Response to comments

Anonymous Referee #1

- Section 4.1. Data: how many data points? How many data points for each depth interval? I know it is all described in the Padarian et al 2017, but it is quite important information to evaluate data augmentation effects. It is also mentioned in section 4.5, but without depth information.

We added extra information about the dataset.

- Section 4.2. Data augmentation: I think the comparison should show the results of cubist with data augmentation. Data augmentation is something often associate with neural networks modelling, but it could be used for any models. I do not think it is a level comparison if one model has X sampling points and the other has X*4 sampling points. Figure 4 shows the reduction of RMSE with data augmentation. However in figure 5 the results without data augmentation are not compared with cubist. I also think that the authors should at least discuss (if not test) the potential effects of data augmentation on the spatial auto-correlation of the data. Unfortunately, I do not think that the results presented by the authors are enough to describe the effect of data augmentation on the final outcomes.

Data augmentation can certainly be used with other machine learning techniques, not only CNNs. It is reasonable to think that we should use data augmentation in the comparison but the fact is that the previous study (Padarian et al., 2017) and most researchers use point information (a vector of covariates) as input for the Cubist model. Transforming that vector into the same 3D structure that we used in this study generates a 1x1 image with n channels, where n is the number of covariates. Rotating that image by 90, 180 and 270º generates exact copies of the image, which is oversampling and not data augmentation.
A possible alternative is to use some context (pixels from the vicinity), but Cubist is not designed to handle that type of information, making the comparison even less level. If the 3D array is flattened as a vector, adding more context actually increases the prediction error.

In terms of the autocorrelation of the data, we are assuming that there is no variance when distance=0 by adding samples with exactly the same SOC content. That is theoretically true if we consider that the distance is exactly equal to 0. In practical terms, when calculating the semivariogram, the semivariance value of the first bin will be lower, but that does not significantly affect the final model. We will expand the discussion to clarify this point.

- Section 4.5 Training and validation: I am a bit confused about how the various datasets are named (see comments for figure 5 as well). My understanding is (further comments are based on this): Test : 10% of data never used in the fitting of the model. Remaining data (90%): bootstrap splitting in training and validation (1/3)

Test dataset: 10% of original data (n=49)
We augmented the remaining 90% (n=436) obtaining 1,744 samples. With those samples, we ran a bootstrapping routine. At each repetition, we sampled with replacement (obtaining 1,744 samples) which we used as a training dataset. A sampling with replacement usually draws around 2/3 (~66.66%) of the samples, hence excluding ~33.33% of the samples (which we used as a validation set and to select hyperparameters). That is the meaning of the 1/3. Just to clarify, 1/3 is not a 1:3 ratio between the size of the training and validation sets.

- Figure 4: The largest reduction in RMSE happened for the deepest depth interval. I assume that this interval was also the one with fewest data points (as it is usually in soil datasets). Is this an effect of data augmentation? I.e. stronger effect when there is the lowest number of points?

Because we are using a multi-task CNN, we excluded the samples with depths lower than 100cm. In consequence, all depths have the same number of observations (we will add a clarification to make it more explicit).

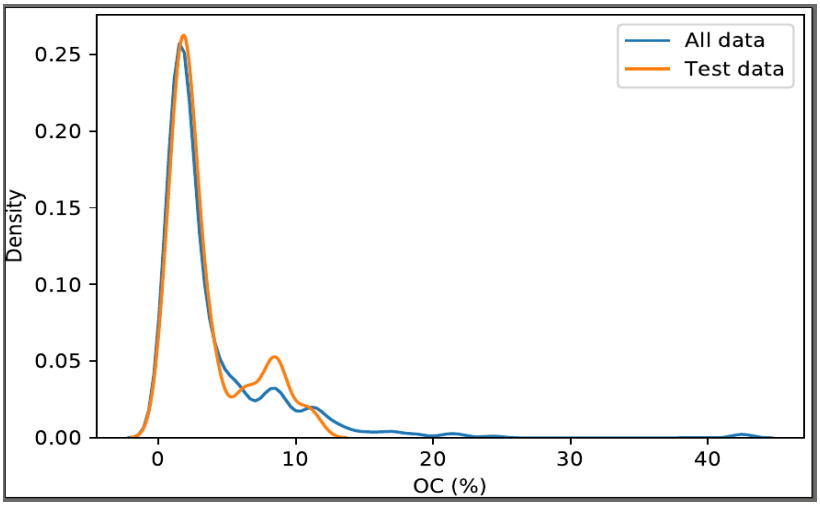
- Figure 5: How is it possible that the test dataset (i.e. the one that was never used in the model fitting) has a lower error than train itself or even validation? (see above comment for my interpretation of these labels) If this is not a label problem, I think the authors should really explain and discuss this, because, to me, it seems a problem of overfitting. I would also like to see a comparison with cubist model with data augmentation.

Actually, it is perfectly possible to have a lower error in the test dataset. That would be the case if the SOC content of the samples are relatively low (which is the case for this test dataset), because the error is higher in samples with larger SOC content.
A lower error in the test dataset is not an indication of overfitting. A larger error would be an indication of overfitting.

Reponse from round 2: If the test dataset has a lower mean value than the dataset on which the model is fitted, then there are some potential problems in how the test dataset was selected. Test dataset should be representative of the whole population. I think this is fundamental to test, after the comments from the authors. Also the test dataset is only 49 sampling points versus the 1744 (with augmented data). Was the split repeated? If the test dataset is representative of the whole population and still has lower errors, then could the authors please find some other cases in the literature when this happens? A model that performs better on a completely independent dataset it is rather exceptional.

The 49 points were selected on the original data, before data augmentation, in order to have the same test for both cases.
The split was not repeated because the idea is to exclude the test set from the training. In a repeated sample, some of the test data would be present in the training or validation sets, defeating the purpose of having a "completely independent" test set.

As we mentioned in the manuscript (line 1, p8), we performed a random selection. Since the distribution of the OC is not normal, the distribution of the test set is not exactly the same (see attached figure) and that explains the lower error.



We acknowledge that it would be possible to sample the original data in a different way (sampling by quantiles), and that would increase the error shown in Figure 5 (only in the test set). Most certainly, the error wouldn't be significantly different than the validation error, since the test set would have a distribution just like the rest of the data, which error is accounted during the validation.

We propose adding to the revised version a paragraph explaining this and the attached figure.

- Section 5.6 Uncertainty. The maps show only a small area and the text is a bit confusing. For example: "greater reductions in higher (larger?) areas of the landscape" is not very clear or precise. Maybe it would be better to show a measure of the proportion of the landscape that has a great reduction of intervals width?

We will re-phrase the text to make it clearer. "... higher areas of the landscape" is in terms of elevation, which is where we saw high uncertainty in Padarian et al., (2017). About the proportion of the landscape that has a great reduction of intervals width, we think that a map with a proper colour palette shows exactly that, and much better than a non-spatial calculation.

Anonymous Referee #2

- Is it possible to predict a set of properties at the same time? Eg CEC and Clay and C for example?

It is absolutely possible and it has proven to be quite effective. We just published a study on multi-task CNNs for soil spectroscopy that shows that (DOI: 10.1016/j.geodrs.2018.e00198).

