SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

Multi-source data integration for soil mapping using deep learning [1]

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 covariate pixel values and measured soil properties 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 CNNs significantly increased prediction accuracy as indicated by the coefficient of determination and concordance correlation coefficient, when compared to a conventional DSM technique. [2] Deeper soil layer prediction error decreased, while preserving the interrelation between soil property and depths. The tests conducted using different window size of input covariates matrix to predict organic carbon 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.

1 Introduction

Digital Soil Mapping (DSM) techniques are now 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. [3] In most cases, the relation

1. **deep learning** maybe better use "Convolution Neural Network"? *[tom.hengl]*

2. **Results show that CNNs significantly...** I would expect here much more detail about the statistics, number of points, R-squares etc. *[tom.hengl]*

3. **time.** Maybe cite: McBratney, A.B., Minasny, B., Stockmann, U. (Eds) (2018) Pedometrics. Progress in Soil Science ISBN: 9783319634395, 720 pages. *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

between soil property and the chosen covariates is modelled by a regression which relates either linearly (Wadoux et al., 2018) or non-linearly (Grimm et al., 2008) sampled (point) soil properties and a vector a covariates value extracted at same point location.

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. (2010) 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 this approaches enable to contextualize the spatial information supplied to the regression, they rely either on heavy covariates pre-processing, subjective decisions based on the resolution to which covariates must be treated as input to the model or modeller's choice regarding neighbouring size. In light of these drawbacks, we propose to use Convolutional Neural Network (CNN) as a 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. 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. As for other regression methods, CNN is trained using measured soil properties at point location.

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.

1. **location. Several** keep together same paragraph *[tom.hengl]*

2. **contextual information** please define *[tom.hengl]*

3. **covariates.** ref missing *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

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 Networks for contextual spatial modelling, (ii) account 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.

2 Methodology

2.1 Model definition

We first describe the principle of Neural Networks (NN), 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 an Artificial Neural Network (ANN) model:

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

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 classical geostatistics, measurements of the soil property are assumed independent and identically distributed.

An ANN model is formed of several layers, or "computation steps". The input layer provides the raw information to the network, which is connected to at least one hidden layer, which in turn is connected to an output layer, which outputs the predictions of interest z_{s_i}. Each layer contains units, called neurons. The behaviour of the neurons depends on the activity of the previous layer neurons and the weights between the previous and new 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 gives output z_j by:

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

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 ReLU activation function $\phi(x) = \max(0, x)$ while the output layer for regression uses the linear activation function $\phi'(x) = x$.

1. **account for the soil property measurement...** I was expecting here "develop methodology for multi-source data integration" *[tom.hengl]*

2. **2.1 Model definition** Hmmm model definition stops, then there is a new section 2.2 (not part of the model?). Could be better organized *[tom.hengl]*

3. **ReLU** explain *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

Our neural network will contain more than one single hidden layer. For $k = 1, \ldots, L$ hidden layers,

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

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

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

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 $\boldsymbol{\theta} = (\mathbf{W}^1, b^1, \ldots, \mathbf{W}^L, b^L)$.

2.2 Convolutional Neural Network

In this paper we use the vicinity information of the measured soil property. In this case, ANN is not well adapted because they deal with 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). 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:

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

which can be rewritten with little modification to include the case where we have $c = 3$ environmental covariates. 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 $\mathbf{X}$.

Filters detect features 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:

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

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. Each convolution accepts one input image of a given size and number of channels (*aka* depth), 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. this is a vector that we can pass to a fully connected layer.

1. **h 0**
h is not explained
[tom.hengl]

2. **vicinity information of the measured...**
I have read the paper twice, and I am still not sure how you use the vicinity information. Scales / filters of different size do not explain vicinity effects to specific location but only effect of aggregation. This section could be much improved.
[tom.hengl]

3. **a convolution**
explain what does that mean
[tom.hengl]

4. **pooling layers.**
please explain in more detail
[tom.hengl]

5. **channels.**
note: this is becoming a heavy read "convolution", "channels", "pooling layers"... adding more explanation to soil scientists / environmentalist could help make this text more readable *[tom.hengl]*

6. **channels. this**
typo *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

