Machine learning and soil sciences: A review aided by machine learning tools

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@sydney.edu.au)

Abstract. The application of machine learning (ML) techniques in various fields of science has increased rapidly, especially in the last ten years. The increasing availability of soil data that can be efficiently acquired remotely and proximally, and freely available open-source algorithms, have led to an accelerated adoption of ML techniques to analyse soil data. Given the large number of publications, it is an impossible task to manually review all papers on the application of ML in soil science without narrowing down a narrative of ML application in a specific research question. This paper aims to provide a comprehensive review of the application of ML techniques in soil science aided by a ML algorithm (Latent Dirichlet Allocation) to find patterns in a large collection of text corpus. The objective is to gain insight into publications of ML applications in soil science and to discuss the research gaps in this topic. We found that: a) there is an increasing usage of ML methods in soil sciences, mostly concentrated in developed countries, b) the reviewed publication can be grouped into 12 topics, namely remote sensing, soil organic carbon, water, contamination, methods (ensembles), erosion and parent material, methods (NN, SVM), spectroscopy, modelling (classes), crops, physical and modelling (continuous), c) advanced ML methods usually perform better than simpler approaches thanks to their capability to capture non-linear relationships. From these findings, we found research gaps, in particular: about the precautions that should be taken (parsimony) to avoid overfitting, and that the interpretability of the ML models is an important aspect to consider when applying advanced ML methods in order to improve our knowledge and understanding of soil. We foresee that a large number of studies will focus on the latter topic.

1 Introduction

The application of machine learning (ML) techniques in various fields of science has increased rapidly, especially in the last ten years. Soil science research, in particular, Pedometrics, has used statistical models to "learn" or understand from data how soil is distributed in space and time (McBratney et al., 2019). The increasing availability of soil data that can be efficiently acquired remotely and proximally, and freely available open-source algorithms, have led to an accelerated adoption of ML techniques to analyse soil data. Several well-known ML applications in soils science include the prediction of soil types and properties via

digital soil mapping (DSM) or pedotransfer functions, and analysis of infrared spectral data to infer soil properties. Machine learning analysis of soil data is also used to draw conclusions on the controls of the distribution of the soil.

The definition or what constitutes ML is still contentious or sometimes mistaken. In this work, instead of adding a new argument to differentiate ML from statistical science, we will focus on the view of Jordan and Mitchell (2015) where ML is "lying at the intersection of computer science and statistics". With respect to artificial intelligence (AI), sometimes we have seen the terms ML and AI used interchangeably. This is understandable confusion since ML is a subset of AI, but not everything related to AI falls in the ML category (e.g. expert systems).

There are concerns that ML application ignores soils science knowledge (Rossiter, 2018), and that the results could be misleading and wrong. Nevertheless, many would find that ML methods can help in the scientific process (Mjolsness and DeCoste, 2001; Rudin and Wagstaff, 2014): observations, empirical and theory-based models development, and simulations of soil processes (Rossiter, 2018). For example, exploration of high-dimensional infrared spectral data helps in understanding the horizonation designation in a soil profile (Fajardo et al., 2016). The process of modelling and validation can be used to formulate a model to explain soil distribution (Brungard et al., 2015). Modelling via ML can also be used to improve our understanding of the causes of soil variation. Results from ML models can inform which environmental variables that control soil distribution. New relationships revealed by ML analysis can help to stimulate ideas, generate hypotheses, and formulate future questions for research (Ma et al., 2019).

This paper aims to provide a comprehensive review of the application of ML techniques in soil science. A quick google scholar search of "soil" and "machine learning" resulted in more than 70,000 items, with 16,000 items published in 2018. While we can narrow down a narrative of ML application in a specific research question, such as the application of ML in yield prediction in precision agriculture (Chlingaryan et al., 2018) or DSM, it is an impossible task to manually review all papers on the application of ML in soil science. One ML technique that has not been applied in soil science is topic modelling, a type of quantitative text mining method. Similar to what ML does to numerical data, topic modelling finds patterns in a large collection of text corpus (Blei et al., 2003; Blei, 2012) and it has been used to study the evolution of various disciplines and topics (Zhou et al., 2006; Sugimoto et al., 2011; Wu et al., 2014).

This paper uses topic modelling to analyse the trend in ML application in soil science. The objective is to gain insight into publications of ML applications in soil science, in particular, we will try to answer the following questions:

- Who is using ML? and is the application of ML as ubiquitous as we think?
- Which ML methods are commonly used and how often have they been used?
- In which areas of soil sciences we use ML? and how are they clustered and related?
- Do advanced ML methods performs significantly better than linear or non-linear statistical approaches?
- Can ML methods simulate soil processes in space and time?
- Can we use ML methods to improve our knowledge and understanding of soil?

Throughout this review, we will refer to models as "simple" or "complex/advanced" trusting in the readers' criteria. To illustrate that gradient between simple and complex, we considered a linear model (LM) with 2 variables as simple compared to a LM with 100 variables; a classification and regression tree (CART) with 2 branches as simple compared to a CART with 100 branches; and finally, a CART with 2 branches as simple compared with a LM with 100 variables. We also hope that is clear for the reader that a model such as a deep () has many parameters, hence is more complex than a CART model.

2 Methods

2.1 Articles selection

In order to identify the primary group of articles, we used the term "soil 'machine learning' " to perform a full-text search in databases from different publishers. We selected the publishers based on a) our institution having access to full-text articles, and b) that they provide text-mining permission. We limited our search to English only literature, without fixing a specific time-frame, and completing the search the 1st of February 2019. After performing a screening for the relevance of the initial 3044 matches, we decided to narrow down the selection to the articles containing the word "soil" in their title, yielding a total of 322 articles. The final journal names and number of articles are shown in Table A1.

2.2 Topic modelling

Topic modelling is a probabilistic ML method that aims to discover and annotate large archives of documents with thematic information Blei (2012). By analysing the words contained in a set of documents, these topic modelling algorithms are capable of identifying common themes. These methods allow processing an arbitrarily large number of articles, which can help to reduce part of the bias introduced by only selecting a manageable subset of documents, or by manually assigning documents to topics.

In order to determine in which areas of soil sciences we use ML, we selected an algorithm commonly used in topic modelling called Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to perform the task of allocating the articles into topics. LDA is a probabilistic model that assumes that a number of topics exist in a document collection and each topic is represented by a distribution of words. Each document is represented by a distribution over topics, and each word is a sample over each topic's vocabulary (Fig. 1). For more details about the LDA, we refer the reader to Blei (2012).

Before modelling the topics, we pre-processed the documents in order to reduce the noise of the unstructured texts. We a) removed stop-words (common words such as "from" and "are"), b) we generated bi- and tri-grams, which are groups of 2 and 3 words which commonly appear together in the text (e.g. "remote sensing", "particle size distribution"), and c) removed extremely uncommon (that appear in less than 5 documents) and common words (that appear in more than 50% of the documents), which do not help to differentiate between topics.

The LDA algorithm is capable of learning different topics to which each document is assigned given the words that constitute it. The first challenge is to find the optimal number of topics, which has to be general enough to capture similarities between

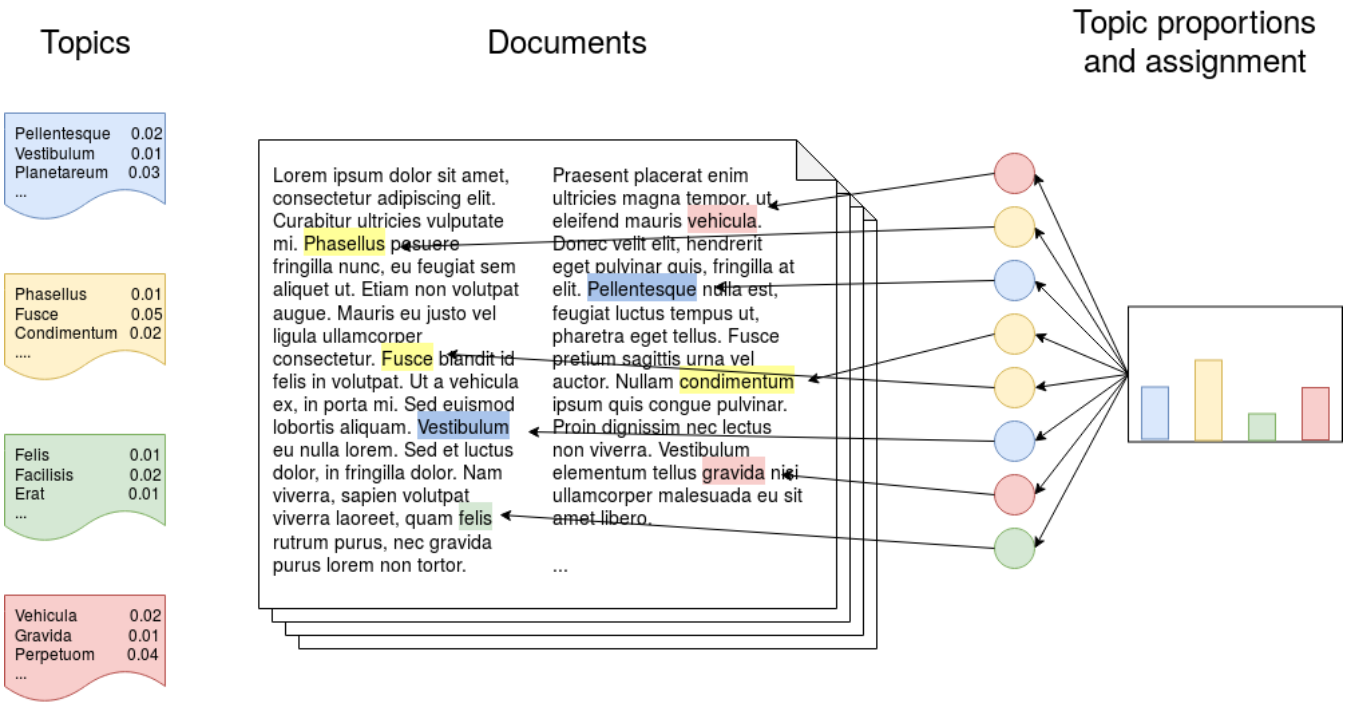

Figure 1. Collection of topics with distribution of words (left), document distribution over topics (histogram, right) and words sampled from the topics' vocabularies (circles). The topics, words and assignment are for illustrative purposes. Adapted from Blei (2012).

articles but with some degree of specificity in order to have a manageable and sensible group of topics. That balance between generality and specificity is key to generate topics that are semantically interpretable by a human (Stevens et al., 2012). One of the measures that is highly correlated with human interpretation of the topics is topic coherence (Stevens et al., 2012). We estimated a coherence measure proposed by Röder et al. (2015) (referenced as CV in their paper) for different models trained with an increasing number of topic, from 2 to 30. CV is an aggregated measure which combines a normalised point-wise mutual information coherence measure, cosine vector similarity and a boolean sliding window of size 110. It ranges from 0 to 1, being 1 the highest coherence. Other parameters of the LDA algorithm such as the threshold of the probability above which a topic considered, and the number of training iterations were set to 0.2 and 1000, respectively, after performing a parameter grid search.