It is important to consider that predicting more properties usually means having a bigger network hence more parameter and the consequent risk of overfitting if there is not enough data. So there is a limit on how much you can predict simultaneously, but it will depend on the task.

- P6 l13 ReLU?

ReLU is an activation function to add non-linearities to the model, commonly used in CNNs. We will add a short description to clarify that.

- P8 L24 this an important step I think. This should be highlighted in the introduction.

The reviewer is referring to "To incorporate contextual information for DSM prediction, we represent the input as image".
We agree that it is an important point, that is why is fully-explained in the rationale section.

- Figure 5: it is very rare to observe lower error in the test dataset than train itself or even validation? Could you comment on that in the paper?

We will comment on that. The error in the test set is lower due to the distribution being different to the training data. We performed 1 random sampling to select the test set and the range OC concentrations is narrower and lower on average. We give a more complete explanation in the response to reviewer 1.

- Section 5.6 The discussion on the prediction of uncertainty needs more global result. I think you can provide a PICP plot using the test dataset to better justify your results.

Very good suggestion. We will include it in the revised version.

Using deep learning for Digital Soil Mapping

José Padarian, Budiman Minasny, and Alex B. McBratney

Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, New South Wales, Australia

Correspondence: José Padarian (jose.padarian@syndey.edu.au)

Abstract. Digital soil mapping has been widely used as a cost-effective method for generating soil maps. However, current DSM data representation rarely incorporates contextual information of the landscape. DSM models are usually calibrated using point observations intersected with spatially corresponding point covariates. Here, we demonstrate the use of the convolutional neural network model that incorporates contextual information surrounding an observation to significantly improve the prediction accuracy over conventional DSM models. We describe a convolutional neural network (CNN) model that takes inputs as images of covariates and explores spatial contextual information by finding non-linear local spatial relationships of neighbouring pixels. Unique features of the proposed model include: input represented as 3D stack of images, data augmentation to reduce overfitting, and simultaneously predicting multiple outputs. Using a soil mapping example in Chile, the CNN model was trained to simultaneously predict soil organic carbon at multiples depths across the country. The results showed the CNN model reduced the error by 30% compared with conventional techniques that only used point information of covariates. In the example of country-wide mapping at 100 m resolution, the neighbourhood size from 3 to 9 pixels is more effective than at a point location and larger neighbourhood sizes. In addition, the CNN model produces less prediction uncertainty and it is able to predict soil carbon at deeper soil layers more accurately. Because the CNN model takes covariate represented as images, it offers a simple and effective framework for future DSM models.

1 Introduction

Digital soil mapping (DSM) has now been widely used globally for mapping soil classes and properties (Arrouays et al., 2014). In particular, DSM has been used to map soil carbon efficiently around the world (e.g. Chen et al. (2018)). DSM methodology has been adopted by FAO (FAO, 2018) so that digital soil maps can be produced reliably for sustainable land management. While DSM can be said is now operational, there are still unresolved methodological issues regarding better representation of landscape pattern and soil processes. Some of the methodological research studies include the use of multiple remotely sensed images (Poggio and Gimona, 2017) or time series of images as covariates (Demattê et al., 2018), testing novel regression and machine learning models (Angelini et al., 2017; Somarathna et al., 2017), and incorporation of spatial residuals of the regression model (Keskin and Grunwald, 2018; Angelini and Heuvelink, 2018).

The formalisation of the DSM methodology was done by the publication of McBratney et al. (2003). Following the ideas of Dokuchaev (1883) and Jenny (1941), they described the *scorpan* model as the empirical quantitative relationship of a soil attribute and its spatially implicit forming factors. Such factors correspond to *a)* s: soil, other properties of the soil at a point; *b)* c: climate, climatic properties of the environment at a point; *c)* o: organisms, vegetation or fauna or human activity; *d)* r: topography, landscape attributes; *e)* p: parent material, lithology; *f)* a: age, the time factor; and *g)* n: space, spatial position. Explicitly, the *scorpan* model can be written as:

$$S_{(x,y)} = f(s_{(x,y)}, c_{(x,y)}, o_{(x,y)}, r_{(x,y)}, p_{(x,y)}, a_{(x,y)}, n_{(x,y)}) + e_{(x,y)} \quad (1)$$

where (x, y) corresponds to the coordinates of a soil observation, and e is the spatial residual.

The usual steps for deriving the *scorpan* spatial soil prediction functions include intersecting soil observations (point data) with the *scorpan* factors (raster images at a particular resolution), and calibrating a prediction function f. In effect, we are only looking at relationships between point observations and point representation of covariates. The *scorpan* factors have implicit spatial information, however the prediction function f does not explicitly take into account the spatial relationship.

Attempts have been made to incorporate more local information in the *scorpan* covariates, in particular topography. Approaches to include covariates information about the vicinity around the observations (x, y) have been devised. One approach is to derive topographic or terrain attributes (e.g. slope, curvature) at multiple scales by expanding the size of the window or neighbour size in the calculation (Miller et al., 2015; Behrens et al., 2010). Another approach includes multi-scale analysis using spatial filters such as wavelets on the covariate raster (Biswas and Si, 2011; Sun et al., 2017). Thus, the raster represents larger spatial support. Studies indicated that, generally, covariates with larger support than its original resolution could enhance the prediction accuracy of the model (Mendonça-Santos et al., 2006; Sun et al., 2017).

DSM can be thought as linking observable landscape structure and soil processes expressed as observed soil properties. To effectively link structure and processes, Deumlich et al. (2010) suggested the use of analysis that spans over several spatial and temporal scales. Behrens et al. (2018) proposed the contextual spatial modelling to account for the interactions of covariates across multiple scales and their influence on soil formation. The authors' approach (e.g.: Behrens et al. (2010, 2014)) derived covariates based on the elevation at the local to the regional extent. Their approaches include ConMap (Behrens et al., 2010) which is based on elevation differences from the centre pixel to each pixel in a sparse neighbourhood, and ConStat (Behrens et al., 2014) which used statistical measures of elevation within growing sparse circular spatial neighbourhoods. These approaches produce a large number of predictors computed for each location, as shown in an example with 100 distance scales (e.g., from 20 m to 20 km) and 1000 predictors per grid cell. These hyper-covariates, solely based on elevation, are used as predictors in a random forest regression model.

Spatial filtering, multi-scale terrain calculation, and contextual mapping approaches require the pre-processing of each covariate independently. The useful scale for each covariate needs to be figured out via numerical experiments and most of the time the process relies on ad hoc decisions. Here, we take advantage of the success of deep learning models that are used for image recognition, as an effective tool in DSM to optimally search for local contextual information of covariates. This work

aims to expand the classic DSM approach by including information about the vicinity around (x, y), to fully leverage the spatial context of a soil observation. The aim is achieved by devising a convolutional neural network (CNN) which can take multiple spatial contextual inputs.

2 Rationale

The theoretical background of DSM is based on the relationship between a soil attribute and soil forming factors. In practice, a single soil observation is usually described as a point p with coordinates (x, y) (Eq. 1), and the corresponding soil forming factors are represented by a vector of pixel values of multiple covariate rasters $(a_1, a_2, ..., a_n)$ at the same location, where n is the total number of covariate rasters.

