

 Open access • Journal Article • DOI:10.1016/J.EARSCIREV.2020.103359

## **Machine learning for digital soil mapping: Applications, challenges and suggested solutions** — [Source link](#)

Alexandre M. J. C. Wadoux, Budiman Minasny, Alex B. McBratney

**Institutions:** University of Sydney

**Published on:** 01 Nov 2020 - Earth-Science Reviews (Elsevier)

**Topics:** Digital soil mapping and Soil map

Related papers:

- [On digital soil mapping](#)
- [Random Forests](#)
- [SoilGrids250m: Global gridded soil information based on machine learning](#)
- [Digital soil mapping: A brief history and some lessons](#)
- [Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/machine-learning-for-digital-soil-mapping-applications-34kozxg724>

This manuscript is a **non-peer reviewed** preprint that has been submitted for publication. Subsequent versions of this manuscript may have updated content. Feedback and comments are welcomed, feel free to contact the corresponding author:

Alexandre Wadoux  
[alexandre.wadoux@sydney.edu.au](mailto:alexandre.wadoux@sydney.edu.au)

# Machine learning for digital soil mapping: applications, challenges and suggested solutions

Alexandre M.J-C. Wadoux<sup>a,\*</sup>, Budiman Minasny<sup>a</sup>, Alex B. McBratney<sup>a</sup>

<sup>a</sup>*Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Australia*

---

## Abstract

The uptake of machine learning (ML) algorithms in digital soil mapping (DSM) is transforming the way soil scientists produce their maps. Machine learning is currently applied to mapping soil properties or classes much in the same way as other unrelated fields of science. Mapping of soil, however, has unique aspects which require adaptations of the ML algorithms. These features are for example, but not limited to, the inclusion of pedological knowledge into the ML algorithm, the accounting of spatial structure present in the soil data, or the desire to increase our scientific understanding of the distribution and genesis of soil from a calibrated ML model. Tackling these challenges is critical for machine learning to gain credibility and scientific consistency in soil science. In this article, we review the current applications of machine learning in digital soil mapping and suggest improvements. We found a growing interest of the use of ML in DSM. Most studies focus on obtaining accurate maps and disregard the characteristics of soil data, such as spatial autocorrelation. Only a few studies account for existing soil knowledge or quantify the uncertainty of the predicted maps. We then discuss the challenges related to the application of ML for soil mapping and offer solutions from existing studies in the natural sciences. The challenges are organized as follows: sampling, resampling, accounting for the spatial information, multivariate mapping, uncertainty analysis, validation, integration of pedological knowledge and, interpretation of the models. We conclude that for future developments, machine learning should incorporate three core elements: *plausibility*, *interpretability*, and *explainability*, which will trigger soil scientists to move beyond model prediction and towards explanation of soil processes.

---

\*Corresponding author: Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, New South Wales, Australia

*Email address:* alexandre.wadoux@sydney.edu.au (Alexandre M.J-C. Wadoux)

*Keywords:* Soil science, Pedometrics, Data mining, Spatial data, Geostatistics, Random forest

---

## 1. Introduction

1 In recent years, soil science has witnessed a considerable increase in digital soil  
2 mapping activities. This is caused by the convergence of several timely factors which  
3 are, among others, a huge demand for quantitative and spatial soil information, the  
4 accumulation of databases of measured or inferred soil properties coupled with ex-  
5 haustively known environmental variables and the development of numerical models  
6 combined with computer resources to mine these stores of soil data. The digital soil  
7 mapping (DSM) framework was formalized by the publication of [McBratney et al.](#)  
8 (2003) which builds on Jenny's  $S = \text{clorpt}$  model ([Jenny, 1941](#)) of soil formation,  
9 where  $S$  is the soil and the acronym *clorpt* stands for climate, organisms, relief, par-  
10 ent material and time, respectively. In short, *clorpt* is a list of variables which, if  
11 they are known without error, are likely to explain the soil variation over a region.  
12 [McBratney et al. \(2003\)](#) supplemented Jenny's formulation with  $n$ , which stands for  
13 spatial position, and advocated the *scorpan* model for soil spatial variation. This  
14 updated equation provides a spatial model to express quantitatively the relationship  
15 between a soil property or class and environmental variables, for a given spatial lo-  
16 cation.

17  
18 Conventionally, spatial prediction of soil has been embedded in the geostatisti-  
19 cal framework ([Heuvelink & Webster, 2001](#)) in which a sample of a soil property is  
20 modelled as a sum of a linear combination of environmental covariates and a spa-  
21 tially autocorrelated (stochastic) residual, and prediction at unobserved locations is  
22 made by kriging. Geostatistical models are often used in soil mapping because they  
23 have several advantages ([Oliver, 1987](#)). First, a statistically sound model is assumed  
24 for spatial variation. This enables interpretation of the underlying physical processes  
25 conveyed (inferred) by the model. Secondly, spatial autocorrelation is explicitly mod-  
26 elled. This is relevant for environmental variables such as soil which vary from place  
27 to place, but exhibit correlation between places. Thirdly, an explicit measure of the  
28 uncertainty is associated with the prediction. In many circumstances such as in a  
29 decision making process, the prediction is not the only interest and uncertainty maps  
30 are required for the evaluation of the map quality or modelling risk.

31  
32 Geostatistical mapping of soil has, conversely, several limitations which have only  
33 partially been resolved in the current literature. To begin, the residuals are as-

34 sumed normally distributed, stationary (with constant mean and unit variance) and  
35 isotropic. Next, modelling the non-linear relation between a soil property or class  
36 and numerous cross-correlated covariates is not straightforward and introduces addi-  
37 tional challenges (e.g. many parameters have to be estimated). Finally, geostatistical  
38 models are computationally demanding if the sample size and/or the number of pre-  
39 diction locations are large (Cressie & Johannesson, 2008).

40  
41 As an alternative, machine learning (ML) emerged in the 1990s as a tool for  
42 spatial prediction and digital soil mapping (Lagacherie, 2008). Machine learning  
43 techniques refer to a large class of non-linear data-driven algorithms employed pri-  
44 marily for data mining and pattern recognition purposes, and now frequently used for  
45 regression and classification tasks in all fields of science. ML algorithms do not make  
46 an assumption of the observations' distribution, unlike geostatistical methods where  
47 transformation of the original observations is often required to satisfy the assump-  
48 tions. ML algorithms can also handle a large number of cross-correlated covariates  
49 as predictor.

50  
51 In parallel, there has been a tremendous increase in the production and availabil-  
52 ity of regional and global soil databases. For example, the Soil and Terrain Digital  
53 Database (SOTER, Oldeman & Van Engelen (1993)) made by FAO-UNESCO com-  
54 piled quantitative information on soil and terrain for different parts of the world while  
55 WoSIS is a harmonised database of more than 6 million geo-referenced soil records  
56 (Batjes et al., 2017). Additionally, numerous spatially exhaustive *scorpan* covariates  
57 are available at global scale for climate (Fick & Hijmans, 2017), elevation (Yamazaki  
58 et al., 2017), and parent material (Hartmann & Moosdorf, 2012). Further potential  
59 covariates are provided by remote sensing such as by the MODIS (Mira et al., 2015)  
60 satellite or Sentinel-2A hyperspectral sensor (Gascon et al., 2017). Soil mappers  
61 are now confronted with an increasing complexity in both soil data and covariates.  
62 Conventional regression techniques seem, to some extent, outdated to accommodate  
63 the increased complexity of soil datasets. This justifies the increasing use of machine  
64 learning algorithms for digital soil mapping.

65  
66 An essential distinction between conventional (statistical and geostatistical) mod-  
67 els and ML algorithms applied in DSM is their purpose. Machine learning algorithms  
68 mostly emphasize prediction accuracy whereas statistical models infer the process  
69 which generated the data through a pre-defined model of spatial variation. In the  
70 latter case, any interpretation is made in light of the model functions and the value of  
71 the covariates or input data. In machine learning, a predictive model is constructed

72 to predict a set of input values to output values using an error-minimization proce-  
73 dure. Since ML algorithms are not conditioned to follow any statistical assumptions,  
74 they often appear more accurate than conventional models. The exact path between  
75 input and output is ignored, and may not resemble an actual process described by  
76 the existing knowledge. In soil science, the explosion of articles using ML algorithms  
77 have made difficult to see the difference between model fitting and inference, and,  
78 as a result between data science and soil science. Research seems to be driven by  
79 the technique rather than by the hypothesis to be tested. This seems a poor bet  
80 for the advancement of knowledge since “almost invariably the technician’s skill is a  
81 solution looking for a problem” (Braben, 1985).

82  
83 In DSM, the use of ML algorithms has led to an increasing number of publica-  
84 tions where prediction (viz. mapping) of a soil property or class is the main interest.  
85 Many “easy-to-follow” software implementations have supported this increase. Dig-  
86 ital soil mapping, however, has unique characteristics which require adaptation of  
87 the ML algorithms. These features are for example, but not limited to, the inclusion  
88 of pedological knowledge in the ML algorithm, the accounting of spatial structure  
89 present in the raw soil data, or the need to increase our scientific understanding of  
90 the soil from a calibrated ML model.