2.3 Text extraction

In order to identify the information required to answer our questions, we used a combination of named-entity recognition and rule-based matching. To extract the MODEL entities, we used a list of modelling methods from the *Outline_of_machine_learning* Wikipedia article in addition to other algorithms that are commonly used in soil sciences and that were not present in the list (e.g. Cubist). After extracting the MODEL entities, we proceeded to extract the abbreviations used to reference those mod-

els. In order to extract the abbreviations, we relied on the commonly seen pattern of writing model names followed by their corresponding abbreviation (e.g. "we used a random forest model (RF)"). By extracting the abbreviations we expected to discriminate between a) models used to generate the results reported in the articles and b) models mentioned to give context to the studies. Extracting abbreviations also allowed us to capture variations of models not present in our original list (e.g. BART for bagged regression trees in Fig. 2).

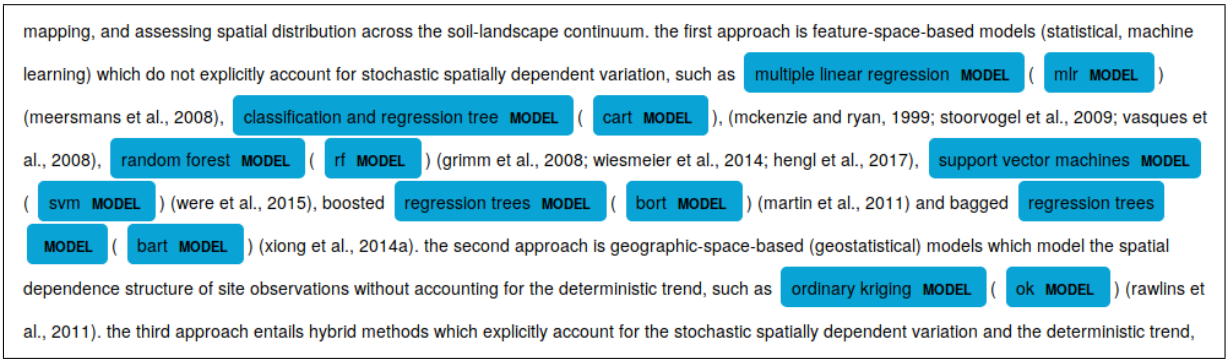

Figure 2. Excerpt from one of the reviewed articles showing named entities recognised as models. Note that the word 'bagged' is not recognised, but the abbreviation 'bart' is.

2.4 Implementation

We performed all our analysis in Python, using the libraries gensim v3.6.0 (Řehůřek and Sojka, 2010) and the MALLET package (McCallum, 2002) for the topic modelling and spacy v2.1.0a6 (Matthew and Honnibal, 2017) for the named entity recognition.

3 Results and discussion

3.1 Who is using machine learning methods?

The first questions related to the current status of the ML literature in soil sciences can be answered after correctly organising all the articles metadata. Regarding the general usage of ML methods, in our review, we observed an expected increment in time in the number of publications using ML to model different aspects of soils (Fig. 3). This increment is most likely due to a combination of increasing computational power and accessibility to high-performance computers, increasing availability of data (e.g. remote sensing) (Jordan and Mitchell, 2015), and the increasing interest in "data science". It is also confounded with the overall increase in the number of publications, which was estimated in 2015 at nearly 2.5 million new publications per year (Ware and Mabe, 2015).

Besides the temporal trend in publishing, we were also interested in how ubiquitous the application of ML methods is. Fig. 4 shows the number of institutions per country ($\log_{10}$) that appeared listed as affiliation in the analysed articles. ML techniques in

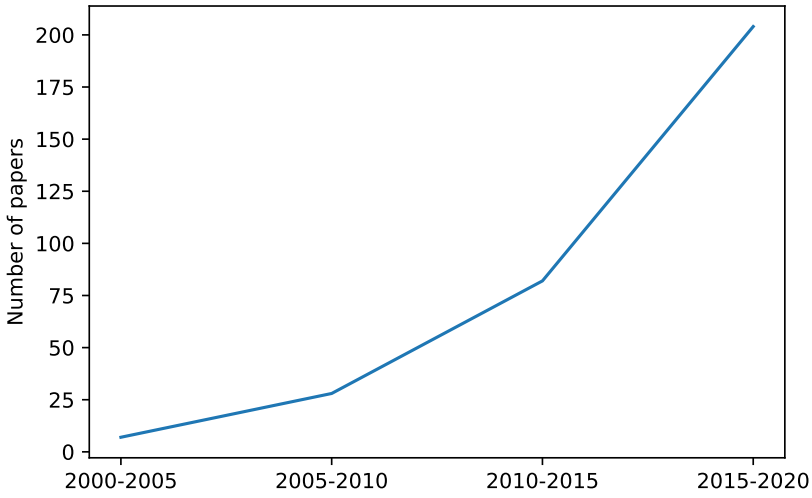

Figure 3. Distribution in time of the articles used in this review.

the context of soil sciences are used in many countries around the world but mostly concentrated in developed countries. This is due to the inseparable relationship between science, technology and development (Sagasti, 1973), which is also related to what is usually called "digital divide" (Rossiter, 2018). Inter-institutional collaboration could be an important aspect of closing this gap (Sonnenwald, 2007). Similar to what is happening in many disciplines (Sonnenwald, 2007), we observed an increase in the number of co-authors per article (Fig. 5), which might be a good sign if we avoid bad practices like "helicopter science" (Minasny and Flantis, 2018).

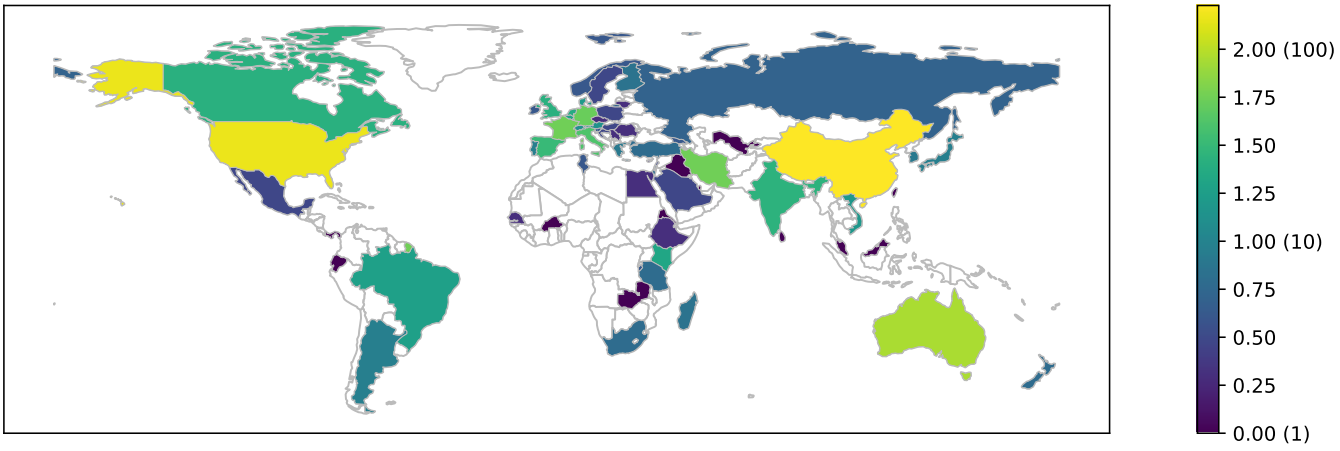

Figure 4. Total number ($\log_{10}$) of institutions per country that participated in articles included in this review. Numbers between brackets are real number of publications. Outlined countries have zero occurrences.

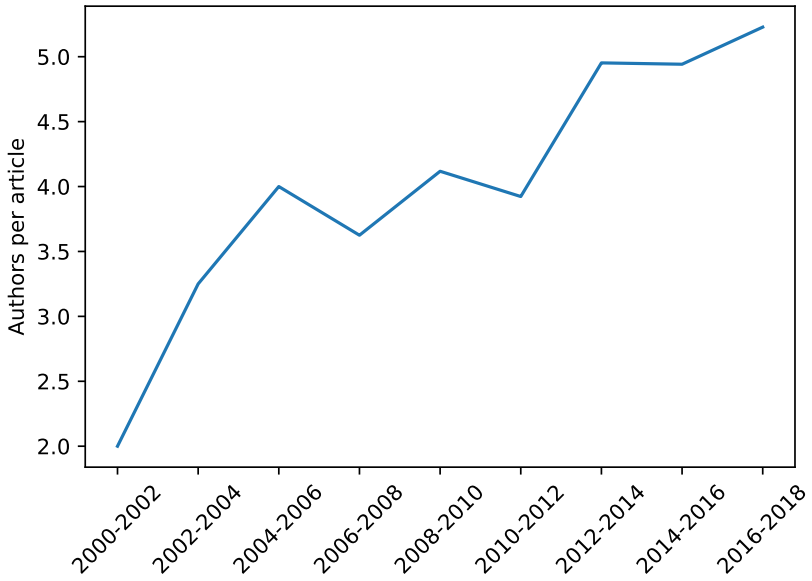

Figure 5. Evolution of the number of authors per publication. Mean values per time-period.

The advance of a discipline is not only measured by the number of publications. Dissemination of knowledge is a key component of research and Open Access (OA) has been recognised as an optimal solution since it is in the best interests of all stakeholders involved in the process (Björk, 2017). In the application of ML in soil sciences, the proportion of OA publications is very low (Fig. 6). This number is in line with the overall OA presence in science (Björk, 2017) but on the opposite side of the general trend in ML where scientists prefer AO (Hutson, 2018).

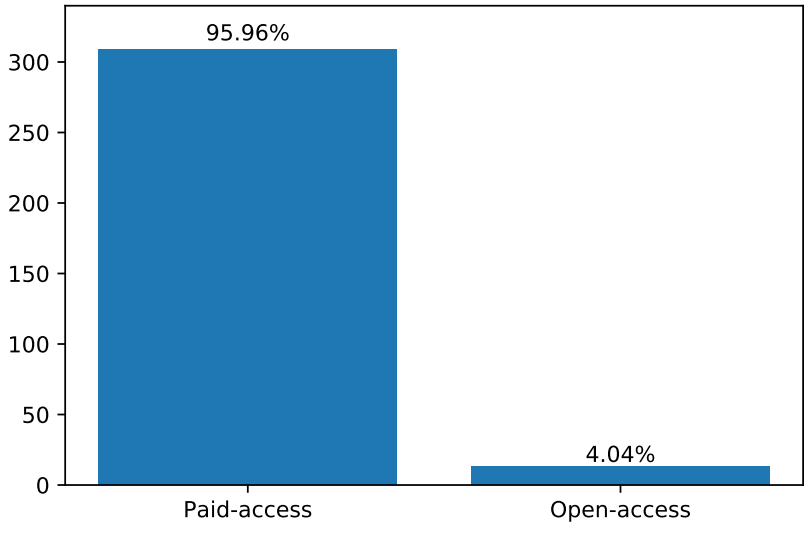

Figure 6. Number of articles published under paid and open access.

