Geoderma 351 (2019) 59–70

Contents lists available at ScienceDirect

Geoderma

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

Using deep learning for multivariate mapping of soil with quantified uncertainty

Alexandre M.J.-C. Wadoux

Soil Geography and Landscape Group, Wageningen University, Droevendaalsesteeg 3, 6708 BP Wageningen, the Netherlands.

ARTICLE INFO

Handling Editor: Cristine L.S. Morgan

Keywords:
Pedometrics
Prediction intervals
Uncertainty quantification
Convolutional neural network
Machine learning
LUCAS

ABSTRACT

Digital soil mapping (DSM) techniques are widely employed to generate soil maps. Soil properties are typically predicted individually, while ignoring the interrelation between them. Models for predicting multiple properties exist, but they are computationally demanding and often fail to provide accurate description of the associated uncertainty. In this paper a convolutional neural network (CNN) model is described to predict several soil properties with quantified uncertainty. CNN has the advantage that it incorporates spatial contextual information of environmental covariates surrounding an observation. A single CNN model can be trained to predict multiple soil properties simultaneously. I further propose a two-step approach to estimate the uncertainty of the prediction for mapping using a neural network model. The methodology is tested mapping six soil properties on the French metropolitan territory using measurements from the LUCAS dataset and a large set of environmental covariates portraying the factors of soil formation. Results indicate that the multivariate CNN model produces accurate maps as shown by the coefficient of determination and concordance correlation coefficient, compared to a conventional machine learning technique. For this country extent mapping, the maps predicted by CNN have a detailed pattern with significant spatial variation. Evaluation of the uncertainty maps using the median of the standardized squared prediction error and accuracy plots suggests that the uncertainty was accurately quantified, albeit slightly underestimated. The tests conducted using different window size of input covariates to predict the soil properties indicate that CNN benefits from using local contextual information in a radius of 4.5 km. I conclude that CNN is an effective model to predict several soil properties and that the associated uncertainty can be accurately quantified with the proposed approach.

1. Introduction

Many agronomic and environmental activities require accurate information about the spatial distribution of soil properties. This information is commonly generated by digital soil mapping (DSM) techniques, whose framework has been formalized by the publication of McBratney et al. (2003). In DSM, prediction is made by exploiting the empirical quantitative relationship between a measured soil properties and one of several environmental covariates chosen to portray the factors of soil formation. The factors correspond to *s*: soil, *c*: climate, *o*: organism/vegetation, *r*: relief/topography, *p*: parent material, *a*: age and *n*: spatial position, which motivate the *scorpan* spatial model of soil variation. Usually, a model is built for predicting each property individually. This can lead to inconsistent prediction (Heuvelink et al., 2016) and rapid increase of computing intensity as the number of models (and therefore parameters to estimate) grows. For example, predicting soil clay, silt and sand would surely benefit from using a common model so as to avoid producing a unrealistic map of soil texture. This has been recognized in many previous DSM studies (e.g. Akpa et al., 2014).

Several methods exist for simultaneous prediction of soil properties, such as co-kriging (Goovaerts, 1997), regression co-kriging (Heuvelink et al., 2016) and structural equations modelling (SEM) (Angelini et al., 2017). These linear methods model the interrelation between properties explicitly but are computationally demanding when the size of the observation dataset is large or the number of properties to predict increases. In addition, they rely heavily on rigid statistical assumptions about the distribution of the soil properties. As an alternative, non-linear techniques such as machine learning have garnered wide interest during the past decade. (Hengl et al., 2018) have promoted random forest for multivariate prediction of soil properties, while (Xu et al., 2013) have adapted support vector machine for predicting many dependent variables simultaneously. Despite those examples, the use of machine learning techniques for multivariate soil mapping have been largely unexplored.

Recently, deep learning (DL) models have shown great potential for

E-mail address: alexandre.wadoux@wur.nl.

https://doi.org/10.1016/j.geoderma.2019.05.012
Received 6 November 2018; Received in revised form 10 May 2019; Accepted 11 May 2019
Available online 22 May 2019

soil mapping. Deep neural network as employed by (Behrens et al., 2018) produced more accurate predictions compared to random forest. Wadoux et al., 2019 and Padarian et al., 2019 used a convolution neural network (CNN) model for mapping soil organic carbon. The authors showed how a single CNN model can be used for predicting at multiple depths, and how the prediction accuracy significantly increased when compared to maps produced by either random forest or cubist regression trees. CNN has the advantage that it uses the contextual information contained in the vicinity of a location by taking a local representation of the input covariates. Mapping using the spatial domain of the covariates isnot new and several approaches have been developed, such as spatial filters using wavelet (Lark et al., 2004) or multiscale analysis (Miller et al., 2015). However, while these approaches contextualize the spatial information supplied to the model, they rely on either subjective modeller's decision or heavy covariates pre-processing, which hamper their use for predicting soil properties in an operational context.

Prediction is not the only interest of map users. Quantifying prediction uncertainty is as important as the prediction itself (Wadoux et al., 2018). Padarian et al., 2019 derived confidence intervals by training several CNN models on bootstrap samples of the input data. A confidence interval reflects the uncertainty around the mean prediction values. In soil mapping, we are rather interested in prediction intervals (Heuvelink, 2014), i.e. the range that is likely to contain the value yet to be observed. As consequence, a prediction interval is always wider than a confidence interval. Estimating only the latter certainly underestimates the total uncertainty of the prediction. For neural network models, several solutions to obtain prediction intervals have been proposed, such as the Delta, Bayesian or bootstrap plus variance estimate methods (Khosravi et al., 2011). To the best of my knowledge, quantifying prediction uncertainty of a neural network model has been yet disregarded in DSM studies.

The objectives of this study were to use a single CNN model for multivariate soil mapping and to quantify the uncertainty of the predictions. The methodology is tested in a potential application scenario, mapping topsoil clay, silt, sand, organic carbon, total nitrogen and pH in $CaCl_2$ solution over France. The predicted soil maps were validated and the uncertainty was estimated of each soil property.

2. Methodology

2.1. Convolutional neural network

Topsoil properties of interest $\mathbf{z}_{s_i}$ at location $s_i (i = 1,..., n; s_i \in \mathscr{F})$ for p soil properties $z_p (p = 1, \dots, P)$ in the study area $\mathscr{F}$ are modelled by a convolutional neural network (CNN):

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

where $\mathbf{X}$ is a 3-D input matrix of size $c \times w \times h$ which contains c environmental covariates of dimension $w \times h$ centred at the measured soil property location s_i. The vector $\boldsymbol{\theta}$ is a set of model parameters used by regression model f to map non-linearly $\mathbf{X} \rightarrow z$ and leaves room for a zero mean random error vector $\boldsymbol{\varepsilon}$. Measurements of the soil properties are assumed independent and identically distributed.

A CNN model is composed of several layers (Goodfellow et al., 2016), among which an input layer supplying the images $\mathbf{X}$ to the network. The input layer is connected to a hidden layer which in turn is connected to another hidden layer or to an output layer. Each layer contains neurons which are independent within the layer but connected to each neurons from the previous and to the next layer. Hidden layers can be classified in three main categories, called convolutional, pooling and fully connected layers.

A convolutional layer has the particularity that it performs a convolution between an input image and a filter. Convolutional layers are placed at the beginning of the network, and take an image of a given size and number of channels (i.e. number of covariates) as input and returns another image with a possibly different size but same number of channels. For a given input image $\mathbf{X}$ and a non-linear function (usually a rectified linear unit (ReLU), $\phi(x) = \max(0, x)$), the convolution outputs the image $\mathbf{X}'$ by:

$$\mathbf{X}' = p(\phi(\mathbf{W} * \mathbf{X} + \mathbf{b})), \tag{2}$$

where $\mathbf{W}$ is a matrix of weights of size $J^k \times J^{k-1}$, i.e. the number of neurons J in the current k layer times the number of neurons in layer k-1. The vector $\mathbf{b}$ are the neuron bias, "$*$" is a convolutional operator over dimensions w and h of $\mathbf{X}$ and $p(.)$ is a pooling function which selects the maximum value in the input image using a given window size (max-pooling). Note that convolutional and pooling layers share weights across the w and h dimensions which greatly reduce the number of parameters to be estimated. The last convolutional or pooling layer returns an image $\mathbf{X}$, which can be converted to a vector $\mathbf{x}$ (the flatten operation) and provided as input to a fully connected layer which outputs $\mathbf{x}'$ as follows:

$$\mathbf{x}' = \phi(\mathbf{W}\mathbf{x} + \mathbf{b}), \tag{3}$$

where ϕ is the ReLU activation function or the linear activation $\phi'(x) = x$ for the output layer. From Eq. (2) and Eq. (3) it follows that the model parameters $\boldsymbol{\theta} = (\mathbf{W}^1 b^1, \dots, \mathbf{W}^L \mathbf{b}^L)$ for $k = 1, \dots, L$ hidden layers. For notational convenience from here on I drop the subscript p so that $z_{p,\, s_i} = z_{s_i}$.