91  
92 This article aims to review the development of ML applied to digital soil mapping  
93 by identifying key challenges and opportunities to solve them from the literature. In  
94 this review, we define ML as the computer assisted practice of using data-driven (and  
95 mostly non-linear) algorithms which resort to a large amount of calibration data to  
96 learn a pattern and make a prediction. We start by reviewing and summarizing the  
97 current use of machine learning in DSM. Based on this summary, we identify gaps in  
98 the knowledge and define areas in which adapting ML algorithms would be beneficial  
99 for their use in DSM. We propose solutions and a framework based on the literature  
100 from different fields of natural science. Finally, we define three core elements that  
101 should trigger soil scientists to move from model prediction to explanation of soil  
102 processes.

## 103 **2. A summary of applications**

### 104 *2.1. Extent, resolution, depths*

105 Table 1 summarizes some recent case studies of digital soil maps that have been  
106 produced using a ML algorithm. There is a large range of case studies, mapping soil  
107 properties or classes from the plot ( $<1 \text{ km}^2$ ) to the global ( $>10^7 \text{ km}^2$ ) scale. Most

108 studies in our literature review predict at a local to regional scale. The mean extent  
109 of the study area is 3,900 km<sup>2</sup>, but most (90%) studies consider a study area smaller  
110 than 650,000 km<sup>2</sup> (equivalent to the size of metropolitan France). Few studies map  
111 at plot or global scales. For example, Pouladi et al. (2019) make a quantitative map  
112 over a 10 ha (0.1 km<sup>2</sup>) field in Denmark while Hengl et al. (2017a) produce quanti-  
113 tative and categorical maps for the whole world.

114  
115 We found a clear correlation between the spatial extent of the study area and the  
116 grid spacing (i.e. the spacing between point predictions) at which the soil property or  
117 class is mapped: the larger the study area, the coarser the resolution. The resolution  
118 spans between 2 m × 2 m (Lacoste et al., 2014) to 1 km × 1 km for large, regional  
119 or continental study areas (e.g. Hengl et al., 2014). Most studies, however, map at a  
120 standard spatial resolution of 30, 90 or 250 m.

121  
122 While most of the studies (70%) predict a soil property or class for a single depth  
123 (topsoil), a number of studies accounts for the soil variation at multiple depths.  
124 Viscarra-Rossel et al. (2015) follow the GlobalSoilMap project specifications (Ar-  
125 rouays et al., 2014) to produce a quantitative three dimensional map of several soil  
126 properties for six depths intervals, namely 0-0.05 m, 0.05-0.15 m, 0.15-0.30 m, 0.30-  
127 0.60 m, 0.60-1.00 m and 1.00-2.00 m. Similar depth intervals are used in Mulder et al.  
128 (2016) and Adhikari et al. (2014) for soil organic prediction in France or Denmark,  
129 respectively. Several other studies (e.g. Grimm et al., 2008; Lacoste et al., 2014) use  
130 standard depth intervals for prediction, based on national mapping requirements or  
131 suitable for their specific case study.

## 132 *2.2. Sampling design, sample size and density*

133 The sampling design is the spatial location of the sampling units used to cal-  
134 ibrate or validate the ML algorithm. Most studies do not specify the sampling  
135 design used to generate the observations. It is speculated that the sample originates  
136 from multiple sources, e.g. legacy data, expert-based designs, and combination of  
137 several surveys, each of which had a different sampling design. When specified, non-  
138 probability sampling such as grid-based sampling designs are by far the most used  
139 (e.g. by Pahlavan-Rad & Akbarimoghaddam, 2018; Sergeev et al., 2019; Shariffar  
140 et al., 2019). Another non-probability sampling design is conditioned Latin Hyper-  
141 cube (cLHS), used to collect a sample in Lacoste et al. (2014); Brungard et al. (2015).  
142 Probability sampling is used in about one fourth of the studies. For example, simple  
143 random sampling is used in Tziachris et al. (2019), while a sample is collected based  
144 on stratified random sampling in Wiesmeier et al. (2011) using land use and topog-

145 raphy as stratifying variables.

146

147 In our literature review, we found that the sample size varies considerably be-  
148 tween studies. While the average sample is composed of 1,000 units, about one third  
149 of the studies use a sample with less than 150 units, mostly for local or small-scale  
150 regional areas. For example, [Blanco et al. \(2018\)](#) use a sample of size 47 for mapping  
151 soil water retention in a 93 km<sup>2</sup> area while [Massawe et al. \(2018\)](#) observed 33 soil  
152 profiles to calibrate a ML algorithm and to predict soil taxa over a 11,600 km<sup>2</sup> area.  
153 As expected, global studies have very large sample sizes. [Hengl et al. \(2017a\)](#) and  
154 [Ramcharan et al. \(2018\)](#) use a sample composed of more than 150,000 units to make  
155 soil property or class maps of the whole world, or of the United States, respectively.

156

157 When the sample size is associated to the extent of the study area, our review  
158 shows that large-scale studies have a very coarse sampling density. While the average  
159 sampling density in our literature is 0.24 units/km<sup>2</sup>, studies by [Beguin et al. \(2017\)](#)  
160 and [Wang et al. \(2017\)](#) have both a sampling density smaller than 3 units/10,000 km<sup>2</sup>  
161 for mapping soil properties in the rangelands of eastern Australia or in the Canadian  
162 boreal forests. Small-scale studies have, conversely, high sampling density. All studies  
163 with area size less than 50 km<sup>2</sup> have a sampling density larger than 7 units/km<sup>2</sup>.

### 164 *2.3. What is mapped?*

#### 165 *2.3.1. Quantitative variables*

166 ML algorithms have been successfully applied for quantitative mapping of vari-  
167 ous soil properties such as soil organic carbon concentration ([Henderson et al., 2005](#);  
168 [Bui et al., 2009](#); [Kheir et al., 2010b](#); [Dai et al., 2014](#); [Siewert, 2018](#); [Pouladi et al.,](#)  
169 [2019](#)) and associated stocks ([Grimm et al., 2008](#); [Adhikari et al., 2014](#); [Ließ et al.,](#)  
170 [2016](#); [Wang et al., 2017](#); [McNicol et al., 2019](#)), to map soil texture (viz. clay, silt  
171 and sand content) ([Ließ et al., 2012](#); [Akpa et al., 2014](#); [Vaysse & Lagacherie, 2015](#);  
172 [da Silva Chagas et al., 2016](#)), pH ([Dharumarajan et al., 2017](#)), or cation exchange  
173 capacity ([Forkuor et al., 2017](#)).

174

175 ML algorithms have also been applied to make maps of soil nutrients such as  
176 nitrogen ([Viscarra-Rossel et al., 2015](#); [Forkuor et al., 2017](#)), phosphorus ([Viscarra-](#)  
177 [Rossel et al., 2015](#); [Hengl et al., 2017b](#); [Song et al., 2018](#)), potassium, calcium or  
178 magnesium ([Hengl et al., 2017b](#)).

179

180 A number of studies have also predicted soil attributes and conditions with ma-  
181 chine learning such as bulk density ([Viscarra-Rossel et al., 2015](#)) or soil pollutants



182 (Kheir et al., 2010a). Wu et al. (2016) map soil background concentrations of arsenic  
183 in the Jiangxi Province in China. Taghizadeh-Mehrjardi et al. (2016) map soil salin-  
184 ity in Iran. Tajik et al. (2019) map soil invertebrate using environmental covariates  
185 in a deciduous forest ecosystem in northern Iran while Malone et al. (2009) map  
186 carbon storage and available water capacity in an area in eastern Australia.

### 187 2.3.2. Categorical variables

188 Compared with continuous soil property mapping, fewer studies apply ML to  
189 categorical variables. Digital mapping of soil classes using machine learning started  
190 in the 90s. Probably the first of its kind, Lagacherie & Holmes (1997) predict soil  
191 classes in a regional area while Cialella et al. (1997) predict soil drainage classes  
192 using remote sensing and elevation covariates. Behrens et al. (2005) map soil units  
193 in a 600 km<sup>2</sup> area of Western Germany. These studies have recently been completed  
194 by a number of publications comparing the maps predicted by a ML model to con-  
195 ventional soil maps (e.g. Zeraatpisheh et al., 2017). Scull et al. (2005); Brungard  
196 et al. (2015); Heung et al. (2016); Hounkpatin et al. (2018) employ machine learning  
197 to classify soil taxonomic units. Vermeulen & Van Niekerk (2017) map salt-affected  
198 areas in irrigation schemes in South Africa. Table 1 provides an additional summary  
199 of case studies.