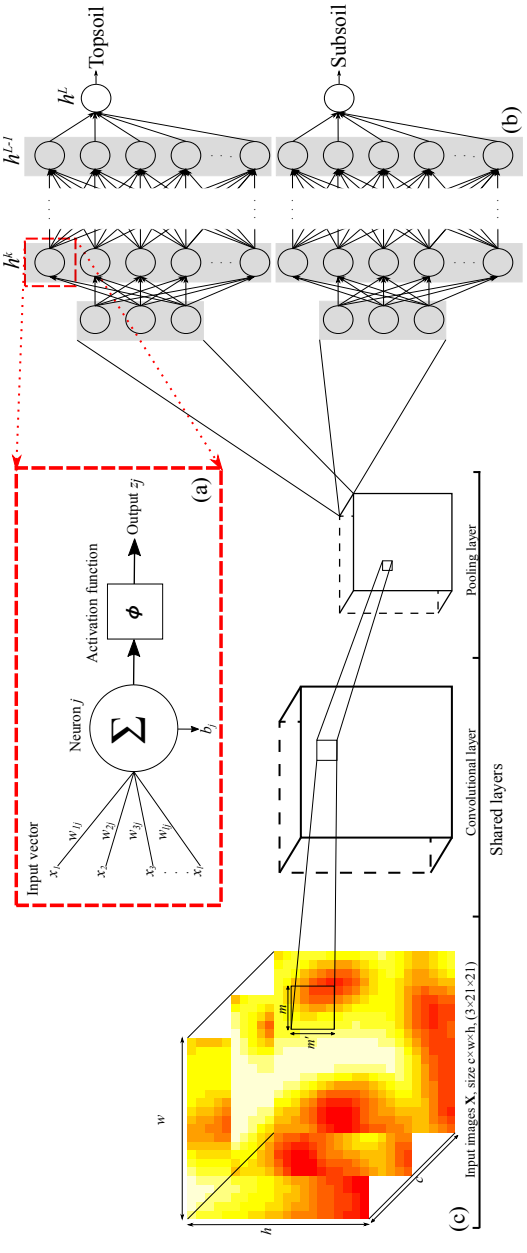

Figure 1. Representation of the CNN architecture developed in this study for (a) a neuron, (b) the ANN architecture or fully connected layer and (c) the convolutional and pooling layers. Note that $\sum = \langle \mathbf{w}_j, \mathbf{x} \rangle$.

1. **Figure 1. Representation**
I am having problems understanding this image. Why is the square smaller in each box? Are models for top-soil and subsoil combined in any way? Maybe split the figure into 2 figures? *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

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:

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

where $\boldsymbol{\delta}$ are the 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 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 weights oscillate.[1]

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 spectroscopic model. This uncertainty must be accounted for when calibrating the CNN model. A solution is to assign a weight to each value of the soil property, depending of its relative error compared to a true measurement of the soil property at the same location.[2] For a vector of a soil property inferred using a spectroscopic model $\mathbf{z}_{IR}$, one can assign a 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 weight δ for a measurement is given by:[3]

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

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. In this work, a true value of a soil property is measured in the laboratory, with assigned weight of 1. The weights for each observation are used in model calibration, by updating objective function to minimize in Eq. 8.

1. **A too small learning rate increases...**
Could this fine-tuning be automated? Maybe also mention in the discussion as general limitation?
[tom.hengl]

2. **location.**
ref missing - or is this is your original suggestion?
[tom.hengl]

3. **The weight $\boldsymbol{\delta}$ for a measurement...**
the important question is whether you derive this per each soil sample or you only use global values?
[tom.hengl]

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

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):

$$\mathrm{RMSE} = \frac{1}{N-n} \sum_{i=1}^{N-n} (z_{s_i} - \hat{z}_{s_i})^2, \tag{11}$$

the R^2:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N-n} (z_{s_i} - \hat{z}_{s_i})}{\sum_{i=1}^{N-n} (z_{s_i} - \overline{\mathbf{z}})}. \tag{12}$$

The bias is assessed by the Mean Error (ME):

$$\mathrm{ME} = \frac{\sum_{i=1}^{N-n} (z_{s_i} - \hat{z}_{s_i})}{N-n}, \tag{13}$$

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

$$\rho = \frac{2\rho' \sigma_{\mathbf{z}} \sigma_{\hat{\mathbf{z}}}}{\sigma_{\mathbf{z}}^2 + \sigma_{\hat{\mathbf{z}}}^2 + (\mu_{\mathbf{z}} - \mu_{\hat{\mathbf{z}}})^2}, \tag{14}$$

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).[2] 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

1. **Concordance Correlation Coefficient** this is practically most important measure of prediction accuracy *[tom.hengl]*

2. **topsoil (0-10 cm) and subsoil (40-50...** any reason why there is a gap between 10 and 40 cm? *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

(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, 2,962 measurements of TC are available for the first depth, among which 722 are from the CNS methods, 1,146 for the NIR and 1,094 from the MIR method. In the second depth, there are 2,500 measurements of the TC: 304 using the CNS method, 1,165 using the NIR method and 1,031 using the MIR method. They are shown in Fig. 2.