Soils are highly dependant on their position in the landscape, and information at a particular pixel might not be sufficient to represent that complex relationship. Our method expands the classic DSM approach by replacing the covariates, usually represented as a vector, with a 3D array with shape (w, h, n), where w and h are the width and height in pixels of a window centred at point p (Fig. 1). Methods commonly used in DSM are not designed to adequately handle the data structure depicted in Fig. 1. The data representation is similar to the network model by Lee and Kwon (2017) which used hyperspectral images for classification purposes.

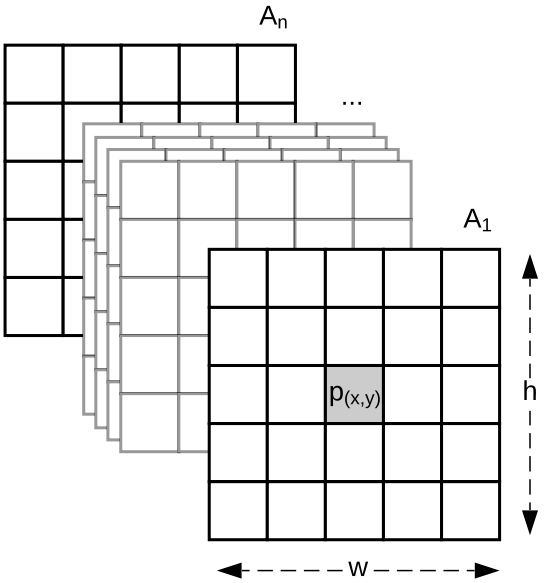


Figure 1. Representation of the vicinity around a soil observation p, for n number of covariate rasters. w and h are the width and height in pixels, respectively.

As described in the introduction, while multi-scale or contextual mapping approaches have been used in DSM, they still rely on a vector representation of covariate and rely on machine learning methods such as random forest to select important

predictors. While deep learning methods have been used in DSM (e.g., (Song et al., 2016)), most studies still use a vector representation of covariate.

In the following sections, we introduce the use of convolutional neural networks (CNNs) to exploit spatial information of covariates that will perform a more effective DSM.

3 Deep learning

Deep learning is a machine learning method that is able to learn the representation of data through a series of processing layers. In agriculture and environmental mapping, it is mainly used in hyperspectral and multispectral image classification problems, e.g. land cover classification (Kamilaris and Prenafeta-Boldú, 2018). We have not seen much application of deep learning in DSM, except for Song et al. (2016) which used the deep belief network for predicting soil moisture.

In this section we briefly introduce CNNs and some associated methods used during this work. For a more detailed and general description about CNNs we refer the reader to LeCun et al. (1990) and Krizhevsky et al. (2012).

3.1 CNN

CNNs are based on the concept of a layer of convolving windows which move along a data array in order to detect features (e.g.: edges) of the data by using different filters (Fig. 2). When stacked together, convolutional layers are capable of extracting features of increasing complexity and abstraction (LeCun et al., 1990). Since CNNs have the capacity to leverage the spatial structure of the data, they have been widely and effectively used in computer vision for image recognition or extraction (LeCun et al., 1995).

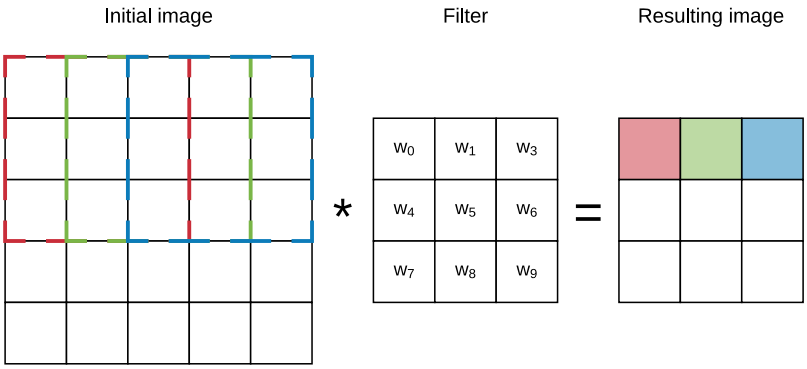


Figure 2. Example of the first 3 steps of a convolution of a 3x3 filter over a 5x5 array (image). The resulting pixel values correspond to the sum of the element-wise multiplication of the initial pixels (dashed lines) and the filter.

A CNN has a number of three dimensional hidden layers, each layer learning to detect different features of the input images (LeCun et al., 2015). In our case, each of the layers can perform one of the two types of operations: convolution, or pooling. Convolution takes the input images through a set of convolutional filters (e.g., a 3x3 size filter), each of which detects and

enhances certain features from the images. Units in a convolutional layer are organised in feature maps (here we used 3 x 3). Each unit of the feature map is connected to local patches in the feature maps of the previous layer through a set of weights. These local weighted sum is then passed through a non-linear transfer function.

A pooling operation merges similar features by performing non-linear down-sampling. Here we used Max-Pooling layers which combine inputs from a small 2x2 window. Pooling also makes the features robust against noise. All the convolutional and pooling layers are finally "flattened" to the fully connected layer. In effect, the fully connected layer is a weighted sum of the previous layers.

To obtain optimal weights for the network, we train the network using a training data set. Weights were adjusted based on a gradient-based algorithm to minimise the error using an Adam optimiser (Kingma and Ba, 2014). We refer to a review by LeCun et al. (2015) on the details of CNN.

3.2 Multi-task learning

CNNs have the capacity to predict multiple properties simultaneously. By doing so, a multi-task CNNs is capable of sharing learned representations between different targets and also use the other targets as "clues" during the prediction process. In consequence, the error of the simultaneous prediction is generally lower compared with a single prediction for each target ~~(?Ramsundar et al., 2015)~~(Padarian et al., 2019; Ramsundar et al., 2015). An additional advantage of using a multi-task CNN is the reported reduction on the risk of overfitting (Ruder, 2017).

In DSM, where the combination of large extents, high resolution, and bootstrap routines, leads to running multiple model realisations on billions of pixels, combined with the fact that CNNs use a group of pixels around the soil observation instead of a single pixel, the time and computational resources required for training and inference is an important factor. Due to the simultaneous training an inference of multiple targets, a multi-task CNN present the advantage of reducing both, training and inference time, compared with a single-task model.

4 Methods

4.1 Data

The data used in this work correspond to Chilean soil information. Since most observations are distributed on agricultural lands, we complemented that information with a second small data collection compiled from the literature and collaborators. We selected soil organic carbon content (%) at depths 0–5, 5–15, 15–30, 30–60 and 60–100 cm as our target attribute. 485 soil profile where used after excluding soil profiles with total depth lower than 100 cm (in order to assure that all the profiles have observations at all depth intervals). For more details about the data and depth standardisation we refer the reader to Padarian et al. (2017).

As covariates, we used a) digital elevation model (HydroSHEDS, Lehner et al. (2008)), which are provided at 3 arc-second resolution, in addition to its derived slope and topographic wetness index, calculated using SAGA (Conrad et al., 2015); and b)

long term mean annual temperature and total annual rainfall derived from information provided by WorldClim (Hijmans et al., 2005), at 30 arc-second resolution. All data layers were standardised to a 100 m grid size.