Parameters $\boldsymbol{\theta}$ can be estimated by training the CNN model to the dataset $\mathscr{D} = \{(\mathbf{X}_{s_i}, \mathbf{z}_{s_i}) \dots (\mathbf{X}_{s_n}, \mathbf{z}_{s_n})\}$ by minimizing the mean squared error (MSE) as objective function, defined by:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (z_{s_i} - \widehat{f}(\mathbf{X}_{s_i}; \widehat{\theta}))^2. \tag{4}$$

Note that one objective function is computed for each soil property p, but only one single multi-task CNN is trained. In this case, the objective function is simply the average of all soil properties objective function. The Adam optimizer (Kingma and Ba, 2015) is used to minimize Eq. (4). Adam computes the derivative of the objective function with respect to the model weights and bias to update their value in a process 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 batches. The number of epochs as well as the batch size is chosen by the user. Another hyperparameter is 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.

2.2. Uncertainty quantification

To estimate the variance of the prediction, I use a two-step method called bootstrap plus variance estimate (Khosravi et al., 2011). Recall from Eq. (1) that errors are assumed to be statistically independent and identically distributed. Using a predicted mean of the soil property of interest by regression model $\widehat{f}(\mathbf{X}, \widehat{\theta})$, shortly denoted $\widehat{f}$ hereafter, one can rewrite Eq. (1) as:

$$z_{s_i} - \widehat{f}_{s_i} = (f_{s_i} - \widehat{f}_{s_i}) + \varepsilon_{s_i}. \tag{5}$$

The first term in the right-hand side of Eq. (5) is the difference between the estimated model $\widehat{f}$ and true regression model f, which relates to the probability distribution $p(f \mid \widehat{f})$ (called confidence interval). The left-hand side of Eq. (5) is the difference between the measured values $\mathbf{z}_p$ and predicted values $\widehat{f}$, which relates to probability distribution $p(\mathbf{z}_p \mid \widehat{f})$ (called prediction interval). Therefore, the total variance of the prediction is formulated as (Khosravi et al., 2011):

$$\sigma_{s_i}^2 = \sigma_{\hat{f}_{s_i}}^2 + \sigma_{\hat{\varepsilon}_{s_i}}^2, \tag{6}$$

where $\sigma_{\hat{f}}^2$ is the error term due to model error, i.e. model miss-specification and model parameter uncertainty and $\sigma_{\hat{\varepsilon}}^2$ is the data error, i.e. the data noise variance. Estimation of these two terms separately is presented in more detail in the two next paragraphs.

1. The model error variance term is estimated by the bootstrap method (Efron and Tibshirani, 1994), which builds an ensemble of CNN models, each with a random initialization of the parameters $\boldsymbol{\theta}$. From the original dataset, B training sets are randomly sampled with replacement, forming $\mathscr{D}_{b=1}^{B}$ training datasets. Next, B CNN models are trained on bootstrapped data $\mathscr{D}_b$. The mean of B model outputs for a soil property p is computed by:

$$\hat{\mathbf{z}}_{s_i} = \frac{1}{B}\sum_{b=1}^{B} \hat{z}_{b,s_i}, \tag{7}$$

where $\hat{z}_b$ is the prediction of the bth bootstrap model for the soil property p. The model error variance term can be estimated using the prediction from b bootstrap models by:

$$\sigma_{\hat{f}_{s_i}}^2 = \frac{1}{B-1}\sum_{b=1}^{B} (\hat{z}_{b,s_i} - \hat{\bar{\mathbf{z}}}_{s_i})^2. \tag{8}$$

Note that models are trained based on the minimization of objective function defined in Eq. (4).

2. The data error variance term is estimated following Nix & Weigend (1994) by assuming normally distributed errors around $\hat{f}$. In this case, the CNN model outputs two values in the final layer, corresponding to the predicted mean $\hat{z}$ and variance ${\sigma_e}^2$ of a Gaussian distribution. The least square regression can be interpreted as maximum likelihood, by minimizing the negative log-likelihood criterion instead of Eq. (4), given by (Lakshminarayanan et al., 2017):

$$-\log(z_{s_i} \mid \mathbf{X}_{s_i}) = \frac{log\ \sigma_{\varepsilon_{s_i}}^2}{2} + \frac{(z_{s_i} - \hat{z}_{s_i})^2}{2\sigma_{\varepsilon_{s_i}}^2} + \text{constant}. \tag{9}$$

The variance term in Eq. (9) is passed through the function log $(1 + \exp(.)$ to enforce the positivity constrain. A small variance term (10^{-6}) is also added to ensure numerical stability (Lakshminarayanan et al., 2017). The optimizer used is the same as for minimizing Eq. (4). Once the CNN model trained, it is used to predict the variance term ${\sigma_e}^2$ at any location.

Assuming both terms independent and Gaussian enables to compute the total variance as the sum of the two terms. The method has the main disadvantage that it requires to build two separate models (one for estimating ${\sigma_e}^2$ and another for $\sigma_{\hat{f}}^2$) and to train $B + 1$ CNN models. Khosravi et al. (2011) note that this method relies on the B bootstrap model estimates. Some of the bootstrap samples may lead to a biased prediction. In consequence, the variance can be underestimated resulting in narrow coverage probability.

2.3. Quality of prediction and estimated uncertainty

For the vector $\mathbf{z}_p$ at N-n test locations where n is the number of calibration location and N is the total number of sampling locations, the quality of the prediction is quantified by the mean prediction error (ME), root mean squared error (RMSE), amount of variance explained by the model (r^2) and concordance correlation coefficient (CCC). The latter is derived as follows (Lawrence and Lin, 1989):

$$\text{CCC} = \frac{2\rho' \sigma_{z_p} \sigma_{\hat{z}_p}}{\sigma_{z_p}^2 + \sigma_{\hat{z}_p}^2 + (\mu_{z_p} - \mu_{\hat{z}_p})^2}, \tag{10}$$

where μ_p and ${\sigma_p}^2$ are mean and variance for the vector of true measurements $\mathbf{z}_p$ of soil property p or for the vector of predicted values $\hat{\mathbf{z}}_p$. The value ρ' represents the correlation between μ_{z_p} and $\mu_{\hat{z}_p}$. The CCC quantifies the agreement of the predictions to the 1:1 line. Its optimal value is 1 and it can be negative.

The quality of the prediction error variance is quantified by the standardized squared prediction error $\boldsymbol{\delta}_p$ (Lark, 2000):

$$\delta_{p,s_i} = \frac{(z_{s_i} - \hat{z}_{s_i})^2}{\sigma_{\hat{\varepsilon}_{s_i}}^2}, \tag{11}$$

which should be distributed as χ^2 with 1 degree of freedom so that its mean $\bar{\delta}$ should be close to 1. Lark (2000) showed that its median $\tilde{\delta}$ is more useful because it is less sensitive to small or large values, $\tilde{\delta}$ should be close to 0.455.

In addition, the prediction intervals coverage probability (PICP) is computed. The PICP is the percentage of observations covered by a defined prediction interval (Shrestha and Solomatine, 2008), in this case corresponding to a 90% probability of occurrence. The PICP is calculated as follows:

$$\text{PICP} = \frac{1}{N-n}\sum_{i=n+1}^{N-n} P_i \cdot 100, \tag{12}$$

for

$$P_i = \begin{cases} 1 & \text{if } q_{(0.05)} \le z_{s_i} \le q_{(0.95)}, \\ 0 & \text{otherwise}, \end{cases} \tag{13}$$

where $q_{(0.05)}$ and $q_{(0.95)}$ are the lower and upper boundaries of 90% probability of occurrence at location s_i.

