-connected layer is an ANN. For now, it is only noted in figure 10s caption.

We define the ANN in the section 2.: 'We first describe the principle of Neural Networks (NN), basis of CNN'. We will refer more to Fig. 1 to make it clear.

You neither define the dropout operation. It should be in the methodology.

We will explain it in the methodology section.

Concerning the parameter estimation:

You define the dataset D as ": : : a 4-D matrix of size n x c w x h : : :", but you only defined n (i.e. the number of calibration points) much later in the paper.

The number of sampling location n is defined at the beginning of the paper, line 15.

You have to be clear and consistent when you use the term "weights" (in this section and the whole paper): the measurement error weights (defined in the 2.5 section) are different from the connection weights (as in the model parameters theta). Using only the term "weights" is confusing.

We will make clear the distinction between the two in the revised manuscript.

It is important that you describe shortly the Bayesian optimization here and not in the case study part (3.2.2 Parameter estimation);

While we may agree on this comment, we would like to keep the Bayesian optimization paragraph in the practical implementation section. This is simply a tuning procedure that we decided to perform but we might confuse the reader by adding it to the methodology section. Therefore, we prefer to keep the paragraph on Bayesian optimization in the implementation section.

You should also clarify the backpropagation concept in connection with the error gradient descent. These are crucial points for neural networks in general.
We will add a few sentences about it.

Comments on the manuscript:

Most comments are straightforward to address. Some small answers below:

We will add a north arrow to the figures as well as a scale.

Concerning the addition of a plot illustrating the correlation between auto-correlation range and window size, we rather do not include it in the article. This paragraph in the discussion is a speculation that requires further investigation (as it is already mentioned).

Thank you Tom for your thorough and constructive review on our paper. We answer to your comments below:

For a comparison to be fair, I would recommend using caret package or tuneRF package to get a better estimate of the random forest parameters, especially "mtry". "mtry" has shown to lead to very different results but can be relatively inexpensively optimized.

We agree that the mtry parameters of the Random Forest model can be optimized and this can lead to better prediction accuracy. We will do so in the revised manuscript. Please note that mtry has already been manually tuned by trying several values. We will use an search algorithm to find maybe more appropriate values.

Points shown in Figure 2 show significant spatial clustering. Again, to get a realistic estimate of mapping accuracy I would recommend using spatial CV (see https://geocompr.robinlovelace.net/spatial-cv.html#intro-cv and https://mlr.mlrorg.com/articles/tutorial/handling_of_spatial_data.html). If it might help you to run spatial CV, we have developed a function for spatially subsetting points per fixed block (http://gsif.r-forge.r-project.org/sample.grid.html).

The sampling locations on this area are not really clustered even though Fig. 2 seems to show spatial clustering. The samples collected are based on a random catena design. Repeating the calibration based on different split between calibration and validation would simply lead to more stable results, which can be useful in some cases, e.g., for a sensitivity analysis. In our case study we do not perform such analysis. In addition, computing time for calibrating a single model (mentioned in the discussion) does not allow repeating this procedure several dozens of time. This is a limitation that we recognize in the Discussion and that must be addressed by further research. The calibration and validation sets are used to validate both random forest and CNN so that no method is disadvantaged using this split. A random split of the calibration and validation sets are very common in the soil mapping literature.

Any CV to be stable should be repeated e.g. 5-10 times. Single split could by accident lead to over-optimistic/pessimistic results.

See answer to comment 2.

Other comments:

Otherwise, I would highly recommending rewriting methods section, primarily to add more detailed explanation to non-CNN experts. I have read the paper two times and I have to admit that it was still not totally clear to me whether you fit models for top and sub-soil separately or at once. Figure 1 could probably be split into 2 figures and then you can explain some concepts of CNN somewhat clearer. You use many abstract terms from ML and these probably need a gentle intro for soil scientists.

The Method section will be modified in view on your comments and other reviewers comments. Many useful suggestions for improvements were made and we will account for them while revising the manuscript. In particular, technical terms will be clarified to non-expert readers.

Your Paper is, in essence, twin-paper of Behrens et al. (2018) and leads to similar results/conclusions (Behrens use the term "Gaussian scale space" to characterize scaling), so please emphasize more what is really different in your approach as compared to their approach. For example, Behrens et al. (2018) run more complex terrain analysis for a range of aggregated DEMs, but they do not deal with NIR/MIR data specifically etc.