4.2 Data augmentation

Deep learning techniques are described as "data-hungry" since they usually work better with large volumes of data. The direct effect of data augmentation is to generate new samples by modifying the original data without changing its meaning (Simard et al., 2003). To achieve this, we rotated the 3D array shown in Fig. 1 by 90, 180, and 270 degrees, hence quadruplicating the number of observations. It is important to note that the central pixel preserves its initial position.

A secondary effect of data augmentation is regularisation, reducing the variance of the model and overfitting (Krizhevsky et al., 2012). Data augmentation also induces rotation invariance (Vo and Hays, 2016) by generating alternative situations (rotated data) where the model response should be similar to the original data (e.g: a soil profile next to a gully is expected to be similar to a profile next to the opposite side of the gully, *ceteris paribus*).

4.3 Network architecture

The multi-task CNN used in this study (Fig. 3; Table 1) consists of an input layer pass through a series of convolutional and pooling layers with a ReLU activation function, which adds a non-linearity by passing the learned weights through the function $f(x) = max(0, x)$. The initial common/shared network has a function of extracting features shared between the five target depth ranges. Next, the common features are propagated through independent branches, one per depth range, of 3 fully-connected layers.

Table 1. Sequence of layers used to build the multi-task neural network

Layer type	Kernel size	Filters	Activation
†Convolutional	3x3	16	ReLU
†Max-Pooling	2x2	-	-
†Dropout (0.3)	-	-	-
†Convolutional	3x3	32	ReLU
‡Fully-connected	-	10	ReLU
‡Dropout (0.3)	-	-	-
‡Fully-connected	-	10	ReLU
‡Fully-connected	-	1	ReLU

†Shared layers; ‡for each property

The multiple connection between the layers generates a high number of parameters. In order to reduce the risk of overfitting, we introduce a dropout rate. In between the layers, 0.3 of the connections were randomly disconnected (Nitish et al., 2014). We added this dropout operation in the shared layer and another dropout before the output.

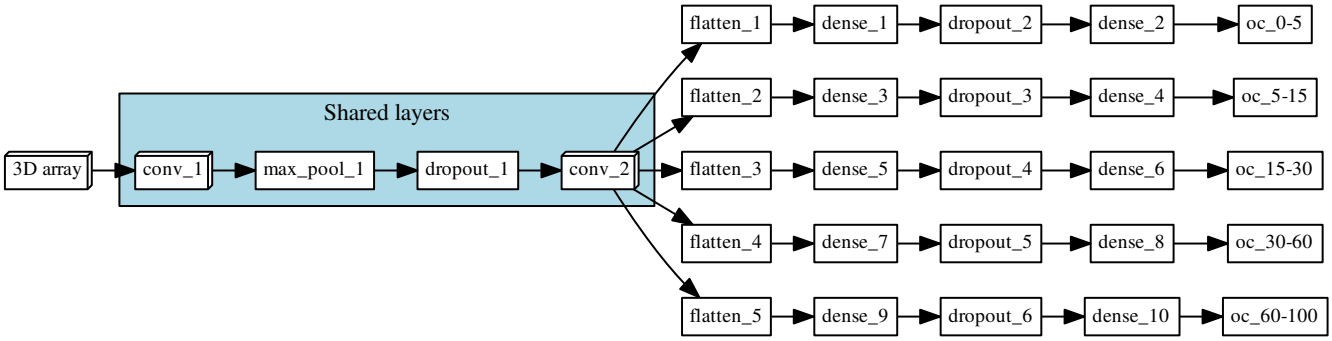


Figure 3. Architecture of the multi-task network. "Shared layers" represent the layers shared by all the depth ranges. Each branch, one per depth range, first flattens the information to a 1D array, followed by a series of 2 fully-connected layer and a fully-connected layer of size=1, which corresponds to the final prediction.

4.4 Inputs

As explained in Section 2, our method uses a window around a soil observation which encloses a group of pixels instead of the single pixel that coincides with the observation. Most likely, the extent or size of that window will affect the model performance. To assess this effect, we compared the results of different models trained with a window size of 3, 5, 7, 9, 15, 21 and 29 pixels.

As the vicinity size increases, so does the number of parameters of the network (considering a fixed network architecture) and the risk of overfitting. To minimise overfitting, we modified the architecture of the network depending on the vicinity size (Table 2).

Table 2. List of modifications made to the base network architecture for specific input window sizes.

Window size	Changes
15x15	• Extra Max-Pooling(2x2) after last Convolutional layer
21x21	• Extra Max-Pooling(2x2) after last Convolutional layer • Extra Convolutional(3x3, 16 filters)
29x29	• Extra Max-Pooling(2x2) after last Convolutional layer • Extra Convolutional(3x3, 64 filters) • Droputs changed to 0.5

4.5 Training & Validation

First, 10% (n = 49) of the total dataset was randomly selected and used as a test set. The remaining 90% of the samples (n = 436) were augmented (see Section 4.2) obtaining a total of 1,744 samples. Following the data augmentation, we performed a bootstrapping routine (Efron and Tibshirani, 1993) with 100 repetitions, where the training set is obtained by sampling with replacement, generating a set of size 1,744. The samples which were not selected, about 1/3 (one-third), correspond to the out-of-bag validation set.

As a control, we compared our results with a previous study by Padarian et al. (2017) where we used a Cubist regression tree model (Quinlan et al., 1992) to predict soil organic carbon (SOC) at a national extent. Cubist has been used in many other DSM studies due to its interpretability and robustness. In that study, we used the same set of soil observations and covariates described in Section 4.1.

4.6 Uncertainty analysis

In this work (and in Padarian et al. (2017)), the uncertainty is represented as the 90% prediction interval derived from the 100 bootstrap iterations. To estimate the upper and lower prediction interval limits, we used the formula:

$$PIL = \bar{x} \pm 1.645\sqrt{\sigma^2 + MSE} \tag{2}$$

where $\bar{x}$ and σ^2 are the mean and variance of the 100 iterations, and MSE is the mean square error of the 100 fitted models.

4.7 Implementation

The CNN was implemented in Python (v3.6.2; Python Software Foundation, 2017) using Keras (v2.1.2; Chollet et al., 2015) and Tensorflow (v1.4.1; Abadi et al., 2015) backend. Computing was done using the University of Sydney's Artemis high performance computing facility.

5 Results and discussion

5.1 Data augmentation

To generalise and improve the CNN model, we created new data using only information from the training data by rotating the original image input. Data augmentation was effective at reducing model error and variability (Fig. 4). It was possible to observe a decrease of the error, by 10.56, 10.56, 11.25, 14.51, and 24.77% for 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively. The results are in accordance with image classification studies which generally showed that data augmentation increased the accuracy of classification tasks (Perez and Wang, 2017). It is hypothesised that by increasing the amount of training data, we can reduce overfitting of CNN models.