Finally, a visual assessment of the quality of the estimated uncertainty is provided by an accuracy plot as first proposed by Deutsch (1997). Having normally distributed error at each test location with known mean and variance allows computing a symmetric interval around the predicted values by calculation of the $(1 - q)/2$ and $(1 + q)/2$ quantiles, for a number of q intervals. One can count the proportion of observed values at test location included for each q interval. If the uncertainty is correctly modelled, the proportion of observations covered by a q interval at test locations is approximately equal to the value of q, for all q. The values of q can be plotted in a scattergram against the actual proportion of observation in each q. Ideally all points in the plots are on the 1:1 line. Deviation from the 1:1 line is due to overestimation or underestimation of the modelled uncertainty, depending whether the points lie above or below the line, respectively.

3. Case study

3.1. Study area and data

The methodology is tested on the metropolitan territory of France which is about 543,965 km^2 excluding Corsica and other islands. France has a very diverse landscape and climate. The altitude ranges from 0 to more than 4500 m in the Alps. The climate is Mediterranean in the South and temperate in the North, influenced by a West-East gradient of decreasing precipitation and temperature due to the increasing distance from the Atlantic ocean. France is mostly covered by soils from calcareous rocks such as the Rendzic, calcaric Leptosol and Calcisols on the Champagne region or in Argonne. Soils from clayey sediments such a Haplic and Vertisols are found in patches north of Massif Central and inthe southern Alps mountains with fertile loess soils (Luvisols) in the North. Sandy soils such as Podzols are found in the Landes or Sologne while large areas in the Massif Central and Brittany are covered by dystric Cambisols originating from moderate weathering of different types of parent materials (Jones et al., 2005).

In this study I used the soil measurements from the land use and cover area frame statistical survey (LUCAS) covering the French

Table 1
List of *scorpan* environmental covariates used with unit and associated reference when applicable.

Factor of soil formation	Predictor variable	Unit	Reference
Soil	Average soil and sedimentary-deposit thickness	metre	Pelletier et al. (2016)
	Landsat Band 3 (red) for year 2014	–	Tucker et al. (2004)
	Global Water Table Depth	metre	Fan et al. (2013)
	Landsat Band 4 (NIR) for year 2014	–	Tucker et al. (2004)
	Landsat Band 5 (SWIR) for year 2014	–	Tucker et al. (2004)
	Landsat Band 7 (SWIR) for year 2014	–	Tucker et al. (2004)
Climate	Long-term averaged mean annual surface temperature (daytime) MODIS	Kelvin	U.S. Land Process archive centre
	Temperature seasonality	Celsius	Karger et al. (2017)
	Precipitation of driest month	mm	Karger et al. (2017)
	Total annual precipitation	mm	Karger et al. (2017)
	Temperature annual range	Celsius	Karger et al. (2017)
Organisms/vegetation	Global tree cover	%	Hansen et al. (2013)
	Cultivated land cover for year 2010	%	Chen et al. (2015)
	Forests cover for year 2010	%	Chen et al. (2015)
	Grasslands cover for year 2010	%	Chen et al. (2015)
	Shrublands cover for year 2010	%	Chen et al. (2015)
	Wetland cover for year 2010	%	Chen et al. (2015)
Relief	DEM	metre	Robinson et al. (2014)
	Terrain slope	radians × 100	–
	Multiresolution Index of Valley Bottom Flatness (MRVBF)	metre × 100	–
	SAGA Wetness Index	metre × 10	Olaya and Conrad (2009)
	Landform class: Breaks/foothills	%	U.S. Geological Survey
	Landform class: Flat plains	%	U.S. Geological Survey
	Landform class: High Mountains/Deep Canyons	%	U.S. Geological Survey
	Landform class: Hills	%	U.S. Geological Survey
	Landform class: Low hills	%	U.S. Geological Survey
	Landform class: Low mountains	%	U.S. Geological Survey
	Landform class: Smooth plains	%	U.S. Geological Survey
Parent material/age	Rock type: Acid plutonics	%	Hartmann and Moosdorf (2012)
	Rock type: Carbonate sedimentary	%	Hartmann and Moosdorf (2012)
	Rock type: Metamorphics	%	Hartmann and Moosdorf (2012)
	Rock type: Siliciclastic sedimentary	%	Hartmann and Moosdorf (2012)
	Rock type: Mixed sedimentary	%	Hartmann and Moosdorf (2012)
	Rock type: Basic volcanics	%	Hartmann and Moosdorf (2012))
	Rock type: Unconsolidated sediment	%	Hartmann and Moosdorf (2012)
Geographical position	X-coordinates	metre	–
	Y-coordinates	metre	–

territory. The LUCAS dataset is a harmonized dataset of about 20,000 topsoil (0–10 cm) samples covering the whole Europe. The LUCAS dataset has been used in many previous studies on soil spatial distribution (e.g. Ballabio et al. (2016)). In this study are predicted six soil properties from the LUCAS dataset covering France. The properties are the soil organic carbon in g kg^{-1}, particle size fraction (clay, silt and sand in %), pH in $CaCI_2$ solution and total nitrogen content in g kg^{-1}, denoted OC, clay, silt, sand, pH and N hereafter. For more information about the LUCAS dataset I refer to Tóth et al. (2013). In addition, a set of thirty-seven readily available environmental covariates were assembled to represent the *scorpan* factors of soil formation. Table 1 lists the covariates with their unit and original reference. Any covariate which did not conform with the target grid resolution of 1 km × 1 km was either resampled using bilinear interpolation or aggregated. Categorical covariates such as landforms and geology were transformed to dummy variables to allow subsequent analysis such as standardization.

3.2. Practical implementation

The original dataset was randomly split between calibration (80%) and test (20%) sets (Fig. 1). All six soil properties (clay, silt, sand, OC, pH and N) were simultaneously selected for either test of calibration. Each soil property was normalized between 0 and 1 while the covariates were standardized by subtracting their mean and dividing by their standard deviation. Recall that categorical covariates are converted to dummy variables prior the standardization. Two 4-D matrices are created, one for calibration and one for test, by extracting covariates values in the vicinity of the sampling location. They are of size $n \times c \times w \times h$ for the calibration set where n is the number of sampling location, $c = 37$ is the number of covariates and $w = h$ are the dimensions (in number of pixels) of the covariates input images. For the test set, the 4-D matrix has size $N - n \times c \times w \times h$.

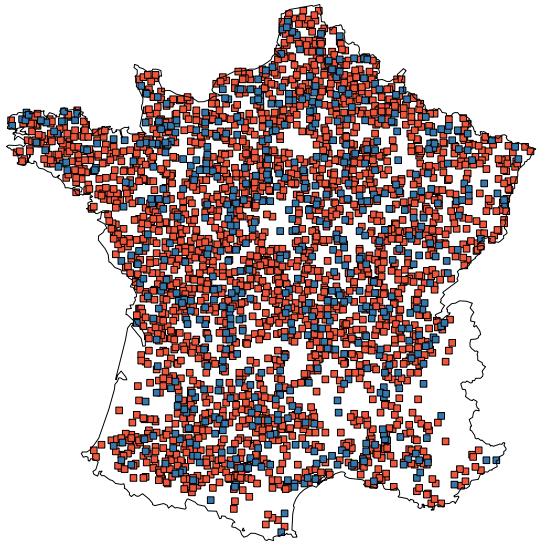

Fig. 1. LUCAS sampling locations in France. The squares in blue are the test locations and in red the calibration locations. Note that a square is used to illustrate the amount of covariate contextual information included into the model.

Table 2
Layers used in the sequential model built for clay, silt, sand, OC, pH and N prediction. A graphical representation is given in Fig. 2.

Layer type	Shared	Filter size	Number of filters/ neurons	Activation
Convolutional	Yes	3 × 3	64	ReLU
Max-pooling	Yes	2 × 2	–	–
Convolutional	Yes	2 × 2	40	ReLU
Dropout (0.1)	Yes	–	–	–
Convolutional	Yes	2 × 2	22	ReLU
Convolutional	Yes	2 × 2	30	ReLU
Dropout (0.1)	Yes	–	–	–
Flatten	Yes	–	–	–
Fully-connected	No	–	160	ReLU
Dropout (0.3)	No	–	–	–
Fully-connected	No	–	20	ReLU
Dropout (0.2)	No	–	–	–
Fully-connected	No	–	1 or 2	Linear
Concatenate & Softmax	clay, silt, sand	–	–	–

Next, a sequential multi-task CNN model is built for predicting jointly the six soil properties. The CNN model has a common architecture for the convolutional and pooling layers followed by separated fully connected layers which output either one (model for bootstrap) or two (model for variance estimate) values per soil property. The model specifications are reported in Table 2 and a graphical representation is given in Fig. 2. Zero padding is always applied to the convolutional layers to preserve the original size of the input image and conserve information at an early stage of the network.