200  
201 A special case of categorical mapping occurs when the map of soil class already  
202 exists but needs to be disaggregated. Bui et al. (1999) and Moran & Bui (2002)  
203 use a decision tree to disaggregate an existing map and obtain a realization of the  
204 disaggregated soil class distribution. With multiple realizations, the most probable  
205 soil class is obtained for a given location. This is further investigated by Hansen  
206 et al. (2009) to disaggregate a reconnaissance soil map using a binary decision tree.  
207 A similar approach with decision tree is used in Häring et al. (2012) to downscale  
208 soil types within existing map unit boundaries. More recently, Odgers et al. (2014)  
209 use ML to model and disaggregate soil classes and report the probability associated  
210 to each soil class at a given location in the area of interest. A growing number of  
211 publication exploits the DSMART approach proposed by Odgers et al. (2014) (e.g.  
212 Holmes et al., 2014; Vincent et al., 2018; Ellili et al., 2019).

### 213 2.4. Covariates

214 Environmental covariates are used as predictors in ML algorithms. They are  
215 supposed to explain part of the physical and chemical process governing soil spa-  
216 tial variation. Most studies use about 20 covariates. Only a few use less than five  
217 (e.g. Dai et al., 2014; Padarian et al., 2019) while other use more than 100 (e.g Hengl

218 [et al., 2017a](#); [Ramcharan et al., 2018](#)). Since the covariates represent soil forming fac-  
219 tors, numerous studies (e.g. [Viscarra-Rossel & Chen, 2011](#); [Wang et al., 2018](#); [Gomes](#)  
220 [et al., 2019](#); [Szatmári & Pásztor, 2019](#)) logically select the covariates to represent the  
221 key factors of the *scorpan* model of soil spatial variation. The most common ones  
222 are existing soil property or class maps, (long-term) average annual precipitation  
223 and temperature, remote sensing images (e.g. SPOT satellite images or vegetation  
224 indices derived from satellite images), elevation, terrain attributes (e.g. slope, local  
225 curvature, topographic wetness index) and existing geological maps.

226  
227 Covariates representing *scorpan* factor of soil variation might not be available or  
228 easily obtainable in all case studies. In some cases, covariates are chosen based on  
229 expert knowledge. A number of studies therefore calibrate machine learning algo-  
230 rithms using sets of climatic variables, remote sensing images or terrain attributes  
231 only, or a combination of them. For example, [Mansuy et al. \(2014\)](#) use a set of  
232 eight climatic and eight terrain attribute variables to map C, N and soil texture in a  
233 large area in Canada. [Shariffar et al. \(2019\)](#) use six terrain attributes as predictors.  
234 There are chosen from a large set of environmental covariates using knowledge on  
235 the expected relationship between the covariate and the soil property to be mapped.  
236 We note that a few studies (e.g. [Hengl et al., 2018](#); [Miller et al., 2015a](#)) consider  
237 that if a sufficiently large ( $> 100$ ) number of covariates is used for calibration, the  
238 machine learning algorithm learns a representation of the spatial pattern and pre-  
239 dicta a realistic spatial pattern. This large amount of covariates relies mostly on  
240 remote sensing images, e.g. MODIS land products (long-term averages, several near-  
241 or mid-infrared bands) or Landsat products (near-, short-wave near-infrared, or  $\gamma$   
242 radiometric bands, bare ground images).

243  
244 A few studies account for the multi-scale variation of the environmental covari-  
245 ates. In other words, terrain derivatives may well be aggregated to account for  
246 physical processes in soil that are not visible are finer scale. Examples of studies us-  
247 ing multi-scale covariates for mapping with machine learning algorithms are [Behrens](#)  
248 [et al. \(2010\)](#), [Miller et al. \(2015b\)](#) or more recently [Behrens et al. \(2018a\)](#). [Miller](#)  
249 [et al. \(2015b\)](#), for example, use a total of 412 covariates, several of which are derived  
250 from the aggregation of terrain attributes from a fine (i.e. a grid cell size of  $2 \text{ m} \times$   
251  $2 \text{ m}$ ) elevation map.

252  
253 A growing number of studies have advocated the use of spatial surrogate covari-  
254 ates as an indicator of spatial position in the *scorpan* model of soil variation The  
255 most common surrogate is the use of geographical coordinates (easting and northing)

256 as covariates in the model. Maps of distances from observation locations, or group  
257 of locations, have been used by [Hengl et al. \(2018\)](#). They are categorized into Eu-  
258 clidean, downslopes or “resistance” distances. More recently, [Behrens et al. \(2018b\)](#)  
259 use Euclidean distance fields, which are maps of distance from reference locations in  
260 the study area such as a corner or a center.

### 261 *2.5. Covariate selection*

262 Covariate (*aka* feature) selection aims at reducing the number of covariates used  
263 to calibrate the machine learning models. While most ML models are robust to mul-  
264 ticolinearity between covariates, there are several reasons for selecting a subset of  
265 covariates to calibrate the model. Some of them are: (i) to calibrate the ML model  
266 faster, (ii) to reduce complexity, (iii) to increase the prediction accuracy or (iv) to  
267 prevent over-fitting of the ML model, i.e. to prevent poor prediction accuracy on  
268 unseen data. In our literature review, about one third of the studies apply covariate  
269 selection. Two main categories of covariate selection techniques are found. The first  
270 applies the covariate selection as a pre-processing step, i.e. before calibrating the ML  
271 model. This is the case in [Zhu et al. \(2019\)](#); [Hamzhepour et al. \(2019\)](#); [Zeraatpisheh  
272 et al. \(2019\)](#). [Hamzhepour et al. \(2019\)](#) select the covariates to be used in calibration  
273 by computing the Pearson’s  $r$  correlation coefficient between the covariates, and by  
274 discarding the ones that were highly correlated, while [Mosleh et al. \(2016\)](#) select the  
275 covariates based on the Pearson  $r$  correlation coefficient between the soil property  
276 values and the covariates, and select a subset of covariates which are strongly corre-  
277 lated with the property. The second type of covariate selection are called “wrapper”  
278 methods and rely on the inference made by a calibrated ML model to determine  
279 whether covariates are important. By re-calibrating a ML model several times, each  
280 time removing the least important covariate, one may expect to reduce considerably  
281 the overall number of covariates with little or no decrease in model prediction accu-  
282 racy. Examples on the use of “wrapper” methods are found in [Taghizadeh-mehrjardi  
283 et al. \(2016\)](#); [Shi et al. \(2018\)](#); [Rudiyanto et al. \(2018\)](#); [Tajik et al. \(2019\)](#) or [Gomes  
284 et al. \(2019\)](#). The most used of “wrapper” methods is an optimization algorithm  
285 called recursive feature elimination.

### 287 *2.6. Machine learning models*

288 A large number of ML algorithms and their variants have been used in the DSM  
289 literature. For quantitative mapping, tree-based algorithms are the most popu-  
290 lar ones, the simplest version of which is the regression tree, used for example by  
291 [Taghizadeh-Mehrjardi et al. \(2016\)](#). Regression tree is known to be sensitive to the

292 calibration sample. To solve this problem, the bagging (bootstrap and aggregating)  
293 procedure (Breiman, 2017) has been introduced in random forest (RF). Our litera-  
294 ture review shows that RF is currently the most popular ML algorithm for regression  
295 purposes. Example of case studies using RF for mapping are Tziachris et al. (2019);  
296 Vaysse & Lagacherie (2015); Forkuor et al. (2017); Dharumarajan et al. (2017); Liu  
297 et al. (2019). More recently, Vaysse & Lagacherie (2017) introduced a variant of  
298 random forest, called quantile regression forest, as a method to map the uncertainty  
299 associated with the prediction of the soil property. Another tree-based method is  
300 cubist, employed in about 10% of the reviewed literature (e.g. by Mulder et al., 2016;  
301 Viscarra-Rossel et al., 2015; Miller et al., 2015a). A few studies (less than five) use  
302 boosted regression tree (Yang et al., 2016; Beguin et al., 2017). In addition, a num-  
303 ber of studies use neural networks (Lamichhane et al., 2019) algorithms (Aitkenhead  
304 & Coull, 2016; Guevara et al., 2018), such as artificial neural networks (Dai et al.,  
305 2014). A relatively small number of studies use alternative algorithms such as sup-  
306 port vector machines (Guevara et al., 2018),  $k$ -nearest neighbours (Mansuy et al.,  
307 2014) or generalized boosted regression (Tziachris et al., 2019; Gomes et al., 2019).

308  
309 For classification purposes, tree-based algorithms are also the most popular ones.  
310 About 80% of the case studies used at least one tree-based algorithm such as regres-  
311 sion tree (e.g. Taghizadeh-Mehrjardi et al., 2019b; Heung et al., 2016), random forest  
312 (e.g. Häring et al., 2012) or boosted regression tree (e.g. Lorenzetti et al., 2015). Al-  
313 ternatively, gradient boosting is used by Hengl et al. (2017a),  $k$ -nearest neighbors by  
314 Vermeulen & Van Niekerk (2017) and compared to support vector machines. The  
315 latter algorithm is also used in Taghizadeh-Mehrjardi et al. (2019b). Neural networks  
316 is also popular and used in Behrens et al. (2005); Heung et al. (2016).

317  
318 Recent studies have proposed to use model ensemble techniques to improve the  
319 predicted map of several individual models in terms of accuracy. Taghizadeh-Mehrjardi  
320 et al. (2019a) combined seven ML model predictions for soil class mapping in a case  
321 study in Iran while Song et al. (2020) implemented a weighted ensemble learning  
322 model to map soil organic carbon in consideration of pedoclimatic zones in China.  
323 Ensembles are also considered in Hengl et al. (2017a) for global soil mapping.