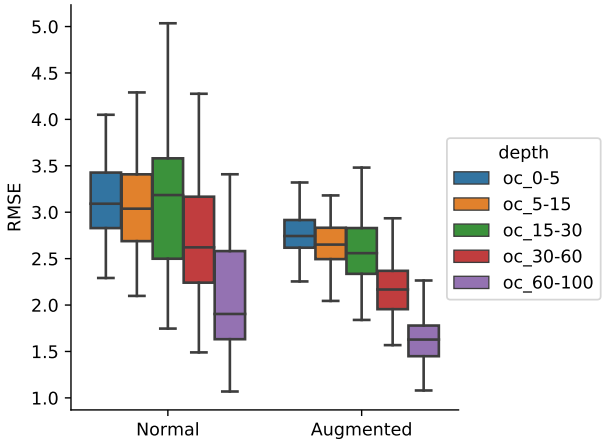


Figure 4. Effect of using data augmentation as a pre-treatment on a 7x7 pixels array.

In terms of the data spatial autocorrelation, we need to consider that after augmenting the data we have 4 samples in the same locations with exactly the same SOC content, therefore assuming that there is no variance when distance=0. That is theoretically true if we consider that the distance is exactly equal to 0. In practice, when calculating the semivariogram, the semivariance value of the first bin will be lower, but that does not significantly affect the final model.

5.2 Vicinity size

To incorporate contextual information for DSM prediction, we represent the input as image. The image is represented as observation in the centre, with surrounding pixels in a square format. The size of the neighbourhood window (vicinity) has a significant effect on the prediction error (Fig. 5). There is not significant difference when using a vicinity size of 3, 5, 7 or 9 pixels, but sizes above 9 pixels showed an increase in the error. Is possible to observe a lower error in the test dataset, compared with the training and validation, due to the slight differences in the dataset distributions (Fig. 6). Since the SOC distribution is right-skewed, the random sampling used to generate the training dataset does not completely recreate the original distribution, excluding samples with very high SOC values. This should not significantly affect the conclusions given that the error for the samples with hight SOC values is accounted during the bootstrapping routine and reflected in the training and validation curves of Fig. 5. In this example, for a country-scale mapping of SOC at 100 m grid size, information from 150 to 450 m radius is useful. A similar influence distance was obtained by Jian-Bing et al. (2006) and Sun et al. (2003) whom reported a medium-scale spatial correlation range for SOC in China of around 300 m and 550 m, respectively; Rossi et al. (2009) and the 190 m reported for a Coastal forest in Tanzania; and Don et al. (2007) of around 200 m in two grassland sites in Germany. A similar spatial correlation range was reported for croplands in an review by Paterson et al. (2018), where, based on 41 variograms, the authors estimated an average spatial correlation range of around 400 m.

As described in Section 4.3, we slightly modified the architecture of the network as input window size increased, in order to minimise the risk of overfitting and isolate the effect of the vicinity size. As we increase the vicinity size, we give the model a

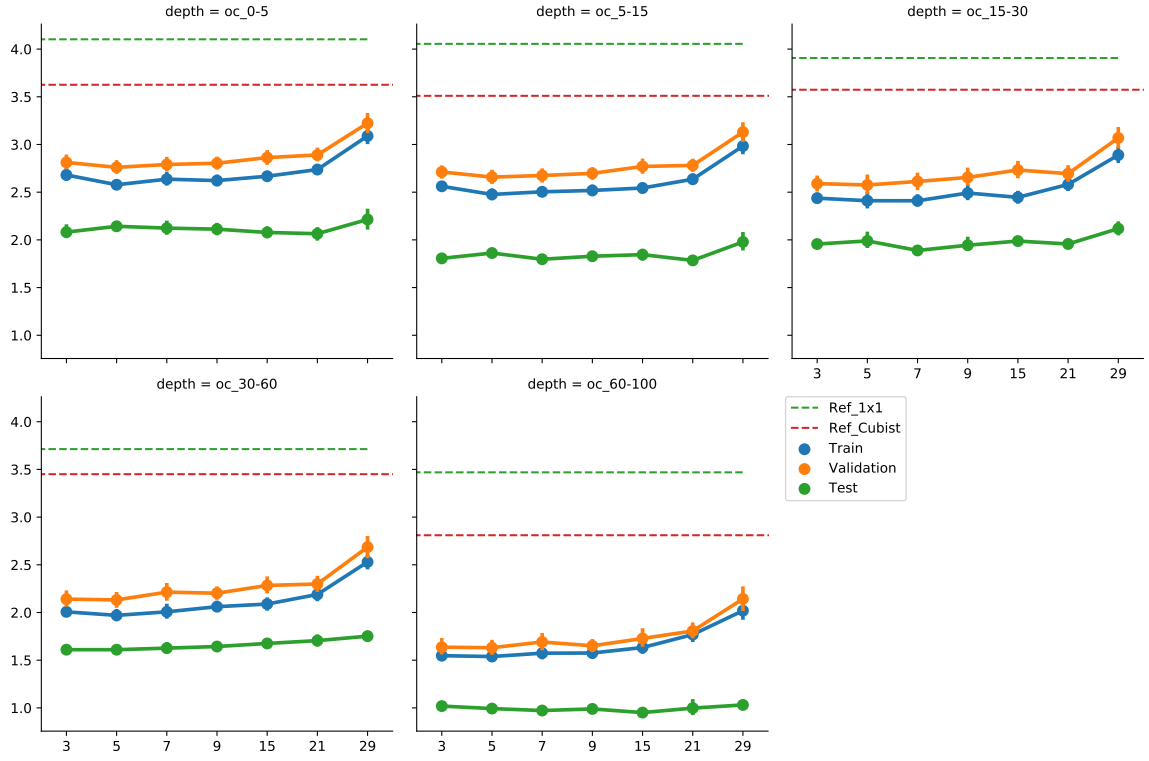


Figure 5. Effect of vicinity size on prediction error, by depth range. Ref_1x1 corresponds to a fully connected neural network without any surrounding pixels. Ref_Cubist corresponds to the Cubist models used by Padarian et al. (2017).

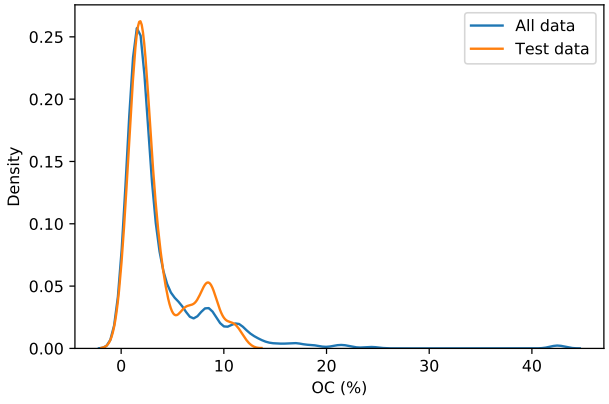


Figure 6. Distribution of the original dataset and the test dataset. The random sampling excludes some observation with high SOC values.

broader spatial context. Our results show that just a small amount of extra context provides enough information to improve the model predictions, and a larger amount of neighbouring information acts as noise, impairing the generalisation of the model. Since we used the relatively large resolution of 100 m, it is hard to tell specifically what is the minimum amount of context

needed to improve SOC predictions. We believe that using higher resolutions ($<$ 10m) could produce more insights about this matter.