Soil fraction clay, silt and sand are reported in percent, which must sum to 100. To handle this compositional constrain, the prediction from the output layer of clay, silt and sand is passed through a softmax layer. A softmax layer takes as input a vector and returns a vector of the same length, where each value is in the range $(0, 1)$ and the vector adds up to one. The output values of clay, silt and sand are then multiplied by 100. The softmax function is very similar to the additive log-ratio transform generally applied to compositional variables in soil mapping studies.

The CNN model is trained using different window size of image input ($h = w$) of 3, 6, 9, 15, 21 and 29 pixels. The optimal window size is chosen based on the averaged soil properties RMSE of the test set. The parameters were estimated by minimizing either Eq. (4) or Eq. (9) using the Adam optimizer. It is of common practice to further separate the calibration set into calibration (90%) and validation set (10%) before training the model. The calibration set is used to find optimal values of $\boldsymbol{\theta}$ while the validation set is used to ensure that the model is not overfitting during the parameter optimization process. Since the test set is used only for computing the accuracy measures it cannot be used for this purpose. Overfitting is prevented by adding dropout layers which deactivate neurons of a given layer at random during each batch. A total of 100 bootstraps are made so that $B = 100$. Predictions are made on the centre cells of a 1 km × 1 km grid.

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. Training a single model for 500 epochs, a batch size of 350, an input window size of 9 × 9 for 37 covariates and 2357 sampling locations took approximately 2 hours in parallel on a Linux server 4.4.0–38-generic Ubuntu SMP with 48 cores.

Random forest (RF) is used as a reference model to compare the predictions made by CNN. Random forest has been widely used in DSM studies to model non-linear relationships between soil properties and environmental covariates (Hengl et al., 2018). For a fair comparison between CNN and RF, the same calibration and test sets are used for both models. RF forest is trained for each soil property separately using 1000 trees and fine-tuned parameter values.

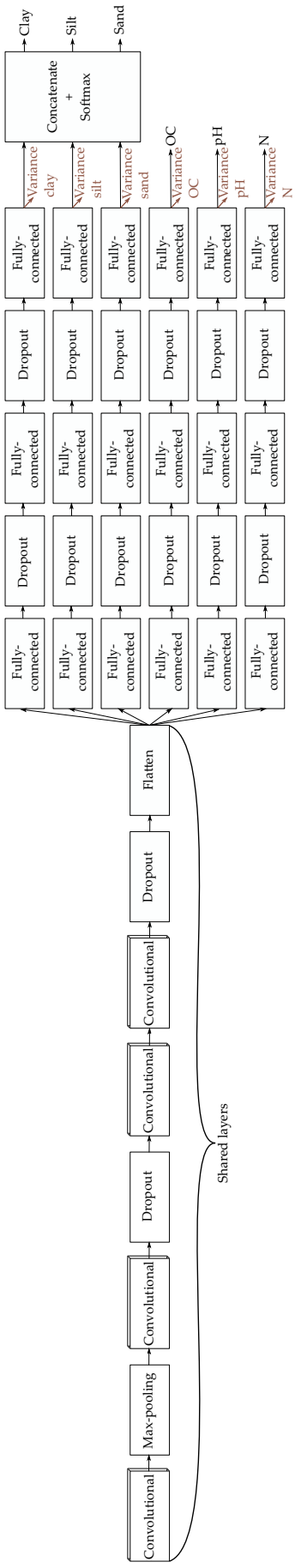

Fig. 2. Graphical representation of the CNN models built in this study. Network represented in black is used for the bootstrap while network combining black and red colours is used for the variance estimate.

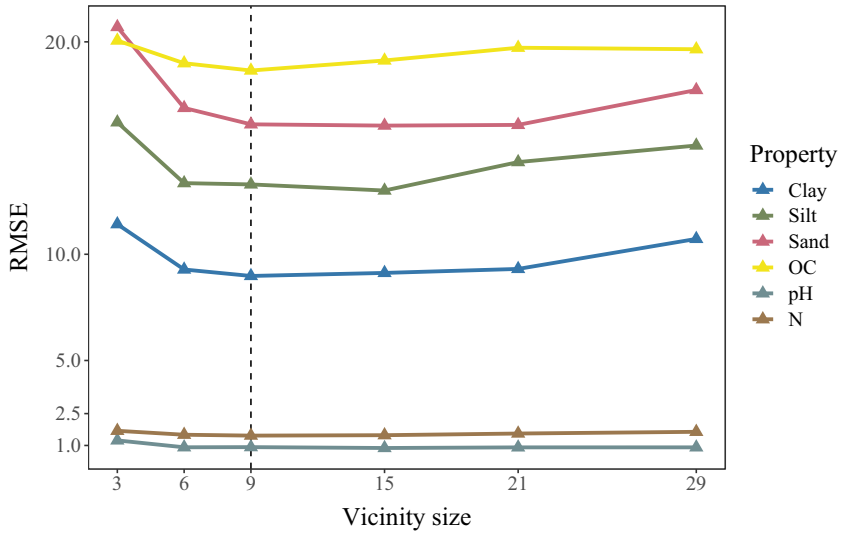

Fig. 3. Effect of the vicinity size of the input image. The RMSE corresponds to the error between the predictions and measured values in the test set.

4. Results

Fig. 3 shows the RMSE of each soil property for different window size of input images. On average, the RMSE decreases for increasing window size from 3 × 3 to 9 × 9 pixels. Recall that one pixel has a resolution of 1 km so that a window size of 9 × 9 pixels includes contextual information surrounding the soil property measurement up to (9 × 1)/2 = 4.5 km. Fig. 3 shows that using a window size larger than 9 × 9 pixels does not lead to further decrease of the soil properties RMSE, with the exception of pH for which the lowest RMSE is found using a window size of 15 × 15 pixels. CNN does not allow for different input window size while predicting using a multi-task model, therefore all results presented hereafter come from using an input window size of 9×9 pixels. The scatterplots of measured against predicted by CNN soil properties are shown in Fig. 4. For all soil properties, the predictions generally follow the 1:1 line. The agreement between measured and predicted values of clay and silt is satisfactory because most values are along the 1:1 line. For sand the predicted values are more scattered for large values of measured sand. The bulk of OC and N predictions are gathered on the 1:1 line, with the exception of large measured OC and N values which are strongly underestimated (e.g. measured OC = 200 and predicted OC = 50 g kg^{-1}). Scatterplot of pH shows a different pattern. Prediction of small pH values is more dispersed than prediction of larger ones. The measured pH values do not exceed 8 while the predictions are often greater than 8 and sometimes close to 9.

CNN predictions are compared to those made by RF (Table 3) as a baseline. CNN accuracy measures are on average equivalent or better than those of RF. Clay, silt, sand and pH are better predicted by RF than CNN, as shown by the r^2 and RMSE. CNN does better for predicting OC and N with a significantly larger r^2 for OC (r^2 for CNN is 0.15 and r^2 for RF is 0.12). This is confirmed by the CCC values which are greater for OC and N predicted by CNN. While clay, silt, sand and pH predicted by RF have a larger correlation coefficient with the measured values than those predicted by CNN, the CCC shows that CNN predictions are either better or equal to those of RF. Since the CCC assesses the deviation of the predictions with respect to the 1:1 line it is a useful measure to serve model comparison (Lawrence and Lin, 1989). The ME values show that prediction are relatively unbiased, with the exception OC predictions which are either slightly positively of moderately negatively biased for RF and CNN, respectively. This bias is certainly due to the measured OC values greater than 60 g kg^{-1} that are systematically underestimated. This is visible in Fig. 4.

The prediction and prediction standard deviation maps for clay, silt and sand are presented in Fig. 5 and for OC, pH and N in Fig. 6. For all soil properties, the maps of the mean in the left-hand side show a smooth but detailed pattern with significant spatial variation. Recall that prediction of clay, silt and sand are made with the constraint that they must sum to 100% at any prediction location. The clay, silt and sand maps have different range of predicted values. Variation of silt content is large, with values ranging from 13 to 97%. In contrast, clay values do not exceed 45%. Large clay content is found in the East of France, while the map of silt shows large silt content in the topsoils of the North-West part of France (Britain included) and a low content in the Massif Central mountains. Sand has its largest proportion in the latter mountains and in the Landes, while being almost unexistent in the upper North of France. Large values of OC (> 200 g kg^{-1}) are found in mountainous areas such as in the Alps, the Massif Central and in the Pyrenees. This is similar to the spatial pattern of the total Nitrogen (N) which in addition has large values in Britain (> 4 g kg^{-1}). High values of pH (> 8), indicating alkaline soils, are found in most parts of France, with the exceptions of Britain, the Vosges Mountains and the Massif Central.