We will put emphasise on the differences with the paper of Behrens (2018). Basically Behrens et al.'s deep learning is a multi-layer feedforward artificial neural network, while our model is a

convolutional neural network. In addition, Behrens requires a pre-processing step: creation of multiscale data as inputs to the ANN, while CNN just accepts images as inputs and performs the multiscale analysis directly from the input.

Comments on the manuscript:

Most of them are useful small text editing comments. Some comments are similar to those made by other reviewers.

a) Line 9 page 9: We used a univariate RF model for each soil depth, i.e. a single model is trained by depth, while CNN predict both depths using a single calibrated model. We will make it clearer in the text.
b) Fig. 3: We would like to keep this figure. We believe it is interesting to visualize the covariates for interpretation of the predicted maps.
c) Line 15 page 12: for CNN we built one single model for both depths.
d) Line 5 page 14: Thank you for the suggestion. We will modify the text accordingly.
e) Line 22 page 16: Maps of the prediction error could be obtained, but this is not straightforward and difficult to implement while accounting for the measurement error. Deriving such maps would certainly need an important research effort.
f) Line 6 page 17: We agree that this is well known. The first paragraph in the discussion already mentioned this clearly and the relevant literature.

Dear Thorsten, thank you for the thorough revision of our manuscript. We address your comments below:

General remarks:

The paper is not easy to read (at least for a non-native speaker) and requires some restructuring and revision.

The text of the methodology section will be improved based on many comments that we received. We will clarify and explain terms that are not familiar to non-experts in deep learning.

The title starts with "Multi-source data integration". CNNs are explained in detail, but the integration of multi-source data is not adequately explained and discussed. The explanation of the CNNs lacks some detail.

We will try to expand and better describe the integration of multi data sources.

Why did you only use three covariables?

We used only three covariates because these were found to be the most important predictors in our case study at the fine spatial resolution. The same covariates are used in a recent paper of Somarathna et al (2018) https://doi.org/10.1016/j.scitotenv.2018.02.302

What is the difference between a window and a filter?

The difference between a window size and a filter will be better emphasized in the revised manuscript.

Specific comments:

Abstract:

The abstract is very condensed.
At the end you mention "different window size of input covariates matrix". This is not clear and needs an explaination.

We like the abstract but we will try to add a bit more detail on the results. We will rephrase the last sentences.

Introduction:

What was the purpose of DSM techniques before they were used to predict soil properties?

We will make small change in the text.

Can you explain this a little? The only point I really see is the size of the neighborhood.
Some approaches require less preprocessing than traditional terrain analysis. And you can simply set neighbourhood size as large as possible. For most approaches the resolution is the same for all input data. So I don't see any real "drawbacks" here. It's probably the same as setting up a CNN.

A bit more explanation will be provided.

In your paper you use three covariates. I assume that processing large spatial data sets with a CNN is a challenge. However, some of the other methods quoted here have no problem handling hundreds of covariables/scales. So not sure if they really have "drawbacks".

In view of your comment and those made by the other referee, we will and a short paragraph in the discussion to explain the limitation of our study using three covariates, and how CNN can handle more covariates.

What is the difference to approach presented by Padarian et al. (2018) especially when you write that "Padarian et al. (2018) have shown that it is possible to use CNN for soil mapping while accounting for contextual covariate information"?

The approach presented in our paper is an extension of the paper of Padarian (2018). Since using CNN is relatively new, we still need to develop approaches for mapping using CNN. In addition, Padarian (2018) use CNN for a country extent mapping, while we use a small, regional case study. We will mention it in the introduction.

Model definition:

Is this different with CNN?

Yes, in geostatistics, the measurements of the soil properties are a realization of an underlying Gaussian field. This is clearly not the case in deep learning, and generally in machine learning, where the measurements of the soil properties are spatially independent. We will add 'in space' to the sentence you mention.

Please explain the ReLU activation function and why you chose it.

ReLU activation function will be explained in the revised version of the manuscript.

CNN:

Small comments will be addressed directly in the revised manuscript. We agree with the reviewer comments and we will address them in the revised version. Most of them are also commented by the other referees.