Soil forming factors interact in complex ways and affect soil properties with different strength. At local scale, a broader context (e.i.: larger vicinity size) does not necessarily provides extra information to the model, for instance when one of the factors is relatively homogeneous. The extra information could be even detrimental if the vicinity size is well beyond the area of influence of a factor, which is what probably happened when we increased the vicinity size above 9 pixels (radius$\approx$450m). Representing this complexity in numerical terms would imply varying the size of the input array, such as each forming factor has a different vicinity size, most likely also varying depending on the spatial location of the soil observation (e.g.: smaller vicinity for homogeneous areas, larger for heterogeneous areas). This is technically possible but considerably increases the complexity of the modelling.

5.3 Comparison with other methods

We compared our approach with the Cubist model used in our previous study (Padarian et al., 2017), where we did not use any contextual information. We observed a significant decrease in the error (Fig. 5) by 23.0, 23.8, 26.9, 35.8, 39.8% for the 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively. Most current DSM studies rely on punctual observations without contextual information and, given the improvements shown by our approach, we believe there is a big potential for CNNs to be used in operational DSM.

To compare our results with a method that uses contextual information, we ran a test using wavelet decomposition as per Mendonça-Santos et al. (2006). In addition to the five covariates, we used their approximation coefficients from the first, second and third levels of a Haar decomposition (Chui, 2016; Haar, 1910). The results including wavelet decomposed variables were similar to ones obtained with the Cubist model. The CNN approach reduced the error by 24.8, 24.7, 28.5, 28.6, 23.5 for the 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively. Mendonça-Santos et al. (2006) reported an average improvement of 1% for the prediction of clay content. In our case the wavelet decomposition reduced the error of SOC content by 5.1%, in average, compared with Cubist but the reduction was only observed in depth (2.4, 1.2, 2.3, -10.1, -21.4% error change for the 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively), where SOC content is low, hence reducing the effect in applications such as carbon accounting. Our approach showed greater error reductions and through the whole profile.

5.4 Prediction of deeper soil layers

Our approach uses a multi-task CNN to predict multiple depths simultaneously in order to produce a synergistic effect. Compared with predicting each depth range in isolation by training a network with the same structure (Section 4.3) but with only one output, our approach reduced the error by 1.5, 6.7, 6.6, 8.9, 13.0% for the 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively. In this case, the reduction was modest and we believe the effect can be greater when more soil observations are available.

In DSM, there are two main approaches to deal with the vertical variation of a soil attribute: 2.5D and 3D modelling. In the first one, an independent model is fitted for each depth range. The latter explicitly incorporates depth in order to obtain a single model for the whole profile. Interestingly, both approaches show a decrease in the variance explained by the model as the prediction depth increases. In a 3D mapping of SOC for a 125 km^2 region in the Netherlands, Kempen et al. (2011) presented R^2 values of 0.75, 0.23 and 0.09 for the 0–30, 30–60, and 60–90cm depth ranges, respectively. Our previous study (Padarian et al., 2017) the 2.5D mapping showed R^2 values of 0.39, 0.39, 0.27, 0.19, and 0.17 for the 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively. Similar studies show the same trend (Akpa et al., 2016; Mulder et al., 2016; Adhikari et al., 2014), independent of the models used or the soil attribute predicted. This is expected as the information used as covariates usually represents surface conditions. Our multi-task network presented the opposite trend (Fig. 7), showing an increase of the explained variance as the prediction depth increases.

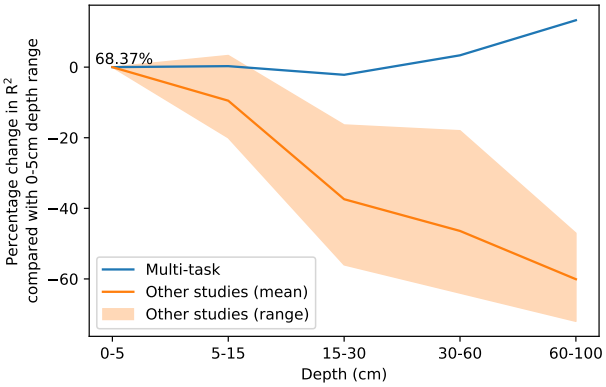


Figure 7. Percentage change in model R^2 in function of depth. The Multi-task model corresponds to a CNN trained using a 7x7 pixels vicinity. Data for "Other studies" correspond to validation statistics from Padarian et al. (2017); Akpa et al. (2016); Mulder et al. (2016) and Adhikari et al. (2014)

The prediction of the adjacent layers served as guidance, producing a synergistic effect. A soil attribute through a profile usually has a predictable behaviour (unless there are lithological discontinuities), which has been described by many authors in the form of depth functions (Kempen et al., 2011; Nakane, 1976; Russell and Moore, 1968). A CNN is capable of generating an internal representation of the vertical distribution of the target attribute, which resembles the observed pattern (Fig. 8).

5.5 Visual evaluation of maps

Visually, the maps generated with the Cubist tree model and our multi-task CNN showed differences (Fig. 9). In an example for an area in southern Chile (around 72.57º S), the map generated with the Cubist model (Fig. 9a) shows more details related with the topography, but also presents some artefacts due to the sharp limits generated by the tree rules. On the other hand, the map generated with the multi-task CNN using a 7x7 window (Fig. 9b) shows a smoothing effect, an expected behaviour consequence of using neighbour pixels.

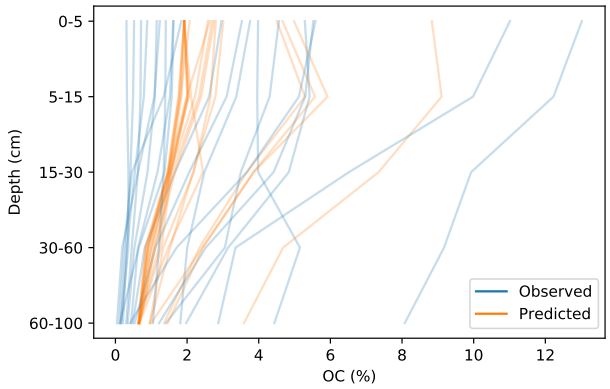


Figure 8. Vertical SOC distribution for 20 randomly selected profiles. Predictions correspond to the multi-task CNN.