3.2 Most used methods

From the huge variety of ML models available, we found over 100 different variants that have been applied in soil sciences. From those, most have been applied experimentally in one or two papers and just a hand-full are consistently used. Fig. 7 depicts the evolution of some selected models. There is an overall increase in the usage of all the models but, proportionally, it is possible to see a decrease in the usage of some models such as support vector machines (SVM), Multivariate Adaptive Regression Spline (MARS) and CART, giving the way to more advanced alternatives such as random forest (RF). The adoption of the later has an accelerated growth and it has been used in a diversity of topics, including mapping and spectroscopy. It is also noticeable the appearance of deep learning, which at the moment has only be used in a few publications related to mapping and spectroscopy.

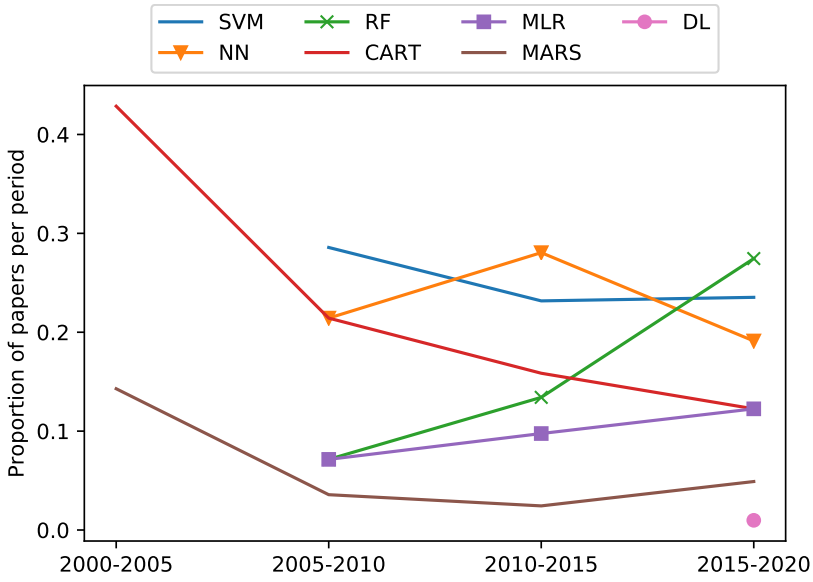

Figure 7. Evolution of model usage in time. SVM: support vector machines; NN: neural networks; RF: random forest; CART: classification and regression trees; MLR: multiple linear regression; MARS: multivariate adaptive regression spline; DL: deep learning

.

3.3 Main topics

As we mentioned in Section 2.2, in order to find the optimal number of topics present in the corpora, we trained models with an increasing number of topics (from 2 to 30) and we plotted the evolution of the CV coherence (Fig. 8). From this curve is possible to select the number of topics that yield the highest coherence, which in this case is 12.

These 12 topics correspond to main soil areas detected by the LDA algorithm where ML is applied. We extracted the most relevant words for each of the 12 topics and we examined the titles of the more relevant papers to identify suitable "topic names". The 12 identified areas were:

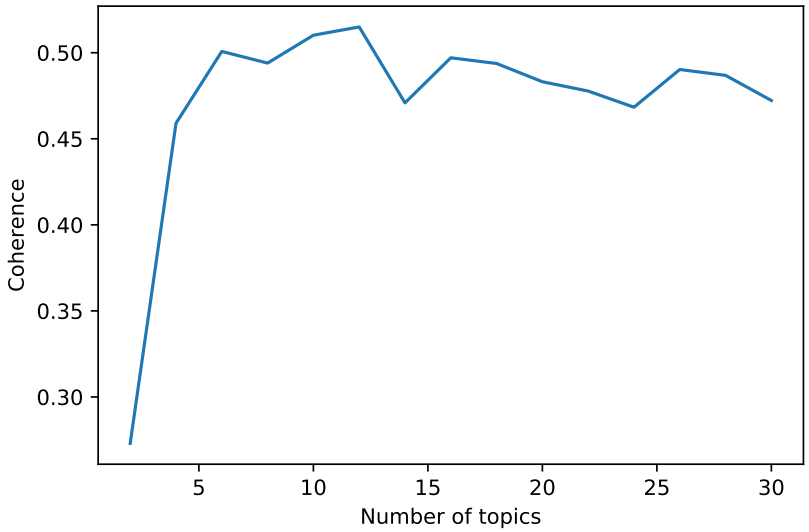

Figure 8. Coherence by number of topics used to train a LDA model.

Remote sensing: Articles heavily based on remote sensing (Grunwald et al., 2015; Xu et al., 2017; Zhang et al., 2018b). Articles related to salinity were also assigned to this group since most of them use remote sensing techniques (Khadim et al., 2019; Zhang et al., 2019).

Soil organic carbon: Articles related to soil organic carbon (SOC) cycles and dynamics, and its relationship with the environment. Carbon stocks in different ecosystems, with particular emphasis in grasslands and topsoil (Rial et al., 2017; Liu et al., 2018; Song et al., 2018; Wang et al., 2018a).

Water: Articles mostly focused on soil water content and its changes over time (Ahmad et al., 2010; Coopersmith et al., 2014; Greifeneder et al., 2018; Han et al., 2018). Other articles in this category are related to soil temperature and CO_2 fluxes (Xing et al., 2018; Oh et al., 2019; Warner et al., 2019; Zeynoddin et al., 2019). All these articles comprise measurements made by "stations".

Contamination: Articles addressing problems related to heavy metals, soil pollution and bio-availability (Costa et al., 2017; Reeves et al., 2018; Wu et al., 2013).

Methods (ensembles): Articles with a focus on model ensembles such as RF (Blanco et al., 2018; Tziachris et al., 2019).

Erosion / parent material: Articles focused on soil formation processes, specifically additions and losses by deposition and erosion, respectively (Geissen et al., 2007; Märker et al., 2011; Martinez et al., 2017). Since soil formation depends on the parent material, articles aiming to characterise it were also included in this category (Kheir et al., 2008; Lacoste et al., 2011).

Methods (NN, SVM): Articles with a focus on methods such asNeural Networks (NN) and SVM (Kovačević et al., 2010; Farfani et al., 2015; Hanna et al., 2007).

Spectroscopy: This topic is related to proximal soil sensing covering different light wavelength sections, from microwave to infrared to gamma (Heggemann et al., 2017; Butler et al., 2018; Xie and Li, 2018).

Modelling (classes): Articles focused on the modelling, especially mapping, of categorical soil properties based on their relationship with environmental covariates (Mansuy et al., 2014; Camera et al., 2017; Dharumarajan et al., 2017; Massawe et al., 2018). In this category is also possible to find articles related to the use of conventional soil maps, especially spatial disaggregation of polygons (Subburayalu et al., 2014; Vincent et al., 2018; Flynn et al., 2019).

Crops: This group of articles focused not merely in soil but in its interaction within the soil-plant continuum. Water and nutrient availability in order to assure crop yields is a key component of this topic (Karandish and Šimůnek, 2016; Ivushkin et al., 2018; Khanal et al., 2018; Leenaars et al., 2018).

Physical: Articles related to the physical properties of soils, including texture and bulk density (Bondi et al., 2018; Naderi-Boldaji et al., 2019), and how they affect aspects of soil such as water retention and flow (Koestel and Jorda, 2014; Gao et al., 2018).

Modelling (continuous): Articles focused on the modelling, especially mapping, of continuous soil properties based on their relationship with environmental covariates, from regional to continental scales (Henderson et al., 2005; Dai et al., 2014; Poggio et al., 2016; Padarian et al., 2019a; Caubet et al., 2019). In this category it is also possible to find articles related to pedotransfer functions (Dobarco et al., 2019).

These topics are not completely independent and they share some commonalities. For instance, Fig. 9 shows an overlap between Topic 12 (Modelling continuous properties) and 9 (Modelling classes) since both are related to mapping using environmental covariables. Both topics are also related to topic 3 (Water) since its articles usually have a spatial component. Something similar occurs between Topic 8 (Spectroscopy) and 1 (Remote sensing) since both are related to spectral data.

Besides the shared features between topics, given that LDA is a probabilistic model, articles also contain features related to more than one topic, i.e. they talk about more than one topic (Fig. 10). For instance, many of the articles related to SOC are also related to soil modelling and mapping (Deng et al., 2018; Wang et al., 2018b; Gomes et al., 2019; Keskin et al., 2019).

3.4 Performance of machine learning models

Our review shows that more advanced modelling techniques usually yield better results compared with simpler approaches. In one of the more extensive comparisons, Sirsat et al. (2018) compared 76 different algorithms, where ensembles of extremely randomised regression trees ranked first when predicting soil fertility indices. Other comparative studies also showed a consistent higher performance of ML methods (NN, SVM, RF) over simpler approaches (Principal Components Regression, Partial