In addition to the TC measurements, three covariates[1] 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.

3.2 Practical implementation

3.2.1 Model definition

The dataset of TC measurements was randomly splitted between test (20%) and calibration (80%) sets.[2] 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 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 $n = 3,519$ for calibration and $(N-n) = 880$ for test, $c = 3$ and $w = h$ of different sizes. When a square of size $w \times h$ is created in the

1. **three covariates**
3 covariates is very modest (would Madlene be able to use your method with this data: https://www.soil-journal.net/4/1/2018/ at all?). Most of PSM expert would recommend that you use at least 10+ covariates (which can be downloaded from various public sources: https://envirometrix.github.io/PredictiveSoilMapping/soil-covs-chapter.html#soil-covariate-data-sources) please discuss *[tom.hengl]*

2. **was randomly splitted between test...**
are all further results based on a single split? repeating CV (with refitting) about 5 times would be highly recommended *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

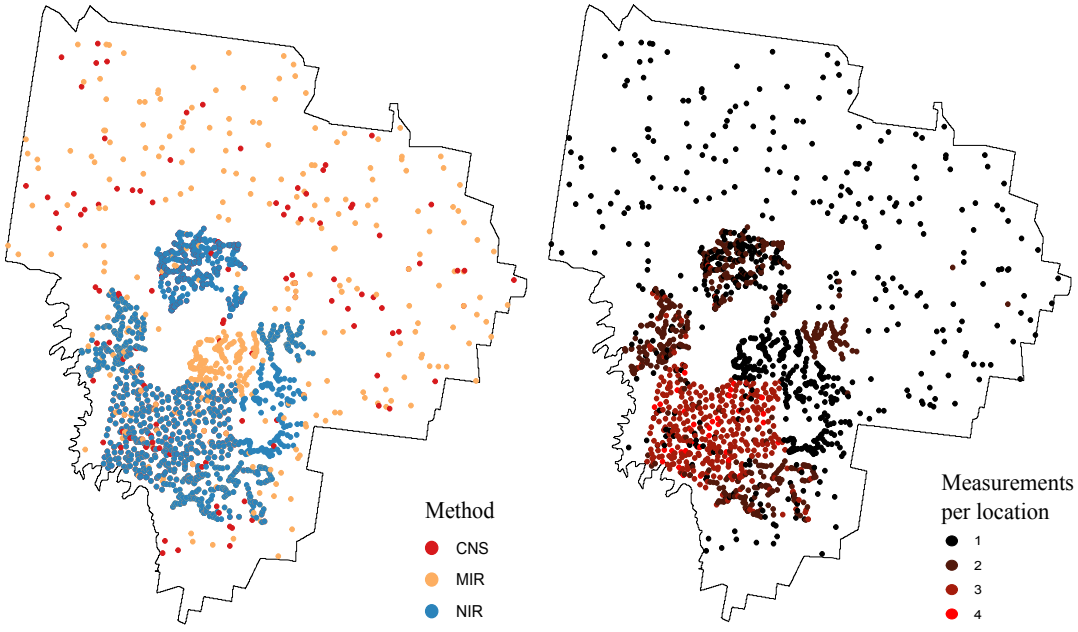

Figure 2. Spatial distribution of the observations for each measurement type (left) for the 0-10 cm topsoil and (right) number of measurements recorded per sampling location for the 0-10 cm topsoil. The subsoil (40-50 cm) map is not shown but closely resembles the topsoil in both number of sampling locations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

vicinity of a soil property at the border of the area, several missing values are reported. Since CNN can not 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.