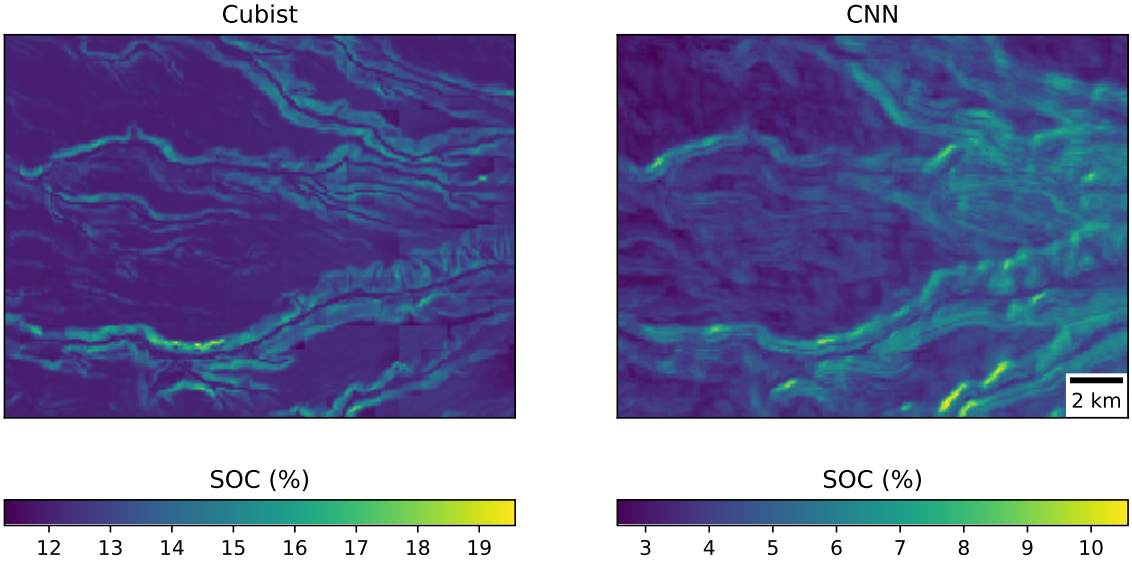


Figure 9. Detailed view of (left panel) map generated by a Cubist model (Padarian et al., 2017) and (right panel) model generated by our multi-task CNN showing the smoothing effect of the CNN.

5.6 Uncertainty

A recommended DSM practice is to present a map of a predicted attribute along its associated uncertainty (Arrouays et al., 2014). ~~In this work (and in Padarian et al. (2017)), the uncertainty is represented as the width of the 90% prediction interval (prediction interval width, PIW) derived from the 100 bootstrap iterations.~~ Our multi-task CNN ~~considerably~~ significantly reduced the prediction ~~uncertainty (Fig 10~~interval width (PIW, Table 3) compared with the Cubist model. ~~The~~ In average, we observed a reduction of 13.1 and 13.8% for the CNN model generated without and with data augmentation pre-treatment~~(~~,

respectively, for the first three depth intervals. Our multi-task CNN model showed a slightly lower prediction interval coverage, but all wider than the proposed 90% coverage.

Table 3. Median prediction interval width (PIW, SOC %) and proportion of observations that fell within the 90% prediction interval (PICP) estimated at the test dataset locations. For the Cubist model, values were extracted from the final maps. For the CNN models, the values correspond to the mean of the 100 bootstrap iterations.

		Cubist	Not augmented	Augmented
0-5 cm	PICP	0.96	0.96	0.94
	PIW	7.96	7.20	7.25
5-15 cm	PICP	0.97	0.96	0.92
	PIW	7.69	6.15	6.06
15-30 cm	PICP	0.97	0.96	0.96
	PIW	7.16	6.47	6.35

In terms of the spatial patterns of the uncertainty (Fig 10~~, left panel)showed reductions mostly in the range of 50–80%, with greater reductions in higher~~), the greater reductions of the PIW where observed in elevated areas of the ~~landscape. By using data augmentation the uncertainty was reduced further (Fig 10, right panel), shifting the distribution to the range of 70–90%~~Andes, followed by the central valleys. A slight increase, in the order of 6-8%, was observed in the western coastal ranges. The reduction of the PIW in the Andes is most most likely due to a more reserved extrapolation by the CNN models compared with Cubist. It is worth noting that the central valleys is were most of the agricultural lands are located and the uncertainty reduction observed in these areas could have important implications.

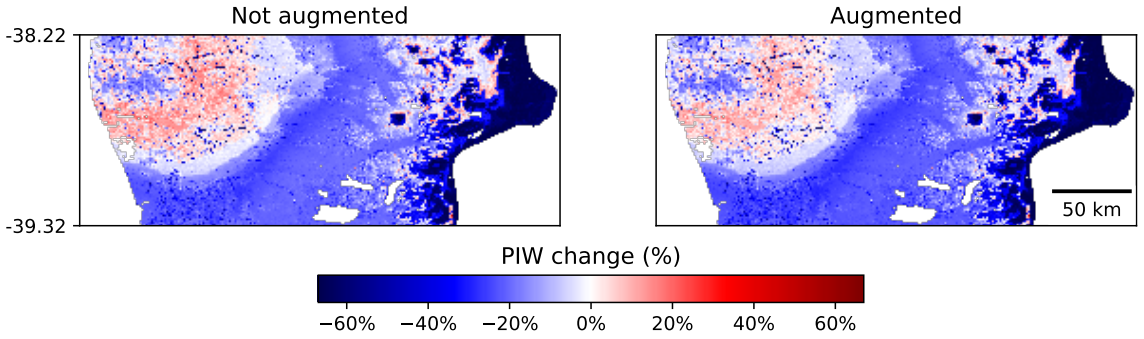


Figure 10. Percentage change on the prediction interval width using as a reference a Cubist model (Padarian et al., 2017) with and without data augmentation pre-treatment (right and left panels, respectively).

6 Conclusions

Incorporating contextual information into DSM models is an important aspect that deserves more attention. Since a soil surveyor will look at the surrounding landscape to make a prediction of soil type, DSM models should also incorporate information surrounding an observation. We demonstrated the use of a convolutional neural network as an efficient, effective, and accurate method to achieve this goal. In particular we introduce a deep learning model for DSM which has the following innovative features:

- The representation of input as an image, which takes into account information surrounding a point observation. CNN is able to recognise contextual information, and extract multi-scale information automatically, which circumvents the need to pre-process the data in the form of spatial filtering or multi-scale analysis.
- The use of data augmentation as a general representation of soil in the landscape, which can reduce overfitting and improve model accuracy.
- The ability to predict different soil depth simultaneously in a model, and thus take into account the depth correlation of soil properties and attributes. In our example, prediction of soil properties at deeper layer, common problem in DSM studies, improved significantly.

The resulting prediction has less uncertainty, further, the use of this data structure with CNN seems to eliminate artefacts generally found in DSM products due to incompatible scale of covariates and sharp discontinuities due to tree models.

A CNN can handle large number of covariates and has advantages over other machine learning algorithms used in DSM, such as random forests, and Cubist regression tree because its architecture is flexible, and explicitly takes spatial information of covariates around observations. While there are attempts to include information surrounding an observation as covariates in a random forest model, those inputs still do not have spatial relationship. CNN does not require pre-processing such as wavelet transformation, rather such function is built in the model. There are other features such as handling missing values via data imputation (Duan et al., 2016) which can be readily added in the network.

The example presented in this paper is for a country-wide modelling at 100 m resolution, and we need to further test such approach in the regional to landscape mapping. The CNN model would be highly suitable for mapping soil class. In addition, the presented model can be used for other environmental mapping.

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

Acknowledgements. This research was supported by Sydney Informatics Hub, funded by the University of Sydney.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/, software available from tensorflow.org, 2015.

Adhikari, K., Hartemink, A. E., Minasny, B., Kheir, R. B., Greve, M. B., and Greve, M. H.: Digital mapping of soil organic carbon contents and stocks in Denmark, PloS one, 9, e105 519, 2014.

Akpa, S. I., Odeh, I. O., Bishop, T. F., Hartemink, A. E., and Amapu, I. Y.: Total soil organic carbon and carbon sequestration potential in Nigeria, Geoderma, 271, 202–215, 2016.