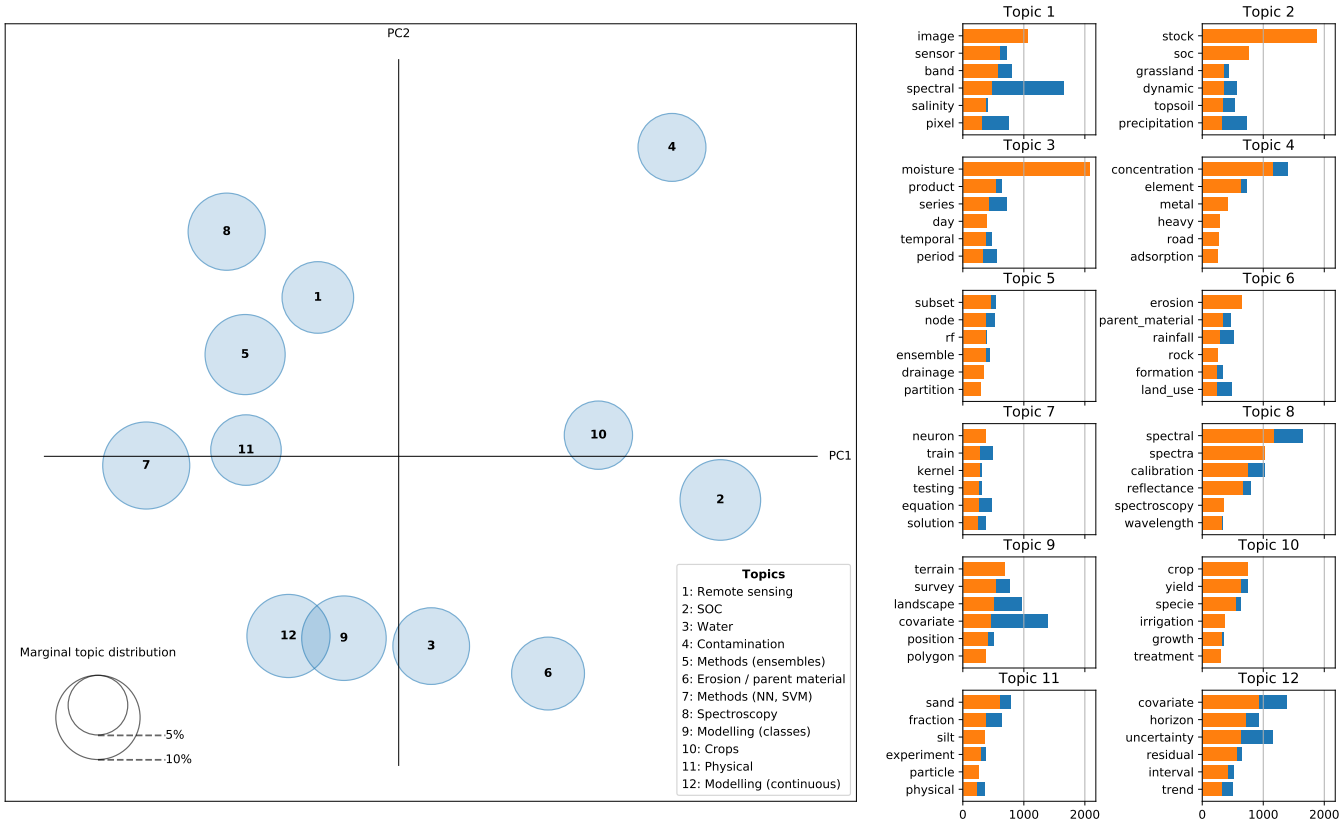

Figure 9. Inter-topic Distance Map. Dimension reduction via Jensen-Shannon Divergence (Lin, 1991) and Principal Coordinate Analysis. Top-6 more relevant words per topic. Complete bars (blue + orange) correspond to overall term frequency and shaded (orange) correspond to term frequency within the selected topic.

Least Squares Regression (PLSR), multiple linear regression (MLR), k-Nearest Neighbours) in applications such as spectroscopy (Viscarra-Rossel and Behrens, 2010; Morellos et al., 2016) and DSM (Brungard et al., 2015; Taghizadeh-Mehrjardi et al., 2015; Camera et al., 2017; Jeong et al., 2017). Most studies mention that the superiority of these algorithms is given by their capability to deal with complex nonlinearities present in the data. Moreover, the better performance of more advanced ML methods is reported in studies related to the prediction of continuous properties and classes.

Regarding the connection between performance and model usage (Section 3.2), we observed that some simpler methods such as MLR, despite their lower performance compared to more advanced models, are very popular. This is expected for statistical models since they have a long tradition in science. On the other hand, we also observed a natural tendency of leaving some model behind despite being used for a long time. For instance, PLSR is very popular and has been used since the 80-90s but, when used in the studies included in this review (mostly published post 2000s), very few studies use it as their main algorithm and, instead, it is used in comparative studies where it is outperformed by more advanced models.

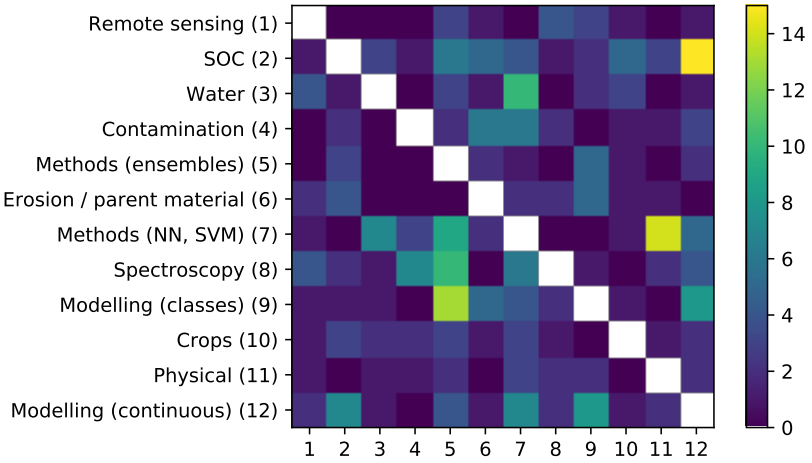

Figure 10. Co-occurrence between the two most likely topics per document. Values correspond to number of papers.

It's worth noting that the final performance is not solely dependent on the selected modelling method. Advanced methods like NNs have a big number of parameters to fit, especially in the context of deep learning. In order to correctly fit those parameters, from a computational and statistically point of view, the size of the dataset is an essential factor (Jordan and Mitchell, 2015). Padarian et al. (2019b) show that a deep CNN trained using a large dataset (around 20,000 soil samples) outperformed methods such as PLS and Cubist when predicting soil properties from spectral data. Using the same method but training on a significantly smaller dataset (390 soil samples), the CNN yielded the worst results.

There is not a clear rule on how big a dataset should be, especially because it certainly depends on the complexity of the underlying problem, but the relationship between dataset size and performance has been shown in many studies, using what is usually known as "learning curves" (Catlett, 1991; Shavlik et al., 1991; Cortes et al., 1994; Perlich et al., 2003; Somarathna et al., 2017). During our review, we observed that the dataset size varied greatly depending on the ML methods (Fig. 11).

Considering that ML models could generate a similar solution to a linear model (e.g. a single-rule tree), it should not be a problem to use any method for any dataset size. However, the main difficulty is that training a complex ML model is not a trivial task, especially to avoid overfitting and to obtain a good generalisation, which becomes challenging in the presence of small datasets (i.e. training and test datasets). Even if a researcher can overcome the training process, it is probable that a simpler model can yield similar results.

3.5 Space-time modelling

Compared with the spatial component of soil variation, which is prominent in the topics found using the LDA algorithm (Section 3.3), the number of studies that address the spatio-temporal dynamics of soil properties using ML methods is still limited. Our findings agree with the review by Grunwald (2009) who characterised studies covering the years 2007 and 2008.

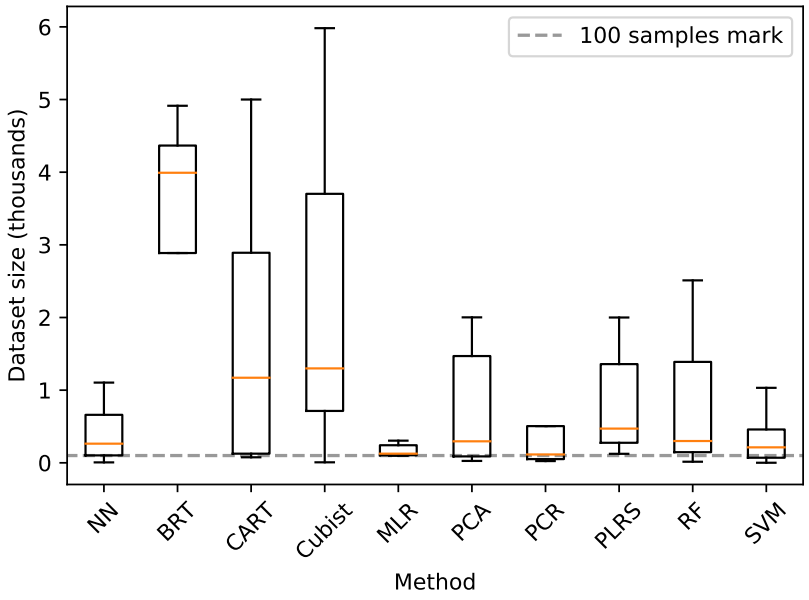

Figure 11. Boxplot of reported dataset sizes grouped by method. Outliers were removed.

A big proportion of studies that deal with the temporal dynamics of soil properties are related to soil-water interactions, as shown in the Topic 3 of our topic detection analysis (Fig. 9).

We found three main approaches to deal with the temporal variation of soil:

Temporal extrapolation: The studies generate models for a specific time-step including one or more predictors that vary on time to then apply that fitted model to another time-step (e.g. Grinand et al. (2017)).

Subtraction: The studies model two or more time-steps independently followed by a change analysis. For instance, Schillaci et al. (2017a) and Zhang et al. (2018b) subtracted the maps of the modelled properties from two different years to compute the change in SOC concentration and pH, respectively.

Dynamics: Studies that model the actual dynamics of a soil property based on some mechanistic or semi-mechanistic method. Stumpf et al. (2018) created yearly land use covers for 8,500 km^2 in Switzerland using a combination of Landsat 5-7-8 and field land use observations in order model the SOC dynamics based on the conversion regimes from their land use sequence patterns (Watson et al., 2014).

Despite there are ML algorithms that have the capacity to capture 4D structures (e.g. Convolutional Recurrent Neural Networks), we did not find studies using ML to continuously model space and time simultaneously. We think the main reason is that soil observations are usually sparse in space-time (Grunwald, 2016) and that is not possible to fulfil the dataset size requirements of such models. That is the reason why we mostly find studies that use a mechanistic or semi-mechanistic approach.

3.6 Uncertainty assessment

Uncertainty assessment is an important requirement for any model, especially if the predictions are going to be used to guide decision-making. In this review, 24% of the studies, among most topics, present uncertainty assessment or mention the importance of considering it (Fig. 12).

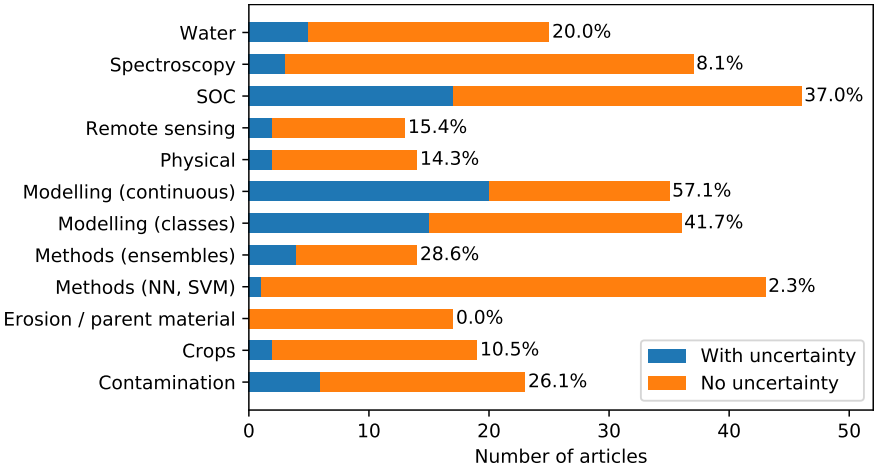

Figure 12. Number of articles per topic that report or mention "uncertainty". Percentage represents proportion of total articles per topic.