There is no equation showing how this is done.

Regarding the case where we have c=3 covariates, we had this equation at an earlier version of the manuscript but we decided to not include it to not confuse the reader.

Figure 1: I suggest to split it in two separate figures. Does the convolution reduce the image size or just the pooling?

The authors would like to keep this figure as it is. We will make a better link with the text to explain each step.

Parameter estimation:

Is this "parameter estimation" or "learning (with backpropagation)"

Both are correct. We rather use the term 'parameter estimation', but learning could also be accepted.

Multi-source data integration:

The term "multi-source data integration" is in the title and is the interesting and important part of this article. However, this section does not show how it really works. How is the function updated? An example would be good.

We are not sure what the reviewer means by 'an example would be good'. We agree that we can add more explanation and we will do so in the revised version.

Quality of prediction:

I'm not sure if the formulas for RMSE are R2 are required.

We can indeed remove r2 and RMSE equations.

Why no cross-validation? Single subsets are generally not recommended for validation.

In our case study, it was not really possible to make a k-fold cross validation because of the computational load of calibrating a single CNN model. This is mentioned in the Discussion. Performing a cross-validation would provide more stable estimates of the validation statistics but would not change drastically the results. Splitting the dataset between validation and calibration sets is very common and widely accepted in soil mapping. The aim is just for a comparison with another method (RF), not for uncertainty estimates. This splitting will not benefit or disadvantage any method as the same dataset is used.

Why only 1000 trees? I suggest setting it to at least 2000 trees and then optimizing mtry and nodesize. However, this is not crucial as it will not significantly increase the prediction accuracy of RF in this study. The difference is the lack of spatial context.

The parameters of the RF model have already been tuned manually. Based on another reviewer comment, we will tune the parameter using a search algorithm for the revised version of the manuscript.

Why 90:10 in this case?

90:10 is an arbitrary choice, but this is common in the deep learning literature.

What "window size"? Filter size? In Table 1 I find 3x3.

The difference between window size and filter size will be better explained in the revised manuscript.

Why is the "window size" optimized separately? What were the inital settings of the hyperparameters when optimizing the window size?

The window size is optimized separately because it relates to the number of contextual information we provide to the model. This is not a model parameter not a hyperparameter of the model architecture.

Figure 6:

The maps look noisy and probably show artificial horizontal and vertical stripes that have nothing to do with the input data and look like artifacts from the convolution. But without examining the data, this is only a guess.

The map is indeed a bit noisy but we cannot see the vertical and horizontal stripes.

Discussion:

Can you produce uncertainty maps reflecting influence of the measurement error?

Producing uncertainty maps would certainly provide a valuable extension of our work. However, this is not straightforward. Padarian (2018) provides a maps of confidence intervals based on

predictions from several CNN models calibrated using a bootstrap sample of the training data. This refers to uncertainty due to the model, and does not include data variance. To obtain prediction intervals (and so the total variance of the prediction), we also need to account for the data noise variance (see second term of Eq. 3 in the paper of Khosravi et al., 2011). This is not simple and deserves an important research effort. In addition, this is not sure that measurement error can also be included because deriving prediction intervals requires another objective function instead of the MSE.

What is the effect on the final maps as well as prediction accuracy when setting all weights to 1?

We believe there is no reason why we should set the same weights as we already know that the measurements have different uncertainty. It is not a meaningful exercise, because we already know that the quality of data is different.

I suggest to put more emphasis on this.

We will put more emphasize on the novelty of our work.

Can you explain the relationship between the window size, the size of the input image, and the amount of contextual information? A figure would be good.

This is closely related to the Fig. 1. Based on your comments and other referee comments, we will put more link between the text and this Figure.

This is not what is shown in the ConMap paper quoted by you, but the multi-scale one in which we have tested comparable window sizes. Multi-scale digital terrain analysis and feature election for digital soil mapping Geoderma 155 (3-4), 175-185).

Thank you for spotting this mistake. We will add the correct reference in the revised version of the manuscript.

Reference:

Khosravi, A., Nahavandi, S., Creighton, D., & Atiya, A. F. (2011). Comprehensive review of neural network-based prediction intervals and new advances. IEEE Transactions on neural networks, 22 , 1341–1356.