### 324 2.7. Parameter tuning

325 The performance of a machine learning model is impacted by the values of its  
326 model parameters. While most ML would perform well on default tuning parameter  
327 values, almost half of the studies perform a search to find optimal values. Padar-  
328 ian et al. (2019) manually decide the artificial neural network neurons number for

329 each layer of the network. This manual search is automated by a so-called grid-  
330 search process. This is by far the most used technique for parameter tuning. In a  
331 grid-search process, a number of parameter values are evaluated based on the model  
332 prediction error. The process is computationally intensive (the ML model must be  
333 calibrated for each parameter set proposal). Examples of studies using a grid-search  
334 to find ML parameter values are [Ottoy et al. \(2017\)](#); [Taghizadeh-Mehrjardi et al. \(2016\)](#);  
335 [Pahlavan-Rad & Akbarimoghaddam \(2018\)](#); [Sergeev et al. \(2019\)](#); [Forkuor et al. \(2017\)](#);  
336 [Ramcharan et al. \(2018\)](#). An alternative to the grid search is to apply  
337 an optimization algorithm, such as the particle swarm method, to find optimal pa-  
338 rameter values. For example, [Wu et al. \(2016\)](#) compare two genetic algorithms and  
339 a grid search process to find the ML parameters. Recently, [Wadoux et al. \(2019b\)](#)  
340 use Bayesian optimization to optimize the number of layers, the neuron number,  
341 the learning rate and the batch size of an artificial neural network for mapping soil  
342 organic carbon.

### 343 *2.8. Validation and uncertainty quantification*

344 In our literature review, all studies compute at least one validation statistic to  
345 assess the quality of the prediction. A list of validation statistics is provided in  
346 [Table 1](#). About 30% of the studies obtain the validation statistics through cross-  
347 validation, while 30% through data-splitting. The remaining studies either repeat  
348 data-splitting several times, validate through visual examination or use a grid-based  
349 sampling design. Only two studies collect an additional probability sample for vali-  
350 dation ([Subburayalu & Slater, 2013](#); [Lacoste et al., 2014](#)).

351  
352 In addition to the validation statistics, about 30% of the studies quantify the  
353 uncertainty associated with the prediction. These studies report confidence inter-  
354 val, obtained by bootstrapping the original set of observations (e.g. [Chen et al., 2019](#);  
355 [Padarian et al., 2019](#); [Hamzehpour et al., 2019](#)). A few studies use the kriging  
356 variance computed on the residuals of a trend obtained by predicting with a ML al-  
357 gorithm (e.g. [Koch et al., 2019](#)), or a combination of bootstrap and kriging variance  
358 (e.g. [Viscarra-Rossel et al., 2015](#)). In three studies, prediction intervals are obtained  
359 through the quantile regression forest. [Wadoux \(2019b\)](#) obtain the prediction inter-  
360 vals following a two-step procedure called mean plus variance estimate for mapping  
361 several soil properties using an artificial neural network.

Table 1: Non-exhaustive list with summary of case studies in which machine learning algorithms are used for digital soil mapping.

Spatial extent <sup>1</sup>	Sample size	Sampling design	Number of covariates	Machine learning model <sup>2</sup>	Covariate selection	Parameter tuning	Validation statistics <sup>3</sup>	Uncertainty quantification	Reference
<b>Quantitative maps</b>									
Plot	285	grid-based	19	cubist, RF	no	no	R <sup>2</sup> , RMSE	no	Pouladi et al. (2019)
Local	47	stratified random	41	RF	yes	yes	RMSE, IQR	yes	Blanco et al. (2018)
Local	70	cLHS	19	cubist	no	no	MAE, RMSE, R <sup>2</sup> , CCC	yes	Lacoste et al. (2014)
Local	75	grid-based	9	ANN	no	no	R <sup>2</sup> , MSE	no	Kalambukattu et al. (2018)
Local	98	varied sources	173	RF	yes	no	RMSE, R <sup>2</sup>	no	Shi et al. (2018)
Local	116	simple random	20	RF	no	no	R <sup>2</sup> , RMSE, CCC	no	Dharumarajan et al. (2017)
Local	117	not specified	13	GBM	yes	yes	R <sup>2</sup> , RMSE, MAE	yes	Hamzhepour et al. (2019)
Local	117	not specified	412	cubist	yes	no	ME, MAE, R <sup>2</sup> , R <sup>2</sup> <sub>adj</sub>	no	Miller et al. (2015b)
Local	120	stratified random	not specified	RF	no	no	ME, RMSE, R <sup>2</sup> , MSE	no	Wiesmeier et al. (2011)
Local	120	stratified random	22	ANN, BRT	yes	yes	R <sup>2</sup> , RMSE, ME	no	Mosleh et al. (2016)
Local	137	systematic random	20	ANN, GEP	yes	yes	RMSE, R <sup>2</sup> , MBE	no	Mahmoudabadi et al. (2017)
Local	138	not specified	15	RF	yes	no	RMSE, R <sup>2</sup> , CCC	no	Zhu et al. (2019)
Local	150	grid-based	not specified	ANN	no	yes	correlation coefficient, R <sup>2</sup> , RMSE, Willmott's index of agreement, RPIQ	no	Sergeev et al. (2019)
Local	151	not specified	not specified		no	no	R <sup>2</sup> , NRMSD	no	Kovačević et al. (2010)
Local	153	grid-based	26	RF	yes	no	RMSE, R <sup>2</sup>	no	Tajik et al. (2019)
Local	159/34	not specified	37	RF, cubist, QRF, NN, avNNNet, ctree, evtree, GBM, <i>k</i> -NN, RT, SVM	yes	no	R <sup>2</sup> , RMSE, MAE, MARE	yes	Rudiyanto et al. (2018)
Local	165	stratified random	18	RF	no	yes	MSE, NMSE	no	Grimm et al. (2008)
Local	173 profiles	cLHS	19	Rf	no	no	ME, RMSE, R <sup>2</sup>	no	Taghizadeh-Mehrjardi et al. (2014)
Local	188 profiles	cLHS	16	ANN, SVR, <i>k</i> -NN, RF, RT	no	yes	RMSE, CCC	no	Taghizadeh-Mehrjardi et al. (2016)
Local	234	not specified	410	cubist	yes	no	MAE, R <sup>2</sup>	yes	Miller et al. (2015a)

<sup>1</sup>Plot: 0-1 km<sup>2</sup>; Local: > 1 km<sup>2</sup>-10<sup>4</sup> km<sup>2</sup>; Regional: > 10<sup>4</sup> km<sup>2</sup>-10<sup>7</sup> km<sup>2</sup>; Global: > 10<sup>7</sup> km<sup>2</sup>.

<sup>2</sup>RF: random forest; ANN: artificial neural networks; CNN: convolutional neural networks; GBM: gradient boosting machine; BRT: boosted regression tree; GEP: gene expression programming; QRF: quantile regression forest; avNNNet: neural networks using model averaging; ctree: conditional inference trees; evtree: evolutionary algorithm for classification and regression tree; NN: neural networks; GBM: generalized boosted regression; *k*-NN: *k*-nearest neighbors; RT: regression tree; SVM: support vector machine; MARS: multivariate adaptive regression splines; SGB: stochastic gradient boosting; CART: classification and regression tree; NSC: nearest shrunken centroids; CT: classification tree; BCT: bagged classification tree; DT: decision tree; LMT: logistic model tree; EGB: extreme gradient boosting.

<sup>3</sup>R<sup>2</sup>: coefficient of determination; R<sup>2</sup><sub>adj</sub>: adjusted coefficient of determination; RMSE: root mean square error; IQR: interquartile range; MAE: mean absolute error; CCC: Lin's concordance correlation coefficient; MSE: mean square error; ME: mean error; MBE: mean bias error; RPIQ: ratio of performance to interquartile distance; NRMSD: normalized root mean squared deviation; MARE: median absolute relative error; NMSE: normalized mean square error; sMAPE: symmetric mean absolute percentage error; SS: skill score; RMSD: minimum root mean square deviation; RPD: residual prediction deviation; SDE: standard deviation of the error; EC: overall ratio; OA: overall accuracy; PA: producer accuracy; UA: user accuracy; AUROC: area under receiver operating characteristic curve; AUC: area under the curve.