The maps of the standard deviation in the right-hand side of Figs. 5 and 6 are different from those of the mean and seem to not share many common spatial features. They have a lower range of values than the maps of the mean. The largest value of the standard deviation is about two times smaller for clay, sand and OC, and about three times smaller for silt, pH and N than their associated mean maps. For clay, silt, OC and N, largest uncertainty is found in mountainous areas in the Alps or in the Pyrenees and in Britain for OC. Surprisingly, areas with the lowest uncertainty for silt is found in the North of France, where the mean map has the largest predicted values. A similar pattern is observed for sand in the Massif Central where largest sand values have small standard deviation. For sand, large uncertainty is observed in the Landes area and in a large patch South of Paris. The pattern of the standard deviation map of pH is rather scattered, with largest uncertainty found in the North-Western cost and in the Massif Central.

The median of $\boldsymbol{\delta}$ in Table 4 shows that the uncertainty is underestimated for all soil properties. Sand, OC and N are slightly underestimated ($\widetilde{\delta} < 0.57$) while clay, silt and pH are moderately underestimated ($\widetilde{\delta} < 0.68$). The 90% PICP shows however that the 90% prediction interval covers satisfactorily the observed values of clay, silt and sand but is too wide for pH and N and slightly too narrow for OC.

This is confirmed by the accuracy plot in Fig. 7: prediction intervals as obtained for the soil properties are too narrow, leading to an overall underestimation of the uncertainty. This is more severe for pH where deviation from the 1:1 line is entirely due to underestimation of the

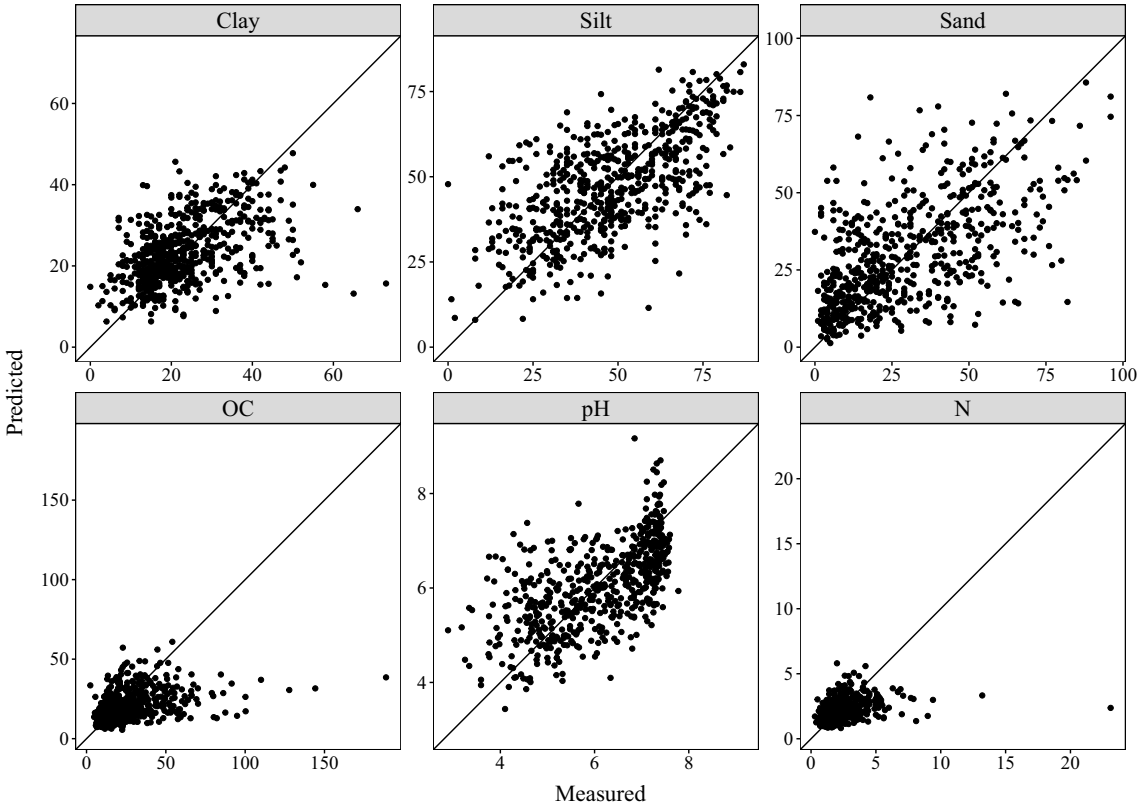

Fig. 4. Scatterplots of the measured against predicted by CNN soil properties. Clay, silt and sand are expressed in percent, OC and N in g kg^{-1}.

Table 3
Prediction accuracy as evaluated on the test set for CNN and RF.

Soil property	r^2		RMSE		ME		CCC	
	RF	CNN	RF	CNN	RF	CNN	RF	CNN
Clay	0.31	0.22	8.90	9.87	0.32	−0.36	0.46	0.45
Silt	0.44	0.40	13.50	14.38	0.93	0.55	0.60	0.62
Sand	0.42	0.36	16.08	17.47	−0.57	−0.37	0.57	0.59
OC	0.12	0.15	24.91	18.65	3.73	−6.32	0.23	0.46
pH	0.50	0.39	0.79	0.90	0.02	0.03	0.64	0.61
N	0.20	0.24	1.54	1.45	0.22	−0.29	0.31	0.20

uncertainty. Uncertainty for OC and N is underestimated for small values of nominal q while it is overestimated for nominal q values in the range $0.6 - 1$.

5. Discussion

5.1. Effect of the input window size

The window size of the input images had a significant impact on prediction accuracy, as shown by the soil properties RMSE on the test set. This is an expected result also reported in Padarian et al. (2019) and Wadoux et al. (2019). The window size closely relates to the amount of contextual information supplied to the model. CNN integrates spatial context by accounting for covariate pixels in the neighbour of a sampling location. More extra context improves the model predictions, but a too large amount of contextual information certainly acts as noise (Padarian et al., 2019). This confirms the study made by Smith et al. (2006) who found that there is an optimal range of window size in a case study deriving terrain attributes from a DEM for use in a soil survey. In our case study mapping soil properties at national scale, a window size between 9×9 pixels and 15×15 pixels was found optimal. Outside this range the prediction can be as much as 15% less accurate than the highest accuracy values. This is because soil property spatial variation is governed by complex relationships with soil forming factors. For example if the soil forming factor (i.e. one of the covariate) is homogeneous, one would like to increase the window size so as to capture sufficient information for the modelling. It is technically possible to integrate a covariate dependent window size in CNN, but it is not obvious that the improvement of the prediction outweighs the substantial increase of computational complexity.

A window size from 9×9 to 15×15 pixels is similar to incorporating contextual information in a radius of a location up to 4.5 to 7.5 km. Several authors (e.g. Padarian et al., 2019; Wadoux et al., 2019) have suggested that the window size could be associated to the range of spatial autocorrelation of the soil property. Wadoux et al. (2019) reported a range of a fitted spherical variogram of 329 m and 275 m for top- and subsoil organic carbon mapping. This was close to their optimal window size radius between 260 and 360 m. Padarian et al. (2019) compared their optimal window size to the variogram range from other studies mapping the same soil property in similar conditions. The authors found similar values of autocorrelation range between 150 and 450 m. This hypothesis was verified by fitting an exponential function to the sample variograms of the soil properties. I found values of the fitted range parameter between 3 km for OC and up to 8 km for sand. While there is large variability, this is close to the radius of the window size that is found optimal. However it is not possible to draw conclusions. Investigating this matter would certainly make a valuable extension for future DSM studies.