In this review, a major contributor to the promotion of uncertainty assessment in soil modelling is the GlobalSoilMap project (Arrouays et al., 2014) which, through specification developed by the DSM scientific community, recommends an uncertainty assessment of all their products. This is evident from Fig. 12, where topics related to DSM show a relatively high proportion of articles mentioning or reporting uncertainty. In the GlobalSoilMap specifications, the proposed uncertainty assessment method is the use of bootstraping when training the model (Stine, 1985), effectively making predictions with many models trained with subsets of the original data, to then estimate the 90% prediction interval (e.g. Castro-Franco et al. (2017); Ma et al. (2017)). Another approach is the use of quantile regression (Koenker and Bassett Jr, 1978) to estimate the complete conditional distribution of the prediction. This method has been recently applied in some DSM studies (Vaysse and Lagacherie, 2017; Sirsat et al., 2018; Cao et al., 2019). Less common approaches are the use of the fuzzy k-means with extragrades (Tranter et al., 2010) algorithm, which defines areas within the covariate space, with different levels of uncertainty, were a new observation (to be predicted) can be placed; and the use of Bayesian optimisation approaches (Snoek et al., 2015; Gal and Ghahramani, 2016).

4 General Discussion

4.1 Interpretability

Based on our findings, it is possible to state that, in general, ML methods have shown superior performance over more traditional methods in terms of predictive power. We now address the last questions from the aims of this paper — does an advanced model provide new insights that improve our knowledge and understanding of soils?

In order for a human to understand the decisions made by the model, the model has to be interpretable. The motivations for interpretability are varied, including trust, causality, transferability, informativeness and fairness (in ethical terms) (Lipton, 2016). In our review, researches usually associate advanced ML models with low interpretability. For instance, Brungard et al. (2015) assigned multiple models to different groups according to their complexity, with NN and SVM categorised as difficult to interpret compared to MLR or CART. Beguin et al. (2017) also mention the lower interpretability of ML models compared with an explicit geostatistical model. Because how to measure interpretability is usually not well defined, there are also contradictory opinions. For instance, RF is mostly considered in the category of low interpretability (Brungard et al., 2015; Were et al., 2015; Taghizadeh-Mehrjardi et al., 2016; Deng et al., 2018) but its use is also sometimes justified due to its ease of interpretability via the use of variables of importance (Jeong et al., 2017).

It is important to clearly define the goal of a modelling exercise. If we want to obtain the model with the greatest accuracy in order to solve a specific problem, maybe interpretability should not be an important factor. If we consider a) that nature is a complex combination of nonlinear phenomena, and b) the limited capacity of humans to understand non-linear relationships (Doherty and Balzer, 1988), by requiring our model complete transparency we are limiting its capability. However, it is important to corroborate that the model is a valid generalisation of the studied phenomenon. If our goal is to obtain new insights, it is important to consider that interpretability goes hand in hand with prior knowledge and biases, and that we could be optimising an algorithm to present misleading but plausible explanations (Lipton, 2016).

4.1.1 How can we increase interpretability?

A common conclusion reported by authors of the reviewed papers is that the selection of the most informative or relevant predictors before training the model can increase interpretability (Xiong et al., 2014; Prasad et al., 2018; Wang et al., 2018a; Keskin et al., 2019), although some authors do not recommend selection of predictors based on the researchers' knowledge since it could lead to biased and suboptimal model performance (Brungard et al., 2015; Keskin et al., 2019). This discordance leads to a large range in the number of the predictors used, with some extreme cases using more than 200 (Xiong et al., 2014; Keskin et al., 2019).

NN are some of the most performing models but, given the complexity of their operation, they are usually labelled as "black-box" models. In consequence, many authors have focused on trying to provide frameworks to interpret the knowledge extracted by these models. For instance, Bau et al. (2017) dissected a CNN to understand how different layers work and which features they favour by visualising their (neurons) activation map. Rauber et al. (2017) used the activation maps projected into a 2D

space in order to visualise and identify confusion zones, outliers, and clusters in the internal representations learned by the model.

In soil sciences, one of the reported methods to interpret ML models is to assess the importance of the variables used, usually derived from the number of times they have been used in the rules generated by tree-like models (Henderson et al., 2005; Martin et al., 2014; Schillaci et al., 2017b; Khanal et al., 2018). Another method to assess the relative influence of predictors in tree-like models is to estimate the average reduction of the error at each split of the tree, for all the predictors (Friedman, 2001). Another alternative, in the context of soil mapping, is to map the rules generated by the model to identify their spatial context or to map where important predictors were used (Bui et al., 2006). For CNNs, by feeding simulated data to a trained model, Ng et al. (2019) explored the most important wavelengths used when predicting multiple soil properties from soil spectral data using a sensitivity analysis. The logic behind their analysis is that modifying unimportant wavelengths should not affect the prediction. By plotting the variance for the predictions by wavelengths it is possible to unveil the most important areas of the spectrum (Fig. 13).

Interpretability is an important concept that should be revisited, since is not absolute nor static, hence a specific model cannot be simply labelled as interpretable or not. Linear models can quickly become unintelligible as we add more variables (Lou et al., 2012), and methods to better understand complex models such as NN are constantly being developed (Bau et al., 2017; Montavon et al., 2018; Zhang et al., 2018a).

4.2 New good practices

Thanks to the effort of some groups to rescue soil legacy data (Arrouays et al., 2017), and cheaper and faster methods to analyse soil samples, there is more soil data available than ever before. This data availability not only allows us to use new ML algorithms, which usually require more observations, but also opens the door to new ways to train those models. An important part of model development is validation. Literature traditionally recommends that an independent, unseen (by the model) dataset should be used as validation (Kohavi et al., 1995). In practice, the data is usually partitioned into training and validation datasets. A more stable solution is the use of k-cross-validation where the dataset is partitioned into k groups, where $k-1$ groups are used for training and 1 group for validation, repeating the training k times, each with a different validation group. When data availability is a limitation, researchers resort to techniques such as n-cross-validation or "leave-one-out" validation to make the most of the available data (Stevens et al., 2008; Pasini, 2015).

A new generation of models based on NNs have been introduced in the later years, which have revolutionised many fields. Deep learning (DL) models, consisting of multiple hidden layers of neurons, have many parameters (from hundreds to millions) which need to be fitted in the training process. This is the reason why they usually need access to large sample sizes. A second characteristic of these models is that they have a considerable number of hyper-parameters. Hyper-parameters are parameters that are not learned from the data during the training phase and include things like number of iterations during the training, learning rate, layer parameters, number of layers, etc. A common practice when training DL models is to split the original dataset into 3 sub datasets: training, validation and test. The training dataset is used to learn the parameters, the validation

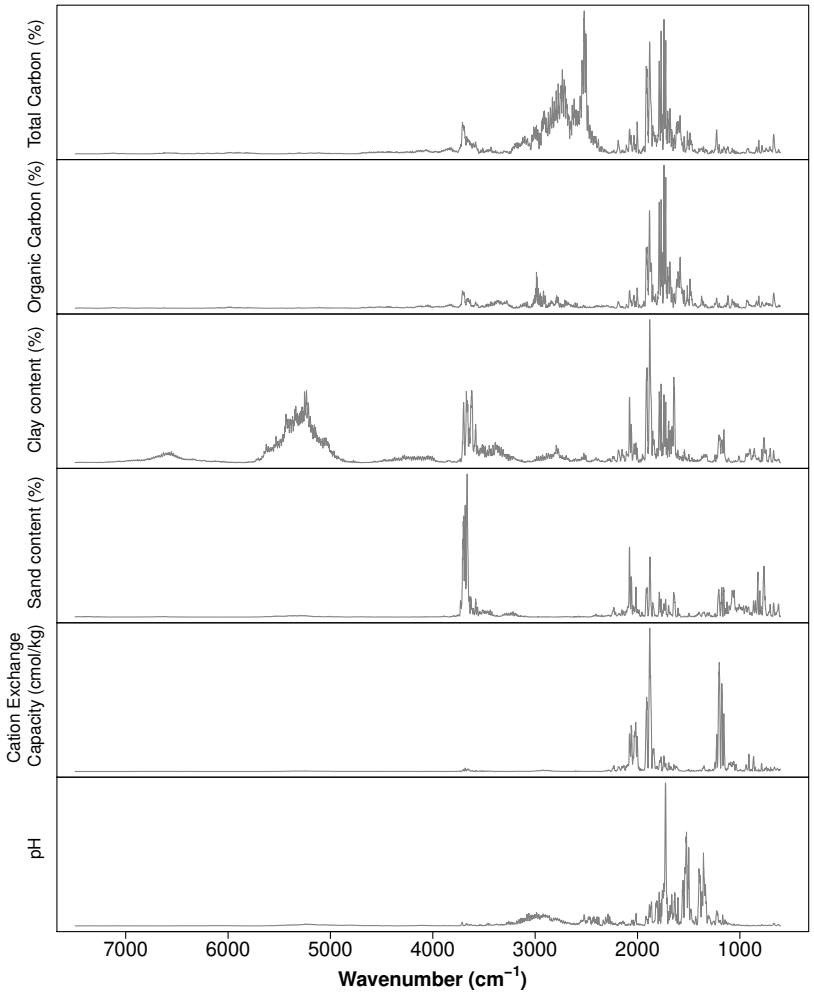

Figure 13. Sensitivity analysis of CNN model prediction as a function of wavelength. Average variance of predictions by wavelength. This analysis allows to explore the most important wavelengths in a CNN model. Adapted from Ng et al. (2019).

dataset to compare models fitted with different hyper-parameters in order to find the optimal combination, and the test dataset as the independent, unseen data.

In soil sciences, ML algorithms are usually trained using the traditional train/validation split or cross-validation (Keskin et al., 2019; Liang et al., 2019), or even no validation (Feng et al., 2019), except for some studies based on DL or with engineering background (e.g. Reale et al. (2018)), including some of our publications on the use of DL for DSM (Padarian et al., 2019c) or soil spectroscopy (Padarian et al., 2019b, a), which use a train/validation/test split. Considering the increasing size of datasets, we think soil scientist should transition towards the implementation of some DL practices such as dataset split and hyper-parameter optimisation (Bergstra and Bengio, 2012; Snoek et al., 2012), not only for NNs but for any algorithm that has hyper-parameters. Some potential candidates are random forest, Cubist, classification and regression trees, and support

vector machines. Most of the implementations of these algorithms have sensible default hyper-parameters but some studies report an important impact of them in their results (Mutanga et al., 2012; Lu et al., 2018). For general hyper-parameter tuning strategies, we refer the reader to Bergstra and Bengio (2012) for simple strategies such as grid or random search. For an in depth report of hyper-parameter tuning and its effects in the context of random forest, we refer the reader to Probst et al. (2019).