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

Angelini, M. E. and Heuvelink, G. B.: Including spatial correlation in structural equation modelling of soil properties, Spatial Statistics, 25, 35–51, 2018.

Arrouays, D., McBratney, A., Minasny, B., Hempel, J., Heuvelink, G., MacMillan, R., Hartemink, A., Lagacherie, P., and McKenzie, N.: The GlobalSoilMap project specifications, GlobalSoilMap: Basis of the global spatial soil information system, 9, 2014.

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., and Rossel, R. V.: Multiscale contextual spatial modelling with the Gaussian scale space, Geoderma, 310, 128–137, 2018.

Biswas, A. and Si, B. C.: Revealing the controls of soil water storage at different scales in a hummocky landscape, Soil Science Society of America Journal, 75, 1295–1306, 2011.

Chen, S., Martin, M. P., Saby, N. P., Walter, C., Angers, D. A., and Arrouays, D.: Fine resolution map of top-and subsoil carbon sequestration potential in France, Science of The Total Environment, 630, 389–400, 2018.

Chollet, F. et al.: Keras, https://github.com/fchollet/keras, 2015.

Chui, C. K.: An introduction to wavelets, Elsevier, 2016.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J.: System for automated geoscientific analyses (SAGA) v. 2.1.4, Geoscientific Model Development, 8, 1991–2007, 2015.

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.

Deumlich, D., Schmidt, R., and Sommer, M.: A multiscale soil–landform relationship in the glacial-drift area based on digital terrain analysis and soil attributes, Journal of Plant Nutrition and Soil Science, 173, 843–851, 2010.

Dokuchaev, V. V.: Russian Chernozem. Selected works of V.V. Dokuchaev. v. 1, Israel Program for Scientific Translations, . Jerusalem (translated in 1967), 1883.

Don, A., Schumacher, J., Scherer-Lorenzen, M., Scholten, T., and Schulze, E.-D.: Spatial and vertical variation of soil carbon at two grassland sites—implications for measuring soil carbon stocks, Geoderma, 141, 272–282, 2007.

Duan, Y., Lv, Y., Liu, Y.-L., and Wang, F.-Y.: An efficient realization of deep learning for traffic data imputation, Transportation research part C: emerging technologies, 72, 168–181, 2016.

Efron, B. and Tibshirani, R. J.: An introduction to the bootstrap, vol. 57, CRC press, New York, 1993.

Haar, A.: Zur theorie der orthogonalen funktionensysteme, Mathematische Annalen, 69, 331–371, 1910.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A.: Very high resolution interpolated climate surfaces for global land areas, International journal of climatology, 25, 1965–1978, 2005.

Jenny, H.: Factors of soil formation: a system of quantitative pedology New York, Macgraw Hill, 1941.

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.

Kamilaris, A. and Prenafeta-Boldú, F. X.: Deep learning in agriculture: A survey, Computers and Electronics in Agriculture, 147, 70–90, 2018.

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.

Keskin, H. and Grunwald, S.: Regression kriging as a workhorse in the digital soil mapper's toolbox, Geoderma, 326, 22–41, 2018.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105, 2012.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D.: Handwritten digit recognition with a back-propagation network, in: Advances in neural information processing systems, pp. 396–404, 1990.

LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks, 3361, 1995, 1995.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, 2015.

Lee, H. and Kwon, H.: Going deeper with contextual CNN for hyperspectral image classification, IEEE Transactions on Image Processing, 26, 4843–4855, 2017.

Lehner, B., Verdin, K., and Jarvis, A.: New global hydrography derived from spaceborne elevation data, EOS, Transactions American Geophysical Union, 89, 93–94, 2008.

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

Mendonça-Santos, M., McBratney, A., and Minasny, B.: Soil prediction with spatially decomposed environmental factors, Developments in Soil Science, 31, 269–278, 2006.

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

Mulder, V., Lacoste, M., de Forges, A. R., and Arrouays, D.: GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth, Science of The Total Environment, 573, 1352–1369, 2016.

Nakane, K.: An empirical formulation of the vertical distribution of carbon concentration in forest soils, Japanese Journal of Ecology, 26, 171–174, 1976.

Nitish, S., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting., Journal of machine learning research, 15, 1929–1958, 2014.

Padarian, J., Minasny, B., and McBratney, A.: Chile and the Chilean soil grid: a contribution to GlobalSoilMap, Geoderma Regional, 9, 17–28, 2017.

Padarian, J., Minasny, B., and McBratney, A.: Using deep learning to predict soil properties from regional spectral data, Geoderma Regional, 16, e00 198, 2019.

Paterson, S., McBratney, A. B., Minasny, B., and Pringle, M. J.: Variograms of Soil Properties for Agricultural and Environmental Applications, in: Pedometrics, pp. 623–667, Springer, 2018.

Perez, L. and Wang, J.: The effectiveness of data augmentation in image classification using deep learning, arXiv preprint arXiv:1712.04621, 2017.

Poggio, L. and Gimona, A.: Assimilation of optical and radar remote sensing data in 3D mapping of soil properties over large areas, Science of the Total Environment, 579, 1094–1110, 2017.

Python Software Foundation: Python Language Reference, Python Software Foundation, https://www.python.org, 2017.

Quinlan, J. R. et al.: Learning with continuous classes, in: 5th Australian joint conference on artificial intelligence, vol. 92, pp. 343–348, Singapore, 1992.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V.: Massively multitask networks for drug discovery, arXiv preprint arXiv:1502.02072, 2015.

Rossi, J., Govaerts, A., De Vos, B., Verbist, B., Vervoort, A., Poesen, J., Muys, B., and Deckers, J.: Spatial structures of soil organic carbon in tropical forests—a case study of Southeastern Tanzania, Catena, 77, 19–27, 2009.

Ruder, S.: An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098, 2017.

Russell, J. and Moore, A.: Comparison of different depth weightings in the numerical analysis of anisotropic soil profile data, Int Soc Soil Sci Trans, 1968.

Simard, P. Y., Steinkraus, D., Platt, J. C., et al.: Best practices for convolutional neural networks applied to visual document analysis., in: ICDAR, vol. 3, pp. 958–962, 2003.

Somarathna, P., Minasny, B., and Malone, B. P.: More Data or a Better Model? Figuring Out What Matters Most for the Spatial Prediction of Soil Carbon, Soil Science Society of America Journal, 2017.

Song, X., Zhang, G., Liu, F., Li, D., Zhao, Y., and Yang, J.: Modeling spatio-temporal distribution of soil moisture by deep learning-based cellular automata model, Journal of Arid Land, 8, 734–748, 2016.

Sun, B., Zhou, S., and Zhao, Q.: Evaluation of spatial and temporal changes of soil quality based on geostatistical analysis in the hill region of subtropical China, Geoderma, 115, 85–99, 2003.

Sun, X.-L., Wang, H.-L., Zhao, Y.-G., Zhang, C., and Zhang, G.-L.: Digital soil mapping based on wavelet decomposed components of environmental covariates, Geoderma, 303, 118–132, 2017.

Vo, N. N. and Hays, J.: Localizing and orienting street views using overhead imagery, in: European Conference on Computer Vision, pp. 494–509, Springer, 2016.