In order to compare the CNN prediction to a reference method, we also calibrated an univariate Random forest (RF) model. 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

1. **Figure 2.** this shows serious spatial clustering, hence I recommend that you repeat CV using spatial CV (https://geocompr.robinlovelace.net/spatial-cv.html#intro-cv) *[tom.hengl]*

2. **univariate** what is here uni-variate? you mean separate models for top and subsoil? *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

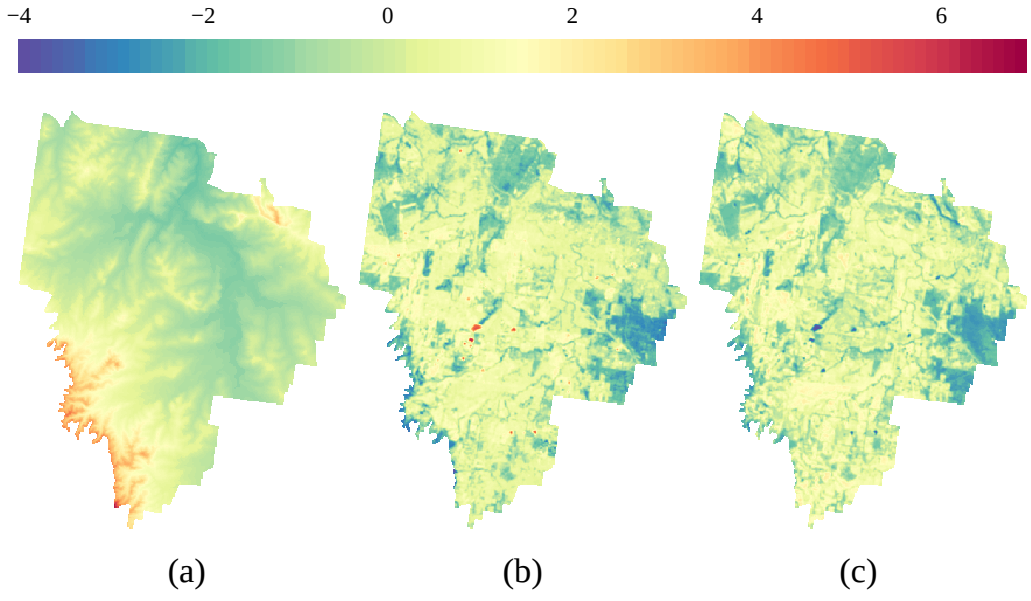

Figure 3. Standardized covariates used in model calibration (a) DEM, (b) Landsat 5 ETM band 5 and (c) NDVI.

1. **Figure 3.**
I would consider omitting this figure *[tom.hengl]*

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

2. **Layers**
quite abstract terms (I apologize for my ignorance) but adding more explanation in the caption could help readers *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

calibration/test sets, with normalized TC measurements and standardized covariates as input. A RF forest model is calibrated for each soil depth, using 1000 trees.[1]

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.[2] The calibration set (90%, 3,167 measurements) was used to calibrate the model and the validation set (10%, 352 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[3] for the input. The comparisons were made based on the depth averaged RMSE of the predictions 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 3,519 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.[4]

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.

4 Results

Based on the procedure detailed in Section 2.4, measurements of TC inferred from NIR spectra were assigned a weight of 0.43 and 0.52 for topsoil and subsoil, respectively.[5] The MIR inferred TC measurements had a 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 weight of 1, as explained in the previous section.

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

1. **A RF forest model is calibrated...**
If you really want a fair comparison, then I recommend fine-tuning RF. I would have not used RF without fine-tuning at least mtry (which can make drastic difference for spatial accuracy) *[tom.hengl]*

2. **We therefore randomly splitted...**
I am not entirely clear why do you need this second split. to estimate the weights / quality per soil sample? *[tom.hengl]*

3. **3, 5, 9, 15, 21, 29 and 35**
raster pixels? hence x 25 m? *[tom.hengl]*

4. **All processing was done in R 3.5.1...**
I would highly appreciated if you could share your code so that we can review the computational steps *[tom.hengl]*

5. **were assigned a weight of 0.43...**
so one weight for all samples of some type? I think pedometricians would prefer to know accuracy of estimate per sample as in e.g. https://www.mdpi.com/2072-4292/8/7/613/htm *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

Table 2. Weights given to the measurements.