Local	330 profiles	not specified	12	BRT, ANN, least-square SVM	no	yes	$R^2$ , $R^2_{adj}$ , RMSE, relative RMSE	no	Ottoy et al. (2017)
Local	330	simple random	10	RF, GBM	no	yes	ME, MAE, RMSE, $R^2$		Tziachris et al. (2019)
Local	334	cLHS	16	cubist, RF, RT	yes	no	$R^2$ , RMSE	no	Zeraatpisheh et al. (2019)
Local	342/321	-	14	MARS, SVR, RF, Cubist, NN	-	yes	$R^2$	no	Behrens et al. (2018b)
Local	399	not specified	12	RF	no	no	$R^2$ , RMSE	no	da Silva Chagas et al. (2016)
Local	440	varied sources	19	RF, SVM, ANN	no	yes	RMSE, ME	no	Were et al. (2015)
Local	460	grid-based	21	RF	no	yes	ME, MAE, RMSE	no	Pahlavan-Rad & Akbarimoghadam (2018)
Local	568	simple random	26	QRF	no	no	$R^2$ , RMSE, range-normalized RMSE, Moran's I	yes	Kirkwood et al. (2016)
Local	1104	expert	29	RF, SVM, SGB	no	yes	RMSE, sMAPE	no	Forkuor et al. (2017)
Local	$\leq 1052/2050/2379$	varied sources	300-500	BRT, RF	yes	yes	bias, RMSE, SS, $R^2$	no	Nussbaum et al. (2018)
Local	2388	varied sources	3	CNN, RF	no	yes	ME, RMSE, $R^2$ , CCC	no	Wadoux et al. (2019b)
Regional	not specified	not specified	20	cubist	no	no	$R^2$ , RMSE, bias, CCC	yes	Mulder et al. (2016)
Regional	125 profiles	purposive	12	BRT, RF	no	no	MAE, RMSE, $R^2$ , CCC	no	Yang et al. (2016)
Regional	244	grid-based	4	ANN	no	yes	ME, MAE, RMSE, CCC	no	Dai et al. (2014)
Regional	339/961	varied sources	40	QRF	no	no	$R^2$ , RMSE	yes	Nauman & Duniway (2019)
Regional	485 profiles	not specified	5	CNN	no	yes	$R^2$ , RMSE	yes	Padarian et al. (2019)
Regional	500	not specified	12	RF, BRT	yes	no	$R^2$ , RMSE	no	Beguín et al. (2017)
Regional	528	subset from a systematic grid	18	$k$ -NN	yes	no	RMSE, $R^2$ , Bias, coefficient of variance	no	Mansuy et al. (2014)
Regional	705	simple random	16	RF, BRT, SVM	yes	yes	$R^2$ , MAE, RMSE, CCC	yes	Wang et al. (2018)
Regional	978 profiles	not specified	24	RF	no	no	$R^2$ , ME, RMSE, CCC	no	Akpa et al. (2014)
Regional	1,014	stratified random	327	CART, BRT, BRT, RF, SVM	yes	no	$R^2$ , RMSD, RPD, RPIQ	no	Keskin et al. (2019)
Regional	1,134	not specified	81	NN	no	no	$R^2$ , ME, MAE, RMSE	no	Aitkenhead & Coull (2016)
Regional	1,300 profiles	not specified	6	RF	no	no	CCC, RMSE	yes	McNicol et al. (2019)
Regional	1,626	not specified	40	SVM	no	yes	$R^2$ , MSE	no	Wu et al. (2016)
Regional	2,024 profiles	legacy data	16	QRF	no	no	ME, RMSE, $R^2$ , accuracy plot	yes	Vaysse & Lagacherie (2017)
Regional	2,024 profiles	legacy data	16		no	yes	MSE, $R^2$	no	Vaysse & Lagacherie (2015)
Regional	2,943	two-stage systematic	37	CNN, RF	no	yes	ME, RMSE, $R^2$ , CCC	yes	Wadoux (2019b)
Regional	4,859	not specified	26	QRF	no	no	ME, RMSE, accuracy plot	yes	Szatmári et al. (2019)
Regional	4,859	not specified	32	QRF	no	no	ME, RMSE, accuracy plot	yes	Szatmári & Pásztor (2019)
Regional	5,386	varied sources	6	cubist, SVM	no	no	$R^2$ , MSE, CCC		Somarathna et al. (2016)
Regional	13,000	not specified	18	RF	no	no	$R^2$	yes	Koch et al. (2019)
Regional	19,790	two-stage systematic	197	RF	no	no	ME	no	Wadoux et al. (2019a)
Regional	37,693	legacy soil data	74	RF, Cubist, SVM	yes	yes	$R^2$ , RMSE, MAE	yes	Gomes et al. (2019)
Regional - Global	2,268-27,262	varied sources	34	cubist	no	yes	CCC, RMSE, SDE, ME	yes	Viscarra-Rossel et al. (2015)

Regional - Global	366,034	varied sources	>200	RF, GBM	no	yes	R <sup>2</sup> , ME, RMSE, MAE	yes	Ramcharan et al. (2018)
Global	11,268	legacy soil data	118	SVM, kernel weighted NN, RF	yes	no	EC, RMSE, R <sup>2</sup>	yes	Guevara et al. (2018)
Global	150,000	legacy soil data	> 200	RF, GBM	no	yes	R <sup>2</sup>	no	Hengl et al. (2017a)
<b>Categorical maps</b>									
Local	-	not specified	125	ANN	no	no	Accuracy, recall, precision	no	Behrens et al. (2005)
Local	33 profiles	not specified	16	RF, J48	no	no	not specified	no	Massawe et al. (2018)
Local	103/297/ 57	cLHS	130	<i>k</i> -NN, NSC, CT, BCT, RF, linear SVM, radial-basis SVM, NN, ANN	yes	yes	Kappa analysis, Brier scores, visual inspection, confusion index	no	Brungard et al. (2015)
Local	125 profiles	cLHS	17	RF	no	no	map purity, Cohen's kappa, Shannon entropy index, relative purity, relative diversity	no	Zeraatpisheh et al. (2017)
Local	151	not specified	not specified	SVM	no	no	NRMSD, micro averaged F1 measure, kappa statistics	no	Kovačević et al. (2010)
Local	175, 63 profiles	varied sources	27	<i>k</i> -NN, SVM, DT, RF	no	no	OA, PA, UA, kappa coefficient, AUROC	no	Vermeulen & Van Niekerk (2017)
Local	452 profiles	regular grid	6	DT, RF	yes	no	OA, UA, PA, Kappa coefficient of agreement	no	Shariffar et al. (2019)
Local	917	grid-based	33	RF	yes	no	Kappa index	no	Hounkpatin et al. (2018)
Local	3,121	by-polygon, equal-class, area-weighted, and area-weighted with random over sampling	20	CART, CART with bagging, RF, <i>k</i> -NN, NSC, ANN, LMT, SVM	no	yes	overall agreement, quantity disagreement, allocation disagreement, total disagreement	no	Heung et al. (2016)
Regional	89,323	random sampling	26	<i>k</i> -NN, RF	yes	no	recall, accuracy	no	Subburayalu & Slater (2013)
Regional	366,034	varied sources	>200	RF, GBM	no	yes	OA, regional dataset	yes	Ramcharan et al. (2018)
Regional	7,664 profiles	varied sources	110	DT, RF, EGB, SVM, <i>k</i> -NN	yes	no	OA, precision, recall, F-score, K-index	no	Taghizadeh-Mehrjardi et al. (2019b)
Regional	9,924	not specified	23	RF	yes	no	error matrix	no	Håring et al. (2012)
Global	150,000	legacy data	>200	RF, GBM	no	yes	map purity, weighted kappa metrics, AUC, True positive rate, scaled Shannon's entropy index	no	Hengl et al. (2017a)



### 362 3. Challenges and opportunities

363 Based on the review, here we identify some knowledge gaps and challenges in  
364 the current use of ML algorithms for DSM. We will outline some opportunities for  
365 research.

#### 366 3.1. Sampling

367 Despite abundant evidence that the sampling design and sample size play a key  
368 role in the resulting map accuracy (De Gruijter et al., 2006), sampling designs suit-  
369 able for mapping with machine learning are yet to be uncovered. The impact of the  
370 sample size for mapping with ML is discussed in Somarathna et al. (2017) where  
371 the efficiency of several ML algorithms are compared for the spatial prediction of  
372 soil carbon. The study shows that having a sufficiently large sample size is more  
373 important than choosing a sophisticated ML algorithm, and that when the sample  
374 size is small, it is best to use simple models. About sampling designs, Brus (2019)  
375 speculates that machine learning algorithms would benefit from a spread of the sam-  
376 pling units in the feature (covariate) space, and suggests the use of feature space  
377 coverage sampling (FSCS) using  $k$ -means clustering or conditioned Latin Hypercube  
378 sampling (cLHS). Both sampling designs aim at covering the space spanned by the  
379 covariates, but in different ways. Experimental results are provided by Wadoux et al.  
380 (2019a) in a study comparing five sampling designs (viz. simple random sampling,  
381 cLHS, spatial coverage sampling (SCS), FSCS and a design optimized in terms of  
382 mean square error) for soil property mapping with random forest. The results show  
383 large differences in mapping accuracy between the designs, and that a FSCS de-  
384 sign optimized in the most important covariates of the random forest model had  
385 the closest match to an optimized design. By performing further diagnostics, the  
386 study concludes that RF does not benefit from a uniform spread of the units in the  
387 geographic/feature space, nor from reproducing the marginal distribution of the co-  
388 variates (as it is done in cLHS). These results apply for RF but there is a need to  
389 further investigate sampling designs for other machine learning algorithms. While  
390 most studies in our literature review (Table 1) use a grid-based sampling or cLHS,  
391 there is now evidence that most conventional sampling designs (e.g. spatial coverage  
392 sampling) are not effective for the purpose of mapping with machine learning.