5.2. Prediction accuracy

For this case study, CNN provided good predictive ability as shown by the RMSE, ME, r^2 and CCC. Compared to RF as a baseline, CNN performs equally on average, while being more accurate for mapping OC and silt. Recent studies using deep learning for soil mapping (e.g. Behrens et al., 2018) have concluded that deep learning outperforms

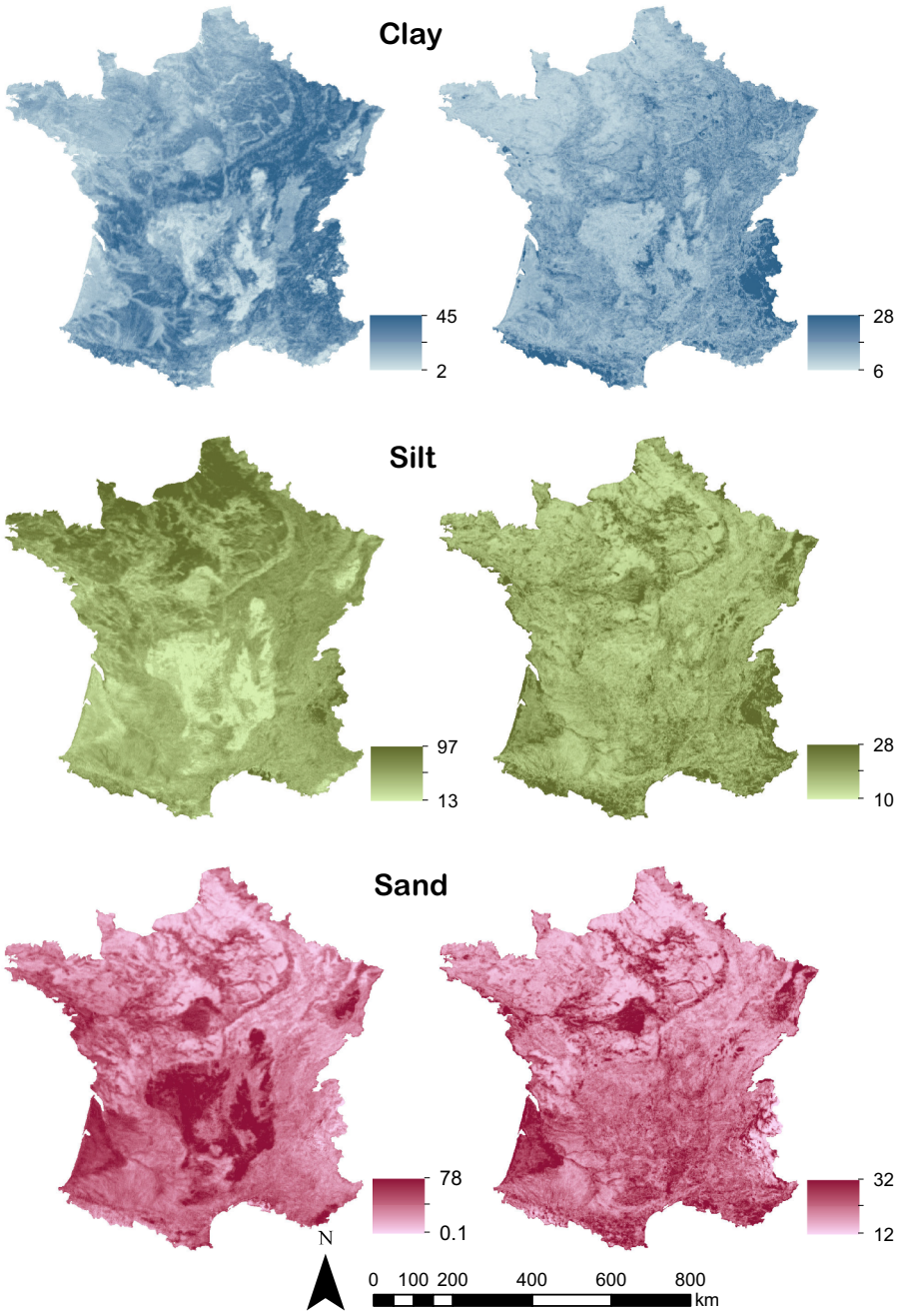

Fig. 5. Maps of the predicted mean (left) and standard deviation (right) for clay, silt and sand. Values are expressed in percent and sum to 100 for the predictions.

RF. This is here clearly not the case. However, those studies were using a small number of covariates to train the model. Padarian et al. (2019) used a DEM with two of its terrain derivatives in addition to rainfall and precipitation while Wadoux et al. (2019) used solely a DEM, a map of NDVI and a Landsat ETM band 5 image. A reason for CNN to perform well with a small number of covariate is possibly that the model creates a large number of hyper-covariates from the original images during convolution, while RF is purely data-driven and empirical on the existing covariates. When the number of covariate is large, this effect becomes negligible. In fact, it is acknowledged that RF is clearly favoured by a large set of covariates (Nussbaum et al., 2018), while being rapidly limited if the number of covariates is too small. However the present study is the first to employ a large number of environmental covariates for soil mapping using deep learning. The impact of the number of covariates on model predictions accuracy has to be further investigated.

5.3. Interpretation of the map features

It is beyond the scope of this study to give an interpretation of all the map features, but I summarize the most striking ones. The soil texture seems to be much influenced by the geology, as already noted by Ballabio et al. (2016). Clay content is high (> 40%) for a large part of Eastern France dominated by early Cretaceous lime and sandstone and middle or late Jurassic limestone. Silt has very high proportion in luvisols on late Cretaceous limestone and on cambisols in Britain. The pattern of sand content is close to that of old massifs such as in the Massif Central of the Vosges characterized by late carboniferous metamorphic bedrocks. In addition, sand content is high in the Landes, constituted of Pliocene fluviatil rocks. The maps of clay, silt and sand closely resemble those made by others using the same dataset (e.g. Ballabio et al., 2016). Predicted maps also share the same pattern than SoilGrid products for France (Hengl et al., 2014). The maps of OC, pH

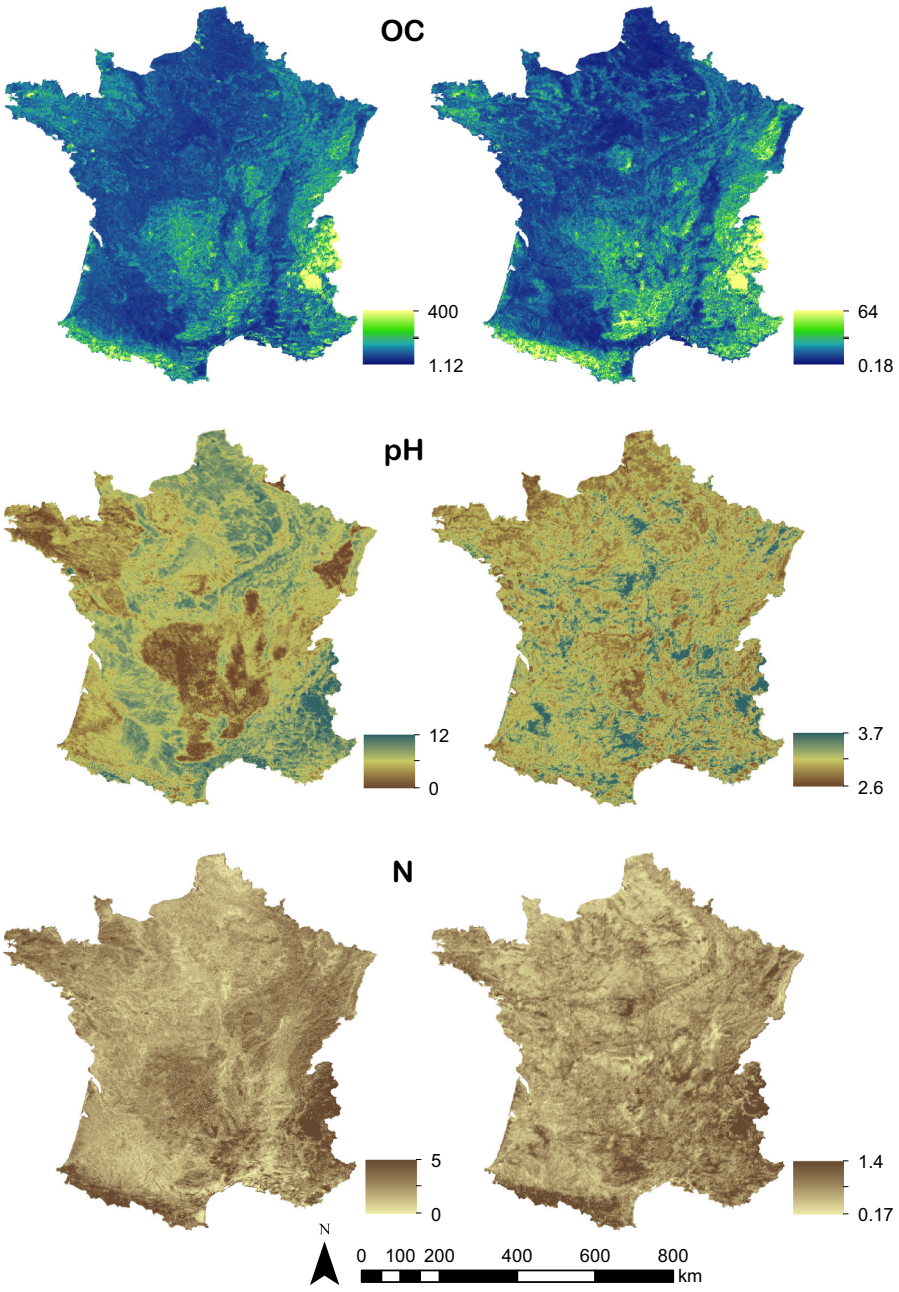

Fig. 6. Maps of the predicted mean (left) and standard deviation (right) for OC, pH and N. Values for OC and N are expressed in g kg^{-1}.