	CNS	NIR	MIR
Topsoil	1	0.43	0.62
Subsoil	1	0.52	0.61

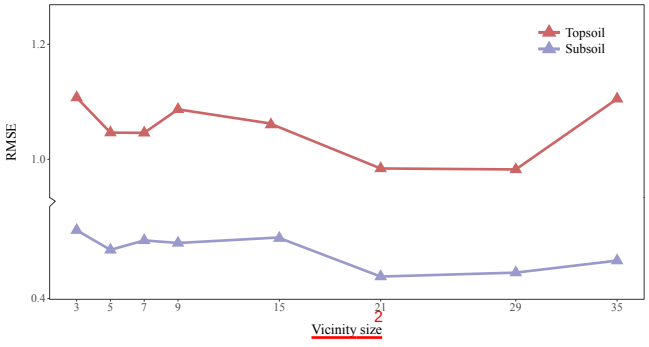

Figure 4. Effect of the vicinity size of input images. The RMSE corresponds to the error between the predictions and measured values in the test set.

resolution of 25 × 25 m so that a window of size 3 × 3 includes contextual information up to 3/2 × 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 × 5. The lowest averaged (topsoil and subsoil) RMSE is found for a window size of 21 × 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 × 29 and 35 × 35. From now on, all the results presented come from using an input window size of 21 × 21 pixels.

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.

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

1. **Table 2. Weights given to the measurements....** please omit and put in text *[tom.hengl]*

2. **Vicinity size** or "the filter size" *[tom.hengl]*

3. **CNN (R 2 = 0.55 for topsoil and...** so even with CNN you build separate models for top and subsoil? *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

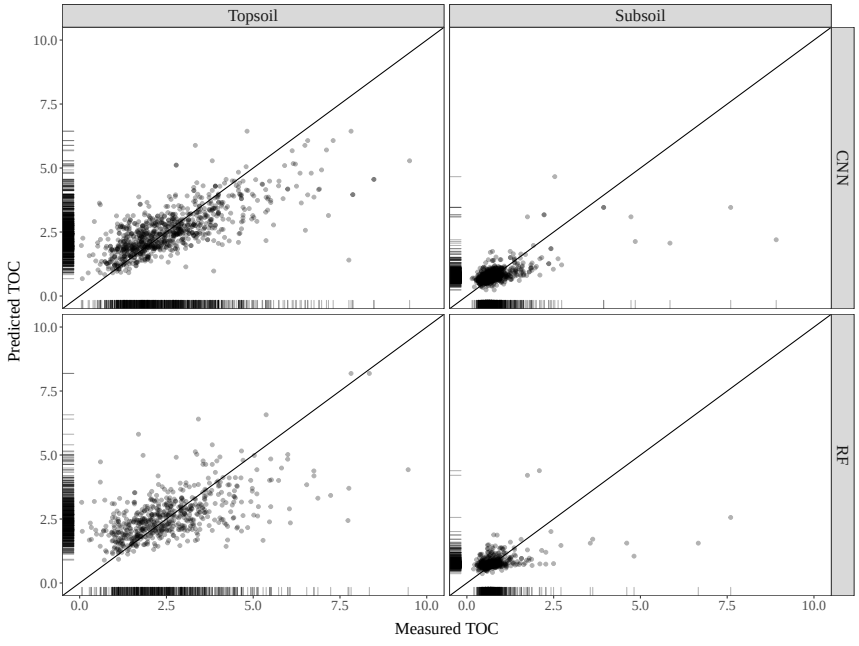

Figure 5. Scatterplot of the measured against predicted topsoil and subsoil TC for the CNN and RF models, along with the 1:1 line. Values are expressed in g/100g^{-1}.

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

	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	−0.05	1.07	0.55
Subsoil	0.21	−0.03	0.54	0.38

to zero in all cases). CNN model provides a significantly smaller accuracy measure (topsoil RMSE of 0.93 against 1.07 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 (>

1. **Figure 5.** I would display this variable in log-scale See e.g.: https://envirometrix.github.io/PredictiveSoilMapping/soilmapping-using-mla.html#prediction-3D But this is of course also question of personal preference *[tom.hengl]*

2. **Scatterplot of the measured against...** officially you should revert x and y, see: https://www.researchgate.net/publication/230692926_How_to_Evaluate_Models_Observed_vs_Predicted_or_Predicted_vs_Observed *[tom.hengl]*

3. **independent test set.** maybe add as a column global SD for topsoil and subsoil *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