393  
394 To discover what makes a good design for mapping with machine learning, one  
395 should ideally derive optimal designs. More importantly, one should investigate the  
396 characteristics of these designs, so that future research can generate simple designs  
397 that resemble the optimal ones (Wadoux, 2019a). It is likely that optimal designs  
398 differ between machine learning algorithms. We speculate that a somewhat uniform

399 spread in the feature (i.e. covariate) space remains important for all ML algorithms  
400 since they all link the covariates and the sample values in a non-linear way, but that  
401 additional considerations might outweigh or overtake this uniform spread. An ex-  
402 ample of optimal design is given by the studies of [Pozdnoukhov & Kanevski \(2006\)](#)  
403 and [Tuia et al. \(2013\)](#) where the sampling configurations are optimized with active  
404 learning for mapping with support vector machines. In the first study, the selected  
405 sampling units are the most beneficial for the algorithm, avoiding mis-classification  
406 between temperature below or above 20°Cs (categorical mapping) by becoming sup-  
407 port vectors. In [Tuia et al. \(2013\)](#), a similar methodology is adopted and tested in  
408 three case studies to subsample an existing sample for quantitative mapping, to add  
409 optimally new sampling units in a continuous map or to define suitable areas for  
410 sampling. In all case studies, the authors obtained a design optimal for the purpose  
411 of mapping with support vectors machine. They conclude that while a sampling  
412 design can be representative of the geographical space, the latter can be judged un-  
413 representative if other dimensions are considered. These results encourage the use  
414 of new methods for sampling design optimization such as active learning. Active  
415 learning is a model-based sequential re-design algorithm. In active learning, the  
416 objective function (e.g. the spatially averaged prediction uncertainty) is explicitly  
417 quantified and used to define the additional sampling units that are the most ben-  
418 efiticial for the model (e.g. the boundary between two classes). In this sense, active  
419 learning is similar to optimization with spatial simulated annealing routinely used  
420 in geostatistical sampling design optimization. Besides the optimization algorithms,  
421 a set of objective functions needs to be tested. [MacKay \(1992\)](#) defined an objective  
422 function that searches for the optimal units in the space spanned by the predictors  
423 (i.e. covariates) for prediction using a neural network algorithm. Taking the latter  
424 considerations and testing active learning for sampling design optimization would  
425 certainly make a valuable contribution to digital soil mapping research.

### 426 *3.2. Resampling*

427 Regional and global scale studies almost invariably use legacy soil data ([Stumpf  
428 et al., 2016](#)). Legacy soil samples provide valuable information on soil classes and  
429 properties but are often highly clustered in areas of specific interest. In modelling  
430 with machine learning, it is assumed that the sample is composed of independent  
431 and identically distributed sampling units whereas soil observations within an area  
432 typically exhibit spatial autocorrelation (i.e. close observations are more similar than  
433 remote ones) This has important implications in terms of sampling, resampling of  
434 the observations and validation of the models. A ML algorithm calibrated with a  
435 spatially clustered sample may lead to biased predictions over the area because of

436 the over-representation in the calibration process of regions of high sampling density.  
437 Despite being critical, this has yet been disregarded in DSM studies. In geostatistics,  
438 spatial declustering has been applied to reduce the effect of clustered data in the  
439 calculation of experimental variogram (Marchant et al., 2013). One form, called cell  
440 declustering involves overlaying a grid over the area and assigning a weight to the  
441 sampling units based on the inverse of the number of units in the cell. In ecology, a  
442 first attempt was made by Bel et al. (2005) and later Bel et al. (2009) to decluster the  
443 sampling units used in the calibration of a CART model. In Bel et al. (2005), weights  
444 are given to the sampling units, where the weights are obtained from a kriging of  
445 the spatial mean. Bel et al. (2009) elaborate a more complex procedure in which  
446 all quantities involved in the CART algorithm (e.g. the proportion of leaves) have  
447 a spatial estimate. This has been further considered by Stojanova et al. (2013) for  
448 both categorical and quantitative mapping of ecological variables. Illés et al. (2019)  
449 applied polygonal declustering technique to spatially clustered samples by assigning  
450 weights on the units based on Voronoi’s area proportion.

451  
452 We point out that the clustering may also occur in the feature (i.e. covariates)  
453 space and speculate that this may also affect the prediction if most units are clustered  
454 at some specific areas of the feature space. For example, a model trained to predict  
455 organic carbon in a mountainous area will exhibit biased prediction if most sampling  
456 units originate from valley, and that elevation is used as predictors. Similar to Bel  
457 et al. (2005), weights can be assigned to the units so down-weight the importance of  
458 over-sampled areas in the feature space. An example method is provided by Carré  
459 et al. (2007). The authors assume that a good sample have a uniform spread in the  
460 feature space and thus covers all strata of a hypercube based on the covariates. A  
461 weight is assigned to each unit in the sample based on the density of the units in each  
462 stratum. The larger the density within the stratum, the smaller the weight assigned  
463 to a single unit.

464  
465 The nature of the legacy soil data in categorical mapping also poses additional  
466 challenges. ML algorithms for categorical mapping rely on balanced sets of units.  
467 In other words, all classes shall comprise a comparable number of sampling units.  
468 Legacy soil samples are considered imbalanced in that all classes are not represented  
469 equally. Most ML algorithms are calibrated by maximizing the average (classification)  
470 accuracy on an independent validation sample. This often results in very low  
471 predictive accuracy for under-sampled classes, and models biased toward the over-  
472 sampled classes (He & Garcia, 2009). In the ML literature, several approaches have  
473 been developed to handle class imbalanced samples. At the higher level, one may

474 distinguish between cost function and resampling based approaches. In the first ap-  
475 proach, the model is penalized for miss-classification to under-represented classes.  
476 This stems from the calibration of ML algorithms, which minimize a loss function  
477 to find optimal parameter values (e.g. in neural networks). In the second approach,  
478 resampling of the sample is performed by either adding units in the under-sampled  
479 class, removing units from the over-sampled class, or a mix of the two. The second  
480 approach has been recently been applied in soil mapping studies, in particular by  
481 [Heung et al. \(2016\)](#) and [Shariffar et al. \(2019\)](#). [Taghizadeh-Mehrjardi et al. \(2019b\)](#)  
482 tested eight resampling approaches and their effect on the prediction accuracy of five  
483 ML algorithms in two large-scale case studies. However, to date resampling tech-  
484 niques are applied the same way as in other disciplines while soil data often presents  
485 spatial autocorrelation which may impact the resampling strategies. This has not yet  
486 been investigated in the literature. The integration of resampling strategies within  
487 a general framework for mapping with ML is provided in [Fig. 3](#).

### 488 *3.3. Accounting for spatial information*

489 Machine learning algorithms do not account for spatial autocorrelation contained  
490 in the raw soil data, unless explicitly specified. [Sinha et al. \(2019\)](#) have tested ran-  
491 dom forest for different scenarios of spatial autocorrelation in the observations and  
492 confirmed that the presence of spatial autocorrelation leads to high variance of the  
493 residuals. ML algorithms accounting for autocorrelated observations have recently  
494 been formulated, such as geographical random forest ([Georganos et al., 2019](#)), or  
495 spatial ensemble techniques ([Jiang et al., 2017](#)). The two methods boil down to  
496 geographically weighted regression by fitting spatially local sub-models using only  
497 neighbouring observations. [Jiang et al. \(2017\)](#) decomposed the area into geographic  
498 disjoint sub-areas, and fitted a local model in each sub-area. [Georganos et al. \(2019\)](#)  
499 fitted a sub-model to each observation using random forest, accounting for both non-  
500 stationarity and spatial autocorrelation.

501  
502 Applying a non-spatial model for digital soil mapping is not a problem in itself.  
503 This is corroborated by the definition of DSM given in [Lagacherie & McBratney](#)  
504 [\(2006\)](#), which gives provision for mapping using “non-spatial soil inference systems”.  
505 In theory, if one includes all relevant environmental variables to model the soil prop-  
506 erty or class, there should be no spatial autocorrelation in the residuals of the fitted  
507 models. If this happens, some important predictors are likely to be missing. More  
508 importantly, this also means that predictions made by the ML algorithm might be  
509 biased or the model underfitted because this is a violation of the assumption of inde-  
510 pendence between data points that is implicitly assumed. [Kühn & Dormann \(2012\)](#)

511 recommend mapping the spatial distribution of the residual autocorrelation to facili-  
512 tate the identification of a missing spatial process. In some cases, a map of residuals  
513 exhibits a clear pattern (e.g. increasing residuals with distance from the river) and  
514 might help to generate a new hypothesis or to refine the existing model (see Fig. 3).  
515

516 Despite the availability of datasets and care made during modelling, residual au-  
517 tocorrelation is still likely to occur. Several authors have advocated the use of spatial  
518 surrogate covariates as an indicator of spatial position in the *scorpan* model of soil  
519 variation or to account for spatial autocorrelation contained in the data. The most  
520 common surrogate is the use of geographical coordinates (easting and northing) as  
521 covariate in the model. This has led to maps with visible artefacts, in particular  
522 when used in combination with tree-based algorithms. Alternatively, maps of dis-  
523 tances from observation locations, or a group of locations, have been proposed by  
524 [Hengl et al. \(2018\)](#). They are categorized into Euclidean, downslopes or “resistance”  
525 distances. Maps of distance to observation locations generally have no direct mean-  
526 ing in terms of soil process over an area (e.g. distance from the river). [Behrens et al.](#)  
527 [\(2018b\)](#) propose to use Euclidean distance fields, which are maps of distance from  
528 reference locations in the study area such as the corner or the centre. The studies  
529 using distance maps as covariates have shown for several case studies an important  
530 reduction of the residual autocorrelation, when compared to a model without dis-  
531 tance maps in the set of covariates.