4.3 Commercial ML applications

This work explores the use of ML in soil sciences by exploring the current scientific literature, but use of ML extends beyond research and companies are very welcoming to this technology, specially in applications such as computer vision, speech recognition, natural language processing, and robot control (Jordan and Mitchell, 2015). It is not hard to imagine a commercial application of approaches such as soil properties prediction using vis-NIR spectroscopy, either in the laboratory or the field. While in research there are some transparency requirements, including describing the methods and data used, companies are usually very secretive about their methods since they are a trade secret that gives them a competitive advantage. Considering that lack of transparency, how can we be sure that the predictions of their models are good? There is not a unique answer but it should include at least some uncertainty assessment (as discussed in Section 3.6) and information about the range of soils used during training.

In terms of the reporting soil types coverage, different approaches can be applied. A simple, perhaps over-confident method can be reporting the geographical extent from where the soil samples used during training were collected (e.g. Tomasella et al. (2000) and Børgesen and Schaap (2005)), or a broad soil classification based in the soil characteristics such as "sandy soils" (e.g Schaap and Bouten (1996) and Shaw et al. (2000)). A better approach, based on the covariate space of the samples used during training is fuzzy k-means with extragrades, which has the benefit of describing both, coverage and uncertainty levels.

Even if uncertainty levels and coverage are reported, another factor to consider is how much we should trust in companies and their reports. Specially for applications involving public funding, but generally as a consumer protection measure, this type of products should be certifiable, in the same way many soil laboratories are. A usual approach is the use of reference materials (Dybczyński et al., 1979; Pueyo et al., 2001; Ahmed et al., 2017), which should be consistent with the model coverage reported. The properties measured in the reference materials should fall within the prediction interval produced by the model, with a confidence defined for each application.

5 Conclusions and recommendations

Aided by a topic modelling approach, we were able to review the status of ML in soil sciences. We observed a general increase in the adoption of ML methods in time, mostly concentrated in developed countries. This gap is probably due to the link between science, technology and development. We believe that proper inter-institutional collaboration plans should be put in place in order to close this gap.

By using topic modelling, we identified twelve categories of studies were ML is commonly used, namely remote sensing, soil organic carbon, water, contamination, methods (ensembles), erosion and parent material, methods (NN, SVM), spectroscopy,

modelling (classes), crops, physical, modelling (continuous). The final topic model successfully captured relationships between topics such as modelling of continuous and categorical soil properties, and water, given that all these topics share a spatial component.

We also found that advanced ML methods usually perform better than simpler approaches thanks to their capability to capture non-linear relationships. However, it is important to note that more advanced methods usually require more data and that some precautions should be taken in order to avoid obtaining misleading results. Considering parsimony is always advised, hence if only a small, simple dataset is available, we recommend using a simple model. This also applies to the number of predictors. In consequence, according to many authors of the reviewed articles, is better to use meaningful predictors instead of relying on the model capabilities to "select the best variables" in order to improve interpretability.

Interpretability is an important aspect to consider when applying advanced ML methods in order to improve our knowledge and understanding of soil. Simpler methods (e.g. linear models) have been used for a long time and the way of interpreting them is well defined. More advanced methods (e.g. neural networks) are usually considered as "black box" models, but that is just a reflection of the current research state and not because it is impossible to interpret them. During our review, we found studies that proposed some solutions to improve their interpretability and we foresee that a large number of studies will focus on this topic.

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

Acknowledgements. The authors acknowledge the University of Sydney HPC service at The University of Sydney for providing HPC resources that have contributed to the research results reported within this paper.

References

Ahmad, S., Kalra, A., and Stephen, H.: Estimating soil moisture using remote sensing data: A machine learning approach, Advances in Water Resources, 33, 69–80, 2010.

Ahmed, O., Habbani, F. I., Mustafa, A., Mohamed, E., Salih, A., and Seedig, F.: Quality assessment statistic evaluation of X-ray fluorescence via NIST and IAEA standard reference materials, World Journal of Nuclear Science and Technology, 7, 121, 2017.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., et al.: GlobalSoilMap: Toward a fine-resolution global grid of soil properties, in: Advances in agronomy, vol. 125, pp. 93–134, Elsevier, 2014.

Arrouays, D., Leenaars, J. G., de Forges, A. C. R., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., et al.: Soil legacy data rescue via GlobalSoilMap and other international and national initiatives, GeoResJ, 14, 1–19, 2017.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6541–6549, 2017.

Beguin, J., Fuglstad, G.-A., Mansuy, N., and Paré, D.: Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches, Geoderma, 306, 195–205, 2017.

Bergstra, J. and Bengio, Y.: Random search for hyper-parameter optimization, Journal of Machine Learning Research, 13, 281–305, 2012.

Björk, B.-C.: Open access to scientific articles: a review of benefits and challenges, Internal and emergency medicine, pp. 247–253, 2017.

Blanco, C. M. G., Gomez, V. M. B., Crespo, P., and Ließ, M.: Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest, Geoderma, 316, 100–114, 2018.

Blei, D. M.: Probabilistic topic models, Communications of the ACM, 55, 77–84, 2012.

Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, Journal of machine Learning research, 3, 993–1022, 2003.

Bondi, G., Creamer, R., Ferrari, A., Fenton, O., and Wall, D.: Using machine learning to predict soil bulk density on the basis of visual parameters: Tools for in-field and post-field evaluation, Geoderma, 318, 137–147, 2018.

Børgesen, C. D. and Schaap, M. G.: Point and parameter pedotransfer functions for water retention predictions for Danish soils, Geoderma, 127, 154–167, 2005.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., and Edwards Jr, T. C.: Machine learning for predicting soil classes in three semi-arid landscapes, Geoderma, 239, 68–83, 2015.

Bui, E. N., Henderson, B. L., and Viergever, K.: Knowledge discovery from models of soil properties developed through data mining, Ecological Modelling, 191, 431–446, 2006.

Butler, B. M., O'Rourke, S. M., and Hillier, S.: Using rule-based regression models to predict and interpret soil properties from X-ray powder diffraction data, Geoderma, 329, 43–53, https://doi.org/10.1016/j.geoderma.2018.04.005, 2018.

Camera, C., Zomeni, Z., Noller, J. S., Zissimos, A. M., Christoforou, I. C., and Bruggeman, A.: A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization, Geoderma, 285, 35–49, 2017.

Cao, B., Domke, G. M., Russell, M. B., and Walters, B. F.: Spatial modeling of litter and soil carbon stocks on forest land in the conterminous United States, Science of the Total Environment, 654, 94–106, 2019.

Castro-Franco, M., Domenech, M. B., Borda, M. R., Costa, J., et al.: Spatial dataset of topsoil texture for the southern Argentine Pampas, Geoderma regional, 2017.

Catlett, J.: Mega induction: A test flight, in: Machine Learning Proceedings 1991, pp. 596–599, Elsevier, 1991.

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

Chlingaryan, A., Sukkarieh, S., and Whelan, B.: Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, Computers and electronics in agriculture, 151, 61–69, 2018.

Coopersmith, E. J., Minsker, B. S., Wenzel, C. E., and Gilmore, B. J.: Machine learning assessments of soil drying for agricultural planning, Computers and electronics in agriculture, 104, 93–104, 2014.

Cortes, C., Jackel, L. D., Solla, S. A., Vapnik, V., and Denker, J. S.: Learning curves: Asymptotic values and rate of convergence, in: Advances in Neural Information Processing Systems, pp. 327–334, 1994.

Costa, J. G., Reigosa, M., Matías, J., and Covelo, E.: Soil Cd, Cr, Cu, Ni, Pb and Zn sorption and retention models using SVM: variable selection and competitive model, Science of the Total Environment, 593, 508–522, 2017.

Dai, F., Zhou, Q., Lv, Z., Wang, X., and Liu, G.: Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau, Ecological Indicators, 45, 184–194, 2014.

Deng, X., Chen, X., Ma, W., Ren, Z., Zhang, M., Grieneisen, M. L., Long, W., Ni, Z., Zhan, Y., and Lv, X.: Baseline map of organic carbon stock in farmland topsoil in East China, Agriculture, ecosystems & environment, 254, 213–223, 2018.

Dharumarajan, S., Hegde, R., and Singh, S.: Spatial prediction of major soil properties using Random Forest techniques-A case study in semi-arid tropics of South India, Geoderma Regional, 10, 154–162, 2017.

Dobarco, M. R., Cousin, I., Le Bas, C., and Martin, M. P.: Pedotransfer functions for predicting available water capacity in French soils, their applicability domain and associated uncertainty, Geoderma, 336, 81–95, 2019.

Doherty, M. E. and Balzer, W. K.: Cognitive feedback, in: Advances in psychology, vol. 54, pp. 163–197, Elsevier, 1988.

Dybczyński, R., Tugsavul, A., and Suschny, O.: Soil-5, a new IAEA certified reference material for trace element determinations, Geostandards Newsletter, 3, 61–87, 1979.

Fajardo, M., McBratney, A., and Whelan, B.: Fuzzy clustering of Vis–NIR spectra for the objective recognition of soil morphological horizons in soil profiles, Geoderma, 263, 244–253, 2016.

Farfani, H. A., Behnamfar, F., and Fathollahi, A.: Dynamic analysis of soil-structure interaction using the neural networks and the support vector machines, Expert Systems with Applications, 42, 8971–8981, 2015.

Feng, Y., Cui, N., Hao, W., Gao, L., and Gong, D.: Estimation of soil temperature from meteorological data using different machine learning models, Geoderma, 338, 67–77, 2019.

Flynn, T., Rozanov, A., de Clercq, W., Warr, B., and Clarke, C.: Semi-automatic disaggregation of a national resource inventory into a farm-scale soil depth class map, Geoderma, 337, 1136–1145, 2019.

Friedman, J. H.: Greedy function approximation: a gradient boosting machine, Annals of statistics, pp. 1189–1232, 2001.

Gal, Y. and Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International conference on machine learning, pp. 1050–1059, 2016.

Gao, M., Li, H.-Y., Liu, D., Tang, J., Chen, X., Chen, X., Blöschl, G., and Leung, L. R.: Identifying the dominant controls on macropore flow velocity in soils: A meta-analysis, Journal of Hydrology, 567, 590–604, 2018.

Geissen, V., Kampichler, C., López-de Llergo-Juárez, J., and Galindo-Acántara, A.: Superficial and subterranean soil erosion in Tabasco, tropical Mexico: development of a decision tree modeling approach, Geoderma, 139, 277–287, 2007.

Gomes, L. C., Faria, R. M., de Souza, E., Veloso, G. V., Schaefer, C. E. G., and Fernandes Filho, E. I.: Modelling and mapping soil organic carbon stocks in Brazil, Geoderma, 340, 337–350, 2019.

Greifeneder, F., Khamala, E., Sendabo, D., Wagner, W., Zebisch, M., Farah, H., and Notarnicola, C.: Detection of soil moisture anomalies based on Sentinel-1, Physics and Chemistry of the Earth, Parts A/B/C, 2018.