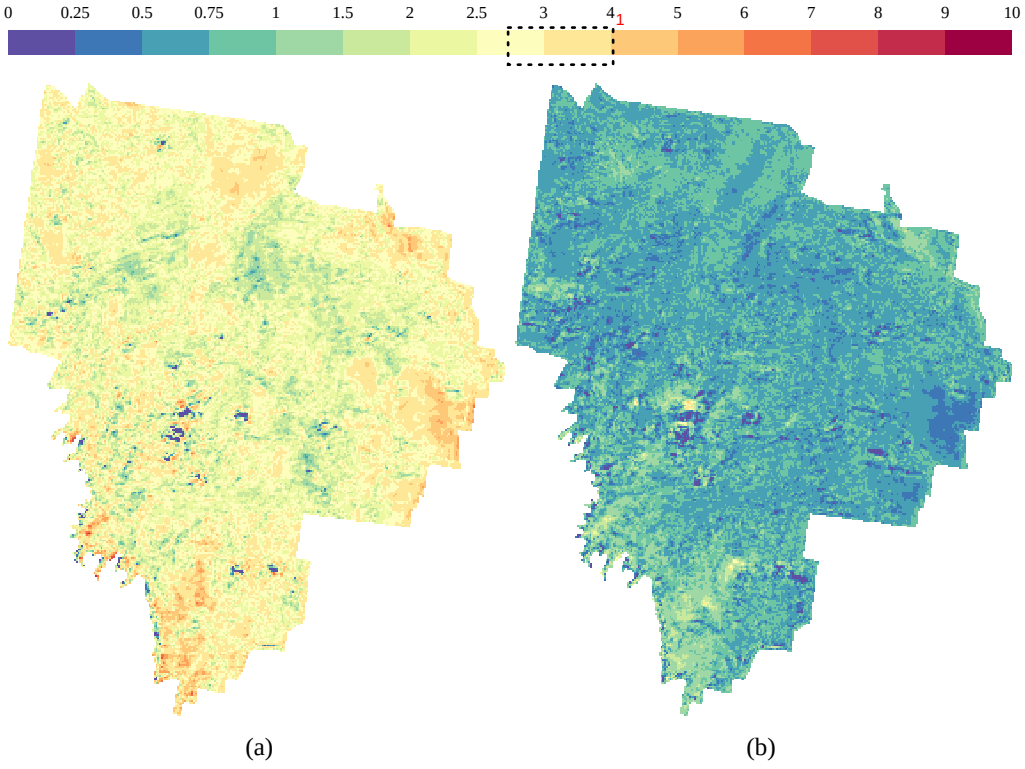

Figure 6. Maps of the prediction of organic carbon for (a) topsoil and (b) subsoil. The values are expressed in g/100g^{-1}.

1. somewhat difficult to read legend. if possible please replace with yellow to brown or similar *[tom.hengl]*

2. **topsoil and (b) subsoil.** 0-10 and 40-50 cm *[tom.hengl]*

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.

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

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

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 the first to account for measurement error for mapping using machine learning.[1]

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 $\times$ 21 and 29 $\times$ 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. (2010). The authors showed that prediction accuracy of topsoil silt content increased remarkably by using larger neighbourhood size.[2] 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.

Using a window size of 21 $\times$ 21 to 29 $\times$ 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.[3] 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 of among 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 Person's r

1. **To the best of our knowledge, 5...**
see section "NRCS data set (weighted regression, 3D)" in https://peerj.com/articles/5518/ I think the AfSIS project has produced something similar but I do not have a reference - please contact Markus Walsh to check *[tom.hengl]*

2. **The authors showed that prediction...**
if you run spatial CV you might get even more interesting results *[tom.hengl]*

3. **The fitted value of range was 329...**
this will always be study area specific, agricultural soils vs forest soils, etc. I can find you areas in Canada where SOC changes drastically every 1 m. *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

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	0.38

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.

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

- 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 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 layers (color channels R-B-G). A solution is to move to could computing or make 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.

1. **Table 4.** single row table - omit and add to text? *[tom.hengl]*

2. **Finally,** In your paper you also do not provide any estimate (map) of the prediction error. Could this be derived and how? *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

Another solution is to extract set of rules (Andrews et al., 1995) for each hidden layer based on the weight vector and associated bias of each unit. Taking these methods into account would certainly make a valuable extension to future CNN studies.

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 most 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.
- 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 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.

1. **by NIR spectroscopy was larger...**
that is well known - maybe cite some classic soil spec work here? *[tom.hengl]*

SOIL Discuss., https://doi.org/10.5194/soil-2018-39
Manuscript under review for journal SOIL
Discussion started: 16 November 2018

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, 2010.

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.[1]

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.

1. **2018.**
even more significant is Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T., & MacMillan, R. A. (2018). Spatial modelling with Euclidean distance fields and machine learning. European journal of soil science, 69(5), 757-770. *[tom.hengl]*