532  
533 In the context of digital soil mapping, we infer that the current use of distance  
534 maps is not satisfactory for several reasons. Including pseudo-covariates with the set  
535 of pedologically relevant covariates can be harmful because it precludes analysis of the  
536 residuals and the generation of new hypotheses from these residuals ([Hawkins, 2012](#)).  
537 It also hampers the interpretation of the most important predictors ([Meyer et al.,](#)  
538 [2019](#)), which is key in several studies on soil mapping. Finally, pseudo-covariates  
539 of distance may well integrate over several pedologically relevant covariates, making  
540 them better predictors or masking the effect of pedologically relevant covariates. In  
541 spatial ecology, alternatives to distance maps are found in the use of spatial eigenvec-  
542 tor maps, spatial filters or trend-surface regression computed on, or optimized for,  
543 the residuals of a model calibrated using ecologically relevant covariates ([Kühn et al.,](#)  
544 [2009](#)). The process is generally in three steps (Fig. 1). In the first step, the variable  
545 of interest is fitted using ecologically relevant covariates, and the (autocorrelated)  
546 residuals are mapped to investigate whether there is an obvious missing spatial pro-  
547 cess in the model. In the second step, spatial surrogate covariates are computed on,  
548 or optimized for, the residuals. Finally, a model is calibrated using the covariates

549 from steps 1 and 2.

550

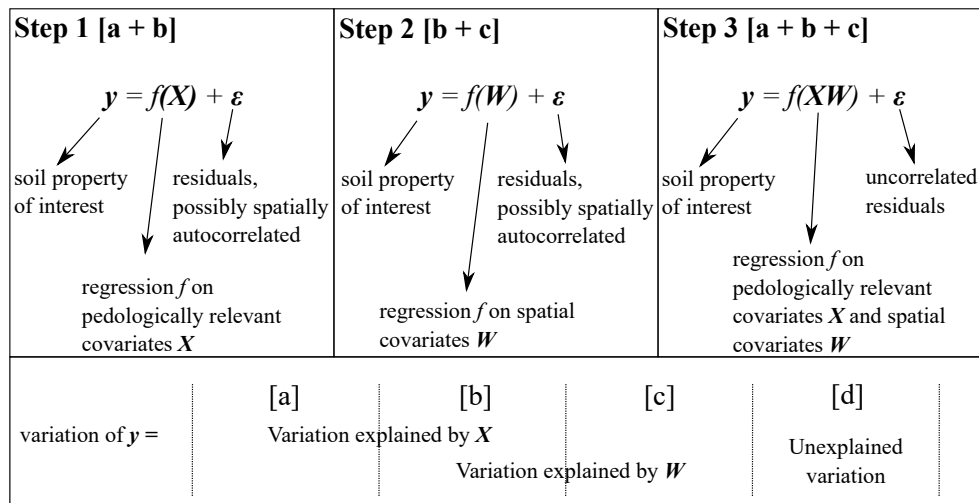


Figure 1: The three steps of variation partitioning between environmental  $\mathbf{X}$  and spatial covariates  $\mathbf{W}$ . The variation of  $\mathbf{y}$  is partitioned into four fractions (Peres-Neto et al., 2006) which are [a] the variation due to the environmental covariates, [b] the variation due to the spatial component of the environmental variables, [c] the spatial component and [d] the unexplained residual variation. Each component is estimated using the amount of variance explained. All [a + b + c + d] sum to 1.

551 The main advantage is to enable subsequent interpretation of the role of environ-  
 552 mental covariates, spatial covariates (most often in the form of Moran’s eigenvector  
 553 maps) and unexplained (uncorrelated) residuals (the “ignorance”) using variation  
 554 partitioning techniques (Peres-Neto et al., 2006). Figure 1 shows that [a + b] is the  
 555 relative influence of environmental variables to the model prediction while [b + c]  
 556 is the relative influence of spatial covariates. The component [b] is the shared vari-  
 557 ation of [a] and [c] because environmental covariates are spatially structured. The  
 558 remaining component [d] computed by  $1 - [a + b + c]$  is the residual fraction of  
 559 the variation. Another benefit of this approach is to have spatial surrogate covari-  
 560 ates with little or no correlation with the meaningful environmental covariates. This  
 561 approach has not yet been tested in DSM, but it would certainly make a valuable  
 562 contribution to increase the interpretability of the ML models and their account of  
 563 the spatial autocorrelation contained in soil data.

### 564 3.4. Multivariate mapping

565 Several authors (e.g. Hengl et al., 2018; Wadoux, 2019b; Wadoux et al., 2019b;  
 566 Padarian et al., 2019) have shown that it is possible to calibrate a single ML model

567 to predict either multiple soil properties or a single soil property at multiple depths.  
568 This reduces the risk of overfitting, computational resources that would be other-  
569 wise required to calibrate several disjoint models (Wadoux, 2019b), and increases  
570 prediction accuracy if there is correlation between the variables to predict. Padarian  
571 et al. (2019) use a multivariate CNN model to predict SOC at multiple soil depths  
572 and report a significant increase of prediction accuracy for the deeper soil depths,  
573 compared to predictions made for each depth separately by a cubist model. Wadoux  
574 (2019b) have shown that for a NN model, it was feasible to constrain the prediction  
575 to avoid inconsistent prediction between compositional soil properties, in particular  
576 soil texture. It was done by adding an additional layer to the model, but we spec-  
577 ulate that this could also be realized by modifying the objective function used to  
578 calibrate the model. Despite a few recent studies, there has been little interest in  
579 multivariate soil mapping using ML algorithms. In the ML literature, it appears  
580 that almost all conventional ML algorithms have a multivariate counterpart. Multi-  
581 variate NNs have already been tested in soil mapping studies. An adaptation of the  
582 RF algorithm for multivariate mapping is proposed by Hengl et al. (2018) but has  
583 several limitations. For example, the calibrated model size increases dramatically  
584 when the number of soil properties to predict also increases and it does not allow to  
585 separate the contribution of the covariates to each predicted property separately. A  
586 theoretical framework for multivariate RF is described by Segal & Xiao (2011) and  
587 was further implemented in the R language by Rahman et al. (2017). For support  
588 vector machines, a multivariate extension is described in Xu et al. (2013).

589  
590 One objective when mapping soil properties or classes is to learn from the cal-  
591 ibrated model. A calibrated multivariate model can provide insights on the soil  
592 property and horizon interrelations. Regrettably, in a multivariate machine learning  
593 model, the correlation between soil properties or depths is not modelled explicitly  
594 (e.g. using a cross-covariance matrix between soil properties). As a result, the corre-  
595 lation between properties or depths cannot be assessed internally and no pedological  
596 interpretation can be derived from the calibrated model. More research is needed on  
597 whether the correlation between original and predicted soil properties (or depths) is  
598 preserved in a multivariate ML model. To model the correlation between properties  
599 explicitly, two solutions are possible. The first is to calibrate additional stochastic  
600 parameters together with the ML parameters (e.g. in a neural network algorithm).  
601 This can take the form of an auto-regressive model between the predictions (Uria  
602 et al., 2016). Another straightforward solution is to calibrate the model with a crite-  
603 rion related to the absolute difference in correlation between the measured properties  
604 and predicted properties. While this is easy to implement in ML calibration based



605 on an objective function (e.g. neural network), this is not straightforward for models  
606 such as RF. Overall, including correlation between properties or depths when predict-  
607 ing with a ML algorithm requires further investigation so as to build pedologically  
608 realistic and interpretable models.