Grinand, C., Le Maire, G., Vieilledent, G., Razakamanarivo, H., Razafimbelo, T., and Bernoux, M.: Estimating temporal changes in soil carbon stocks at ecoregional scale in Madagascar using remote-sensing, International journal of applied earth observation and geoinformation, 54, 1–14, 2017.

Grunwald, S.: Multi-criteria characterization of recent digital soil mapping and modeling approaches, Geoderma, 152, 195–207, 2009.

Grunwald, S.: What do we really know about the space–time continuum of soil-landscapes?, in: Environmental Soil-Landscape Modeling, pp. 16–49, CRC Press, 2016.

Grunwald, S., Vasques, G. M., and Rivero, R. G.: Fusion of soil and remote sensing data to model soil properties, in: Advances in Agronomy, vol. 131, pp. 1–109, Elsevier, 2015.

Han, J., Mao, K., Xu, T., Guo, J., Zuo, Z., and Gao, C.: A soil moisture estimation framework based on the cart algorithm and its application in china, Journal of hydrology, 563, 65–75, 2018.

Hanna, A. M., Ural, D., and Saygili, G.: Neural network model for liquefaction potential in soil deposits using Turkey and Taiwan earthquake data, Soil Dynamics and Earthquake Engineering, 27, 521–540, 2007.

Heggemann, T., Welp, G., Amelung, W., Angst, G., Franz, S. O., Koszinski, S., Schmidt, K., and Pätzold, S.: Proximal gamma-ray spectrometry for site-independent in situ prediction of soil texture on ten heterogeneous fields in Germany using support vector machines, Soil and Tillage Research, 168, 99–109, https://doi.org/10.1016/j.still.2016.10.008, 2017.

Henderson, B. L., Bui, E. N., Moran, C. J., and Simon, D.: Australia-wide predictions of soil properties using decision trees, Geoderma, 124, 383–398, 2005.

Hutson, M.: Boycott highlights AI's publishing rebellion, 2018.

Ivushkin, K., Bartholomeus, H., Bregt, A. K., Pulatov, A., Bui, E. N., and Wilford, J.: Soil salinity assessment through satellite thermography for different irrigated and rainfed crops, International journal of applied earth observation and geoinformation, 68, 230–237, 2018.

Jeong, G., Oeverdieck, H., Park, S. J., Huwe, B., and Ließ, M.: Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain, Catena, 154, 73–84, 2017.

Jordan, M. I. and Mitchell, T. M.: Machine learning: Trends, perspectives, and prospects, Science, 349, 255–260, 2015.

Karandish, F. and Šimůnek, J.: A field-modeling study for assessing temporal variations of soil-water-crop interactions under water-saving irrigation strategies, Agricultural water management, 178, 291–303, 2016.

Keskin, H., Grunwald, S., and Harris, W. G.: Digital mapping of soil carbon fractions with machine learning, Geoderma, 339, 40–58, 2019.

Khadim, F. K., Su, H., Xu, L., and Tian, J.: Soil salinity mapping in Everglades National Park using remote sensing techniques and vegetation salt tolerance, Physics and Chemistry of the Earth, Parts A/B/C, 2019.

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., and Shearer, S.: Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield, Computers and electronics in agriculture, 153, 213–225, 2018.

Kheir, R. B., Chorowicz, J., Abdallah, C., and Dhont, D.: Soil and bedrock distribution estimated from gully form and frequency: A GIS-based decision-tree model for Lebanon, Geomorphology, 93, 482–492, https://doi.org/10.1016/j.geomorph.2007.03.010, 2008.

Koenker, R. and Bassett Jr, G.: Regression quantiles, Econometrica: journal of the Econometric Society, pp. 33–50, 1978.

Koestel, J. and Jorda, H.: What determines the strength of preferential transport in undisturbed soil under steady-state flow?, Geoderma, 217, 144–160, 2014.

Kohavi, R. et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Ijcai, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.

Kovačević, M., Bajat, B., and Gajić, B.: Soil type classification and estimation of soil properties using support vector machines, Geoderma, 154, 340–347, 2010.

Lacoste, M., Lemercier, B., and Walter, C.: Regional mapping of soil parent material by machine learning based on point data, Geomorphology, 133, 90–99, https://doi.org/10.1016/j.geomorph.2011.06.026, 2011.

Leenaars, J. G., Claessens, L., Heuvelink, G. B., Hengl, T., González, M. R., van Bussel, L. G., Guilpart, N., Yang, H., and Cassman, K. G.: Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-Saharan Africa, Geoderma, 324, 18–36, 2018.

Liang, Z., Chen, S., Yang, Y., Zhao, R., Shi, Z., and Rossel, R. A. V.: National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China, Geoderma, 335, 47–56, 2019.

Lin, J.: Divergence measures based on the Shannon entropy, IEEE Transactions on Information theory, 37, 145–151, 1991.

Lipton, Z. C.: The mythos of model interpretability, arXiv preprint arXiv:1606.03490, 2016.

Liu, S., Yang, Y., Shen, H., Hu, H., Zhao, X., Li, H., Liu, T., and Fang, J.: No significant changes in topsoil carbon in the grasslands of northern China between the 1980s and 2000s, Science of the total environment, 624, 1478–1487, 2018.

Lou, Y., Caruana, R., and Gehrke, J.: Intelligible models for classification and regression, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 150–158, ACM, 2012.

Lu, W., Lu, D., Wang, G., Wu, J., Huang, J., and Li, G.: Examining soil organic carbon distribution and dynamic change in a hickory plantation region with Landsat and ancillary data, Catena, 165, 576–589, 2018.

Ma, Y., Minasny, B., and Wu, C.: Mapping key soil properties to support agricultural production in Eastern China, Geoderma Regional, 10, 144–153, 2017.

Ma, Y., Minasny, B., Malone, B., and McBratney, A.: Pedology and digital soil mapping (DSM), European Journal of Soil Science, . (Under review), 2019.

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., and Beaudoin, A.: Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the k-nearest neighbor method, Geoderma, 235, 59–73, 2014.

Märker, M., Pelacani, S., and Schröder, B.: A functional entity approach to predict soil erosion processes in a small Plio-Pleistocene Mediterranean catchment in Northern Chianti, Italy, Geomorphology, 125, 530–540, 2011.

Martin, M., Orton, T., Lacarce, E., Meersmans, J., Saby, N., Paroissien, J., Jolivet, C., Boulonne, L., and Arrouays, D.: Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale, Geoderma, 223, 97–107, 2014.

Martinez, G., Weltz, M., Pierson, F. B., Spaeth, K. E., and Pachepsky, Y.: Scale effects on runoff and soil erosion in rangelands: Observations and estimations with predictors of different availability, Catena, 151, 161–173, 2017.

Massawe, B. H., Subburayalu, S. K., Kaaya, A. K., Winowiecki, L., and Slater, B. K.: Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning, Geoderma, 311, 143–148, 2018.

Matthew and Honnibal, M. I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, . https://github.com/explosion/spaCy/, 2017.

McBratney, A., de Gruijter, J., and Bryce, A.: Pedometrics timeline, Geoderma, 338, 568–575, 2019.

McCallum, A. K.: MALLET: A Machine Learning for Language Toolkit, http://mallet.cs.umass.edu, 2002.

Minasny, B. and Flantis, D.: "Helicopter research": who benefits from international studies in Indonesia?, https://theconversation.com/helicopter-research-who-benefits-from-international-studies-in-indonesia-102165. Accessed: 29/04/2019, 2018.

Mjolsness, E. and DeCoste, D.: Machine learning for science: state of the art and future prospects, science, 293, 2051–2055, 2001.

Montavon, G., Samek, W., and Müller, K.-R.: Methods for interpreting and understanding deep neural networks, Digital Signal Processing, 73, 1–15, 2018.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., and Mouazen, A. M.: Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy, Biosystems Engineering, 152, 104–116, 2016.

Mutanga, O., Adam, E., and Cho, M. A.: High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm, International Journal of Applied Earth Observation and Geoinformation, 18, 399–406, 2012.

Naderi-Boldaji, M., Tekeste, M. Z., Nordstorm, R. A., Barnard, D. J., and Birrel, S. J.: A mechanical-dielectric-high frequency acoustic sensor fusion for soil physical characterization, Computers and Electronics in Agriculture, 156, 10–23, 2019.

Ng, W., McBratney, A., Minasny, B., Padarian, J., Monatzerolghaem, M., Ferguson, R., and Bailey, S.: Deep learning for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra, Geoderma, . (Under review), 2019.

Oh, Y.-Y., Yun, S.-T., Yu, S., Kim, H.-J., and Jun, S.-C.: A novel wavelet-based approach to characterize dynamic environmental factors controlling short-term soil surface CO2 flux: Application to a controlled CO2 release test site (EIT) in South Korea, Geoderma, 337, 76–90, 2019.

Padarian, J., Minasny, B., and McBratney, A.: Transfer learning to localise a continental soil vis-NIR calibration model, Geoderma, 340, 279–288, 2019a.

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

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

Pasini, A.: Artificial neural networks for small dataset analysis, Journal of thoracic disease, 7, 953, 2015.

Perlich, C., Provost, F., and Simonoff, J. S.: Tree induction vs. logistic regression: A learning-curve analysis, Journal of Machine Learning Research, 4, 211–255, 2003.

Poggio, L., Gimona, A., Spezia, L., and Brewer, M. J.: Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA, Geoderma, 277, 69–82, 2016.

Prasad, R., Deo, R. C., Li, Y., and Maraseni, T.: Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors, Soil and Tillage Research, 181, 63–81, 2018.

Probst, P., Wright, M. N., and Boulesteix, A.-L.: Hyperparameters and tuning strategies for random forest, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9, e1301, 2019.

Pueyo, M., Rauret, G., Bacon, J., Gomez, A., Muntau, H., Quevauviller, P., and López-Sánchez, J.: A new organic-rich soil reference material certified for its EDTA-and acetic acid-extractable contents of Cd, Cr, Cu, Ni, Pb and Zn, following collaboratively tested and harmonised procedures, Journal of Environmental Monitoring, 3, 238–242, 2001.

Rauber, P. E., Fadel, S. G., Falcao, A. X., and Telea, A. C.: Visualizing the hidden activity of artificial neural networks, IEEE transactions on visualization and computer graphics, 23, 101–110, 2017.

Reale, C., Gavin, K., Librić, L., and Jurić-Kaćunić, D.: Automatic classification of fine-grained soils using CPT measurements and Artificial Neural Networks, Advanced Engineering Informatics, 36, 207–215, 2018.

Reeves, M. K., Perdue, M., Munk, L. A., and Hagedorn, B.: Predicting risk of trace element pollution from municipal roads using site-specific soil samples and remotely sensed data, Science of the Total Environment, 630, 578–586, 2018.

Rial, M., Cortizas, A. M., Taboada, T., and Rodríguez-Lado, L.: Soil organic carbon stocks in Santa Cruz Island, Galapagos, under different climate change scenarios, Catena, 156, 74–81, 2017.