Table 4
Evaluation of the uncertainty quantification for the CNN model.

Soil property	$\tilde{s}$	PICP 90%
Clay	0.67	90.51
Silt	0.63	89.49
Sand	0.54	92.88
OC	0.55	79.15
pH	0.68	99.83
N	0.57	98.31

and N underline the influence of climate, land use and geology. OC follows the same distribution as the precipitation and temperature with larger OC content in mountainous areas. In Britain, this is due to slurry production related C input (Meersmans et al., 2012). The overall pattern is in accordance with Martin et al. (2010) and Meersmans et al. (2012). Properties pH and N are more difficult to interpret, showing strong correlation with precipitation and temperature, but also geology. PH seems to be acidic in areas of metamorphic and plutonite rocks, while N content is high in all mountainous areas, with smaller temperature and precipitation rate. As for the soil texture, OC and pH maps are close to those produced by SoilGrid (Hengl et al., 2014).

5.4. Uncertainty quantification

It is the first study to propose a method to quantify the uncertainty of the prediction for mapping using a neural network model. On average, the prediction uncertainty was slightly underestimated (Table 4). This is an expected result reported in previous studies (Khosravi et al., 2011). The reason is that estimation of the $\sigma_{\hat{f}}$ is dependent on B neural network. Each network is calibrated using a bootstrap sample of the input data. A specific sample might be unrepresentative of the population and cause inaccurate estimation of the variance parameter and underestimation of the total variance. Other

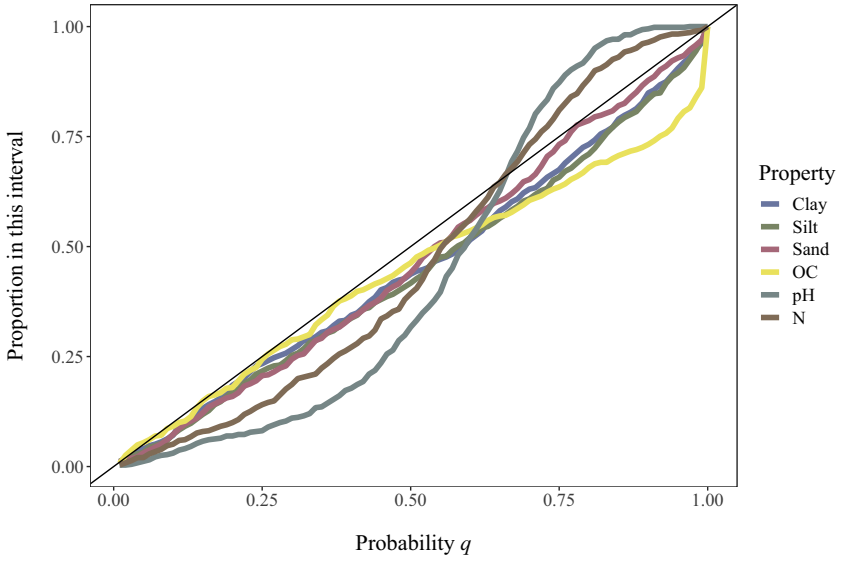

Fig. 7. Accuracy plot of the soil properties.

methods for neural networks uncertainty quantification exist, such as the Delta (De Vleaux et al., 1998), or the Bayesian uncertainty analysis (MacKay, 1992) method, but those have not been tested. Albeit underestimated, the uncertainty of the prediction was accurately quantified, as shown by the accuracy plots and PICP. Caubet et al. (2019) reported the PICP of the LUCAS prediction (from the maps made by Ballabio et al. (2016)) of clay and sand for France and found values of 8 and 15%, respectively. In the present study were obtained 90.51% for clay and 92.88% for sand, which are values very close to the expected value of probability interval. The uncertainty was therefore much better quantified than this existing work using the same dataset in similar conditions.

5.5. Multivariate mapping using CNN

Six soil properties were predicted using a single multi-task CNN model. The CNN shared a common architecture for all soil properties. This reduces both the risk of overfitting and computational resources that would be needed to fit each model separately. This is important for future DSM studies because the number of available geodata is constantly increasing (Nussbaum et al., 2018). In addition, the multi-task CNN architecture can easily be modified to predict at several depths. Padarian et al. (2019) and Wadoux et al. (2019) have shown that this is feasible and lead to a substantial increase of prediction accuracy in deeper soil layers. Several attempts have been made for multivariate soil mapping (e.g. by Angelini et al. (2017)). However, there have been very little, if any, interest for multivariate soil mapping using machine learning techniques. Hengl et al. (2018) was the first to investigate the use of a multi-output random forest. However, the method proposed by Hengl et al. (2018) has the major disadvantage that the data size increases rapidly as the number of outputs expands. In addition, deriving prediction intervals for each output separately remains a challenge. In the case study presented here, it is shown that building a single model for multiple outputs is feasible, and lead to prediction accuracy comparable to those made by a univariate RF model.

5.6. Correlation between outputs

Recent studies (e.g. Angelini et al., 2017) argued that one of the strengths of a multivariate model is its capability to preserve the correlation between soil properties. In Angelini et al. (2017), the correlation between properties is assessed as resulting from a calibrated SEM model. One cannot assess internal correlation between properties in a ML model, but it is possible to measure whether the correlation between original and predicted soil properties is preserved. In this study, the correlation between predicted outputs is on average slightly better preserved by the univariate random forest model than by the CNN model, as assessed by the properties Pearson's r correlation coefficient matrix. This can be explained by the r^2 values of the predictions made by the RF model that are, for most properties, closer to one than those of the CNN model. However, since the correlation between outputs is not modelled explicitly, it would be misleading to conclude that the model retains the correlation between outputs simply by assessing the correlation or covariation between its predictions. I hope this clarifies why the correlation between outputs was not used for interpretation purposes in this study. One can include explicitly the correlation between outputs by calibrating additional stochastic variables together with the CNN model parameters (Uria et al., 2016). Another solution is to modify the loss function so that a criterion related to the absolute difference between the correlation among original and predicted soil properties is minimized jointly with the MSE. Note also that in this study, a CNN with a shared architecture is built, and with separated branches for each soil property. I speculate that the correlation between outputs would be better retained by using a shared architecture only, and by predicting directly after the flattening operation. The effect of the neural network architecture on the predictions and methods to include explicitly correlation between outputs need further investigation. I welcome more research in this area.

5.7. Assumptions made during modelling

The method described in this study for uncertainty quantification assumes independent and normally distributed residuals. The latter assumption was tested by visual inspection of the residuals and computation of a quantile-quantile (Q-Q) plot (not shown). It was found that the Gaussian assumption of the residuals was satisfied. In cases where the Gaussian assumption is too restrictive, one could use a transformation of the data prior to modelling (e.g. logarithm, square-root or Box-Cox transform). Remaining spatial autocorrelation in the residuals was further tested to ensure that the assumption of independence was satisfied. Spatial autocorrelation in the residuals indicates that the prediction might be biased, and that the quantified uncertainty is possibly underestimated. The Moran's I (MI) (Bivand and Wong, 2018) was used to test for spatial autocorrelation of the residuals of the maps presented in Figs. 5 and 6. The values of the MI range from -1 to 1, with 0 indicating no autocorrelation. Ten nearest neighbours

were used to compute the MI on the original values and residuals of the soil properties. Results indicate significant spatial autocorrelation in the original soil properties, with a MI value larger than 0.27 for all soil properties (e.g. MI value for pH is 0.41), while being smaller than 0.1 on the residuals (MI value for residuals of pH is 0.05). Thus, the CNN model was efficient to eliminate spatial autocorrelation. In case there would remain autocorrelation in the residuals, predictions and prediction uncertainty quantification would be improved by kriging the residuals, and by summing the kriged mean and variance maps to the predicted mean and variance maps made by the CNN model. An example can be found in Rossel et al. (2015).