Röder, M., Both, A., and Hinneburg, A.: Exploring the space of topic coherence measures, in: Proceedings of the eighth ACM international conference on Web search and data mining, pp. 399–408, ACM, 2015.

Rossiter, D. G.: Past, present & future of information technology in pedometrics, Geoderma, 324, 131–137, 2018.

Rudin, C. and Wagstaff, K. L.: Machine learning for science and society, 2014.

Sagasti, F. R.: Underdevelopment, science and technology: the point of view of the underdeveloped countries, Science Studies, 3, 47–59, 1973.

Schaap, M. G. and Bouten, W.: Modeling water retention curves of sandy soils using neural networks, Water Resources Research, 32, 3033–3040, 1996.

Schillaci, C., Acutis, M., Lombardo, L., Lipani, A., Fantappie, M., Märker, M., and Saia, S.: Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling, Science of the total environment, 601, 821–832, 2017a.

Schillaci, C., Lombardo, L., Saia, S., Fantappiè, M., Märker, M., and Acutis, M.: Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region, Geoderma, 286, 35–45, 2017b.

Shavlik, J. W., Mooney, R. J., and Towell, G. G.: Symbolic and neural learning algorithms: An experimental comparison, Machine learning, 6, 111–143, 1991.

Shaw, J., West, L., Radcliffe, D., and Bosch, D.: Preferential flow and pedotransfer functions for transport properties in sandy Kandiudults, Soil Science Society of America Journal, 64, 670–678, 2000.

Sirsat, M., Cernadas, E., Fernández-Delgado, M., and Barro, S.: Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods, Computers and Electronics in Agriculture, 154, 120–133, 2018.

Snoek, J., Larochelle, H., and Adams, R. P.: Practical bayesian optimization of machine learning algorithms, in: Advances in neural information processing systems, pp. 2951–2959, 2012.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., and Adams, R.: Scalable bayesian optimization using deep neural networks, in: International conference on machine learning, pp. 2171–2180, 2015.

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, 81, 1413–1426, 2017.

Song, X.-D., Yang, F., Ju, B., Li, D.-C., Zhao, Y.-G., Yang, J.-L., and Zhang, G.-L.: The influence of the conversion of grassland to cropland on changes in soil organic carbon and total nitrogen stocks in the Songnen Plain of Northeast China, Catena, 171, 588–601, 2018.

Sonnenwald, D. H.: Scientific collaboration, Annual review of information science and technology, 41, 643–681, 2007.

Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., and Ben-Dor, E.: Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils, Geoderma, 144, 395–404, 2008.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D.: Exploring topic coherence over many models and many topics, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 952–961, Association for Computational Linguistics, 2012.

Stine, R. A.: Bootstrap prediction intervals for regression, Journal of the American Statistical Association, 80, 1026–1031, 1985.

Stumpf, F., Keller, A., Schmidt, K., Mayr, A., Gubler, A., and Schaepman, M.: Spatio-temporal land use dynamics and soil organic carbon in Swiss agroecosystems, Agriculture, ecosystems & environment, 258, 129–142, 2018.

Subburayalu, S., Jenhani, I., and Slater, B.: Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees, Geoderma, 213, 334–345, 2014.

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., and Ding, Y.: The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation, Journal of the American Society for Information Science and Technology, 62, 185–204, 2011.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., and Triantafilis, J.: Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran, Geoderma, 253, 67–77, 2015.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., and Kerry, R.: Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran, Geoderma, 266, 98–110, 2016.

Tomasella, J., Hodnett, M. G., and Rossato, L.: Pedotransfer Functions for the Estimation of Soil Water Retention in Brazilian Soils, Soil Science Society of America Journal, 64, 327, 2000.

Tranter, G., Minasny, B., and McBratney, A.: Estimating Pedotransfer Function Prediction Limits Using Fuzzy k-Means with Extragrades, Soil Sci. Soc. Am. J., 74, 1967–1975, 2010.

Tziachris, P., Aschonitis, V., Chatzistathis, T., and Papadopoulou, M.: Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters, Catena, 174, 206–216, 2019.

Vaysse, K. and Lagacherie, P.: Using quantile regression forest to estimate uncertainty of digital soil mapping products, Geoderma, 291, 55–64, 2017.

Vincent, S., Lemercier, B., Berthier, L., and Walter, C.: Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships, Geoderma, 311, 130–142, 2018.

Viscarra-Rossel, R. and Behrens, T.: Using data mining to model and interpret soil diffuse reflectance spectra, Geoderma, 158, 46–54, 2010.

Řehůřek, R. and Sojka, P.: Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50, ELRA, Valletta, Malta, http://is.muni.cz/publication/884893/en, 2010.

Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Li Liu, D., Simpson, M., McGowen, I., and Sides, T.: Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia, Ecological indicators, 88, 425–438, 2018a.

Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., and Li Liu, D.: High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia, Science of The Total Environment, 630, 367–378, 2018b.

Ware, M. and Mabe, M.: The STM report: An overview of scientific and scholarly journal publishing, 2015.

Warner, D. L., Guevara, M., Inamdar, S., and Vargas, R.: Upscaling soil-atmosphere CO2 and CH4 fluxes across a topographically complex forested landscape, Agricultural and forest meteorology, 264, 80–91, 2019.

Watson, S. J., Luck, G. W., Spooner, P. G., and Watson, D. M.: Land-use change: incorporating the frequency, sequence, time span, and magnitude of changes into ecological research, Frontiers in Ecology and the Environment, 12, 241–249, 2014.

Were, K., Bui, D. T., Dick, Ø. B., and Singh, B. R.: A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape, Ecological Indicators, 52, 394–403, 2015.

Wu, G., Kechavarzi, C., Li, X., Wu, S., Pollard, S. J., Sui, H., and Coulon, F.: Machine learning models for predicting PAHs bioavailability in compost amended soils, Chemical engineering journal, 223, 747–754, 2013.

Wu, Q., Zhang, C., Hong, Q., and Chen, L.: Topic evolution based on LDA and HMM and its application in stem cell research, Journal of Information Science, 40, 611–620, 2014.

Xie, X.-L. and Li, A.-B.: Identification of soil profile classes using depth-weighted visible–near-infrared spectral reflectance, Geoderma, 325, 90–101, https://doi.org/10.1016/j.geoderma.2018.03.029, 2018.

Xing, L., Li, L., Gong, J., Ren, C., Liu, J., and Chen, H.: Daily soil temperatures predictions for various climates in United States using data-driven model, Energy, 160, 430–440, 2018.

Xiong, X., Grunwald, S., Myers, D. B., Kim, J., Harris, W. G., and Comerford, N. B.: Holistic environmental soil-landscape modeling of soil organic carbon, Environmental Modelling & Software, 57, 202–215, 2014.

Xu, Y., Smith, S. E., Grunwald, S., Abd-Elrahman, A., and Wani, S. P.: Incorporation of satellite remote sensing pan-sharpened imagery into digital soil prediction and mapping models to characterize soil property variability in small agricultural fields, ISPRS journal of photogrammetry and remote sensing, 123, 1–19, 2017.

Zeynoddin, M., Bonakdari, H., Ebtehaj, I., Esmaeilbeiki, F., Gharabaghi, B., and Haghi, D. Z.: A reliable linear stochastic daily soil temperature forecast model, Soil and Tillage Research, 189, 73–87, 2019.

Zhang, C., Mishra, D. R., and Pennings, S. C.: Mapping salt marsh soil properties using imaging spectroscopy, ISPRS Journal of Photogrammetry and Remote Sensing, 148, 221–234, 2019.

Zhang, Q., Nian Wu, Y., and Zhu, S.-C.: Interpretable convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8827–8836, 2018a.

Zhang, Y., Sui, B., Shen, H., and Wang, Z.: Estimating temporal changes in soil pH in the black soil region of Northeast China using remote sensing, Computers and Electronics in Agriculture, 154, 204–212, 2018b.

Zhou, D., Ji, X., Zha, H., and Giles, C. L.: Topic evolution and social interactions: how authors effect research, in: Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 248–257, ACM, 2006.

Acronyms

AI artificial intelligence.

CART classification and regression tree.

DSM digital soil mapping.

LDA Latent Dirichlet Allocation.

LM linear model.

MARS Multivariate Adaptive Regression Spline.

ML machine learning.

MLR multiple linear regression.

NN Neural Networks.

PLSR Partial Least Squares Regression.

RF random forest.

SOC soil organic carbon.

SVM support vector machines.

Appendix A

A1

Table A1: List of journals by publisher and number of articles that marched the term ’soil "machine learning"’ in a full-text search.

Journal	Articles
Geoderma	113
Science of The Total Environment	29
CATENA	18
Geoderma Regional	13
Computers and Electronics in Agriculture	12
Journal of Hydrology	11
Ecological Indicators	8
Remote Sensing of Environment	7
International Journal of Applied Earth Observation and Geoinformation	6
Agriculture, Ecosystems & Environment	5
Soil and Tillage Research	5
Journal of Terramechanics	5
Soil Biology and Biochemistry	5
Agricultural and Forest Meteorology	4
Computers & Geosciences	3
Chemometrics and Intelligent Laboratory Systems	3
Agricultural Water Management	3
Construction and Building Materials	3
ISPRS Journal of Photogrammetry and Remote Sensing	3
Forest Ecology and Management	3
Chemosphere	3
Environmental Modelling & Software	3
Environmental Pollution	3
Advanced Engineering Informatics	3
Geomorphology	3
Computers and Geotechnics	3
Advances in Water Resources	3

Continued on next page

Table A1: List of journals by publisher and number of articles that marched the term ’soil "machine learning"’ in a full-text search.

Journal	Articles
Journal of Environmental Management	2
Soil Dynamics and Earthquake Engineering	2
Applied Soft Computing	2
Ecological Engineering	2
Journal of Geochemical Exploration	2
Sensors and Actuators A: Physical	2
Physics and Chemistry of the Earth, Parts A/B/C	2
Journal of Photochemistry and Photobiology B: Biology	1
Analytica Chimica Acta	1
Applied Geography	1
Applied Ocean Research	1
Chemical Geology	1
Information Processing in Agriculture	1
Geoscience Frontiers	1
Tunnelling and Underground Space Technology	1
Expert Systems with Applications	1
Ecological Modelling	1
Pedobiologia	1
Applied Radiation and Isotopes	1
Spectrochimica Acta Part B: Atomic Spectroscopy	1
iScience	1
Applied Geochemistry	1
Journal of Rock Mechanics and Geotechnical Engineering	1
Measurement	1
Environmental Technology & Innovation	1
Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy	1
Advances in Agronomy	1
Environmental Research	1
Advances in Space Research	1
Journal of Hazardous Materials	1

Continued on next page

Table A1: List of journals by publisher and number of articles that marched the term ’soil "machine learning"’ in a full-text search.

Journal	Articles
Energy	1
Sustainable Computing: Informatics and Systems	1
Chemical Engineering Journal	1
Biosystems Engineering	1
Reliability Engineering & System Safety	1