6. Conclusion

From the results and discussion I draw the following conclusions:

- It is feasible to use convolutional neural network to map soil properties simultaneously. Extensions can further be made to predict at multiple depths. CNN has the advantage that it incorporates explicitly the contextual information of the covariates in the neighbour of a location. In addition, CNN as for other machine learning models, does not rely on rigid statistical assumptions about the distribution of the soil properties.
- The window size of the input covariate images had a significant impact on prediction accuracy. The optimal window size closely relates to the range of the spatial autocorrelation of the soil property. However, there is a need for further research so as to generate rules and enable to choose a priori the optimal size.
- The predicted maps of clay, silt, sand, OC, pH and N have a detailed pattern with significant spatial variation. These maps showed great similarities with those of the same soil properties produced by others using different mapping techniques. Statistical validation of the prediction accuracy indicated that CNN performs on average equally to random forest.
- Bootstrap plus variance estimate was used to quantify the uncertainty of the predictions. Each soil property has its own standard deviation map, which can be used to derive prediction intervals. Validation of the uncertainty assessment indicated that the uncertainty is accurately quantified, albeit slightly underestimated. Comparison with other studies using the same dataset over France showed that the proposed method quantifies much better the uncertainty.

Acknowledgements

Alexandre Wadoux received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607000. The author thanks ISRIC - World Soil Information and Tom Hengl, Envirometrix Ltd., Wageningen, for providing the covariates used in the case study.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: A System for Large-Scale Machine Learning. 16. OSDI, pp. 265–283.

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., 2014. Digital mapping of soil particle-size fractions for Nigeria. Soil Sci. Soc. Am. J. 78, 1953–1966.

Allaire, J.J., Chollet, F., 2018. Keras: R interface to 'Keras'. URL. https://CRAN.R-project.org/package=keras (r package version 2.2.0).

Angelini, M.E., Heuvelink, G.B.M., Kempen, B., 2017. Multivariate mapping of soil with structural equation modelling. Eur. J. Soil Sci. 68, 575–591.

Ballabio, C., Panagos, P., Monatanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110–123.

Behrens, T., Schmidt, K., MacMillan, R.A., Viscarra Rossel, R.A., 2018. Multi-scale digital soil mapping with deep learning. Sci. Rep. 8, 15244.

Bivand, R.S., Wong, D.W.S., 2018. Comparing implementations of global and local indicators of spatial association. TEST 27, 716–748.

Caubet, M., Dobarco, M.R., Arrouays, D., Minasny, B., Saby, N.P., 2019. Merging country, continental and global predictions of soil texture: lessons from ensemble modelling in France. Geoderma 337, 99–110.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., et al., 2015. Global land cover mapping at 30 m resolution: a POK-based operational approach. ISPRS J. Photogramm. Remote Sens. 103, 7–27.

De Vleaux, R.D., Schumi, J., Schweinsberg, J., Ungar, L.H., 1998. Prediction intervals for neural networks via nonlinear regression. Technometrics 40, 273–282.

Deutsch, C.V., 1997. Direct assessment of local accuracy and precision. Geostatistics Wollongong 96, 115–125.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. 57 CRC press, New York.

Fan, Y., Li, H., Miguez-Macho, G., 2013. Global patterns of groundwater table depth. Science 339, 940–943.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep Learning. 1 MIT press, Cambridge.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford Univ. Press, New York.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., et al., 2013. High-resolution global maps of 21st-century forest cover change. Science 342, 850–853.

Hartmann, J., Moosdorf, N., 2012. The new global lithological map database GLiM: a representation of rock properties at the Earth surface. Geochem. Geophys. Geosyst. 13, 1–37.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., et al., 2014. Soilgrids1km-global soil information based on automated mapping. PLoS One 9, e105992.

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

Heuvelink, G.B.M., 2014. Uncertainty quantification of globalsoilmap products. In: GlobalSoilMap: Basis of the Global Spatial Soil Information System. Proceedings of 1st GlobalSoilMap Conference, pp. 335–340.

Heuvelink, G.B.M., Kros, J., Reinds, G.J., De Vries, W., 2016. Geostatistical prediction and simulation of European soil property maps. Geoderma Regional 7, 201–215.

Jones, A., Montanarella, L., Jones, R., et al., 2005. Soil Atlas of Europe. European Commission.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. Scientific data 4, 170122.

Khosravi, A., Nahavandi, S., Creighton, D., Atiya, A.F., 2011. Comprehensive review of neural network-based prediction intervals and new advances. IEEE Trans. Neural Netw. 22, 1341–1356.

Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. In: 3rd International Conference for Learning Representations, San Diego, 2015, pp. 1–15.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, pp. 6402–6413.

Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. Eur. J. Soil Sci. 51, 137–157.

Lark, R.M., Milne, A.E., Addiscott, T.M., Goulding, K.W.T., Webster, C.P., O'Flaherty, S., 2004. Scale-and location-dependent correlation of nitrous oxide emissions with soil properties: an analysis using wavelets. Eur. J. Soil Sci. 55, 611–627.

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

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

MacKay, D.J.C., 1992. The evidence framework applied to classification networks. Neural Comput. 4, 720–736.

Martin, M.P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., Arrouays, D., 2010. Spatial distribution of soil organic carbon stocks in France: discussion paper. Biogeosciences 8, 1053–1065.

McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

Meersmans, J., Martin, M.P., Lacarce, E., De Baets, S., Jolivet, C., Boulonne, L., Lehmann, S., Saby, N.P.A., Bispo, A., Arrouays, D., 2012. A high resolution map of French soil organic carbon. Agron. Sustain. Dev. 32, 841–851.

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

Nix, D.A., Weigend, A.S., 1994. Estimating the mean and variance of the target probability distribution. In: Neural Networks, IEEE World Congress on Computational Intelligence. 1. IEEE, pp. 55–60.

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

Olaya, V., Conrad, O., 2009. Geomorphometry in SAGA. Dev. Soil Sci. 33, 293–308.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning for digital soil mapping. SOIL 5, 79–89.

Pelletier, J.D., Broxton, P.D., Hazenberg, P., Zeng, X., Troch, P.A., Niu, G.-Y., Williams, Z., Brunke, M.A., Gochis, D., 2016. A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling. Journal of Advances in Modeling Earth Systems 8, 41–65.

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

Robinson, N., Regetz, J., Guralnick, R.P., 2014. EarthEnv-DEM90: a nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. ISPRS J. Photogramm. Remote Sens. 87, 57–67.
Rossel, R.A.V., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Research 53, 845–864.
Shrestha, D.L., Solomatine, D.P., 2008. Data-driven approaches for estimating uncertainty in rainfall-runoff modelling. International Journal of River Basin Management 6, 109–122.
Smith, M.P., Zhu, A.-X., Burt, J.E., Stiles, C., 2006. The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma 137, 58–69.
Tóth, G., Jones, A., Montanarella, L., 2013. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environ. Monit. Assess. 185, 7409–7425.
Tucker, C.J., Grant, D.M., Dykstra, J.D., 2004. NASA's global orthorectified landsat data set. Photogramm. Eng. Remote Sens. 70, 313–322.
Uria, B., Côté, M.-A., Gregor, K., Murray, I., Larochelle, H., 2016. Neural autoregressive distribution estimation. The Journal of Machine Learning Research 17, 7184–7220.
Wadoux, A.M.J.C., Brus, D.J., Heuvelink, G.B.M., 2018. Accounting for non-stationary variance in geostatistical mapping of soil properties. Geoderma 324, 138–147.
Wadoux, A.M.J.C., Padarian, J., Minasny, B., 2019. Multi-source data integration for soil mapping using deep learning. SOIL 5, 107–119.
Xu, S., An, X., Qiao, X., Zhu, L., Li, L., 2013. Multi-output least-squares support vector regression machines. Pattern Recogn. Lett. 34, 1078–1084.