### 609 *3.5. Uncertainty analysis*

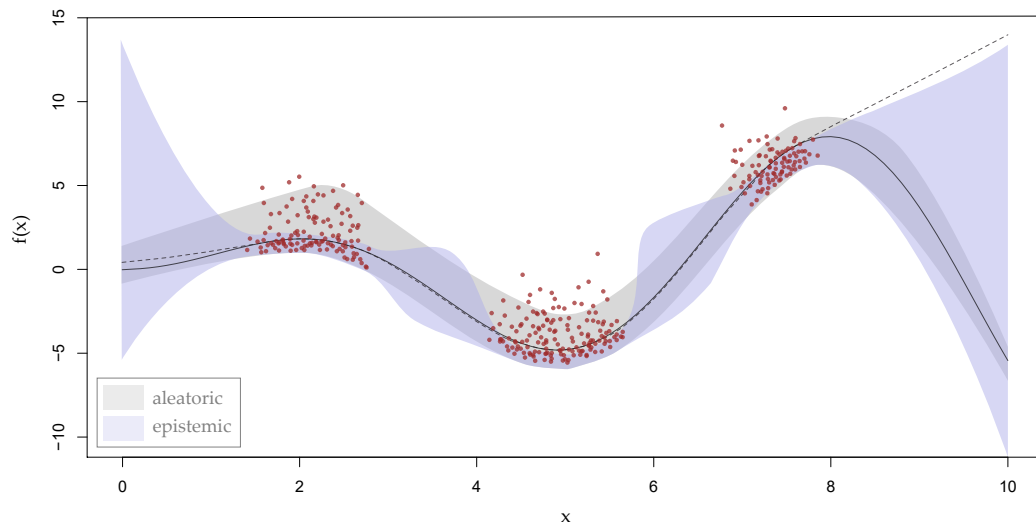
610 Uncertainty analysis in digital soil mapping is crucial to deciding whether the  
611 predicted soil map is reliable to be used for agricultural production systems or de-  
612 cision making. Uncertainty analysis is also about knowing better the limits of the  
613 models and is therefore one step towards model interpretability. At the higher level,  
614 the machine learning literature distinguishes two sources of uncertainty: aleatoric  
615 and epistemic uncertainties (Fig. 2). Aleatoric uncertainty is the data noise variance  
616 (in other terms, the data error), and arises from noise in the data and measurement  
617 error. Epistemic uncertainty refers to model and model parameter uncertainty and  
618 represents our ignorance about a true model that generated the data. While epis-  
619 temic uncertainty is easy to reduce (e.g. by collecting more data at areas of low  
620 sampling density), aleatoric uncertainty is rather difficult to assess (one must repeat  
621 the measurement several times) and even more to reduce. Methods to quantify epis-  
622 temic uncertainty are bootstrapping, or Bayesian modelling. Quantifying epistemic  
623 uncertainty enables to obtain confidence intervals of the prediction. Aleatoric uncer-  
624 tainty is mainly quantified by quantile regression methods, but Monte-Carlo simu-  
625 lation from the probability distribution of the observations might also be a possible  
626 approach. The quantification of both aleatoric and epistemic uncertainty provides  
627 prediction intervals with methods such as quantile regression forest (QRF), the Delta  
628 or Bayesian methods and the mean plus variance estimate (MVE) for neural network  
629 algorithms.

630  
631 The recent development of conditional generative adversarial networks (cGAN)  
632 (Mirza & Osindero, 2014) to generate possible realizations of the observations with  
633 specific conditions or characteristics seem to be of particular interest to include mea-  
634 surement error in DSM. Including measurement error is considered by Wadoux et al.  
635 (2019b) for mapping soil organic carbon using uncertain measurement of the soil  
636 property. However, the authors do not propose a method to quantify the uncer-  
637 tainty of the measurements, nor propagate the measurement error to the predicted  
638 map. With cGAN, a probability distribution of the observations is built, which might  
639 be used for Monte Carlo simulations. Each Monte Carlo sample is used as input in  
640 the ML algorithm, and the final map is the integration of all these simulations. This  
641 would effectively tackle the aleatoric uncertainty of the ML model. More impor-



642 tantly, this would also quantify the uncertainty present in the measurements, which  
643 is currently one of the most important challenges in DSM.

644



661 model with very high predictive abilities, an uncertainty quantification would show  
662 unrealistic predictions characterized by a large uncertainty (see right-hand side of  
663 Fig. 2). This is the results of ML algorithm being very poor predictors for extrapo-  
664 lating to areas of the covariate space that is not comprised in the calibration sample.  
665 Uncertainty quantification that separates out data and model uncertainties is thus  
666 recommended to complete the evaluation of the predicted maps.

667

668 We derive a complementary note about the generation of digital soil maps with  
669 ML by the private sector. Companies and commercial software usually do not report  
670 measures of the uncertainty associated to the maps and there is no transparency  
671 requirement on the methods and quality of soil data. Reporting the uncertainty  
672 associated to the prediction is essential to guide decision-making and political action.  
673 The danger comes from the generated map which gives the appearance of scientific  
674 knowledge where there is none. Making a decision made on maps which are presumed  
675 correct but are in fact away from reality, is presumably worse than making a decision  
676 made in full appreciation of the limits of the map.

### 677 3.6. Validation

678 Studies by [Roberts et al. \(2017\)](#) and [Ruß & Brenning \(2010\)](#) have found that the  
679 estimated performance of the machine learning algorithms applied to spatial data  
680 depends on the validation strategy. In DSM, model performance is usually assessed  
681 using random  $k$ -fold cross-validation (CV) or single random split of a sample into  
682 calibration and validation and/or test subsamples. These strategies give considerably  
683 over-optimistic validation statistics estimates because of the presence of autocorre-  
684 lation in the observations ([Micheletti et al., 2014](#); [Gasch et al., 2015](#); [Meyer et al.,](#)  
685 [2018](#)). Validation statistics estimated from a random split of the master sample as-  
686 sess the ability of the model to reproduce the calibration sample but fail to assess the  
687 model performance in terms of spatial mapping ([Meyer et al., 2019](#)). As an alterna-  
688 tive, several methods ([Brenning, 2012](#); [Le Rest et al., 2014](#); [Pohjankukka et al., 2017](#);  
689 [Meyer et al., 2019](#)) for spatial cross-validation are proposed to account for spatial  
690 autocorrelation of the observations. Two main strategies are adopted. [Roberts et al.](#)  
691 [\(2017\)](#); [Brenning \(2012\)](#); [Meyer et al. \(2019\)](#) use a spatial block approach for  $k$ -fold  
692 CV where the master sample is divided into  $k$  spatially disjoint subsamples using  
693 clustering algorithms on the coordinates or by dividing the spatial domain based on  
694  $k$  cells. In [Le Rest et al. \(2014\)](#) and [Pohjankukka et al. \(2017\)](#), observations from  
695 the calibration subsample that are within a given geographic distance of the valida-  
696 tion subsample are omitted from the calibration subsample, after which the model is  
697 fitted using the remaining observations from the calibration subsample. While these

698 two approaches account for spatial autocorrelation of the observation during valida-  
699 tion, further research is required to provide guideline to select the realistic distance  
700 from which a validation data point is statistically independent from the calibration  
701 sample so as to avoid the opposite effect, i.e. extrapolation and subsequent underop-  
702 timistic validation statistics estimates. Spatial-cross validation is integrated in the  
703 framework presented in Fig. 3.

704

705 Research on spatial cross-validation has drawn attention to the role of autocor-  
706 relation on the calibration on the machine learning algorithms. [Schratz et al. \(2019\)](#)  
707 show that hyperparameter tuning is also impacted by spatial autocorrelation, and  
708 that overoptimistic results are reported when the same data are used for performance  
709 assessment and parameter tuning. They proposed a nested (block) cross-validation  
710 approach for hyperparameter tuning ([Schratz et al., 2019](#)) where spatial block are  
711 split a second time into spatially disjoint geographic subsamples used to optimize the  
712 hyperparameters. The major disadvantage of this method is the dramatic increase in  
713 computing time, which is solved by distributed (parallel) computing solutions. Simi-  
714 larly to the hyperparameter tuning using nested spatial cross-validation, [Meyer et al.](#)  
715 [\(2018\)](#) showed that autocorrelated covariates lead to overfitting and visible artefacts  
716 in the predicted map. The study proposes an iterative procedure for variable se-  
717 lection where a group of two variables is first selected based on the error computed  
718 with spatial cross-validation, and new variables are iteratively added only if these  
719 increase the model performance. The study of [Meyer et al. \(2018\)](#) gives another  
720 argument against the use of covariates describing the spatial dependency as these  
721 lead to misinterpretation of the model’s important contributors and impossibility for  
722 the model to generalize.

723

724 [Meyer et al. \(2019\)](#) emphasize the value of visual examination of the predicted  
725 maps in addition to the statistical validation. In [Meyer et al. \(2019\)](#), two maps with  
726 similar map validation accuracy statistics have a different spatial pattern. The study  
727 shows that this is due to the selected covariates, some having strong spatial autocor-  
728 relation leading to visible artefacts in the predicted map. This highlights the need  
729 for research on the evaluation of predicted maps in terms of spatial pattern. [Poggio](#)  
730 [et al. \(2019\)](#) compare the spatial structure of predicted versus observed values by  
731 computing the area under the curve of variograms fitted on the validation locations  
732 for both predicted and observed probability of having a peat soil. This relies, how-  
733 ever, on the assumption that the variogram of the validation locations represent the  
734 mapped area. More research in this direction will be valuable for future DSM stud-  
735 ies. To date, visual assessment of the map to detect artefacts, and in consideration

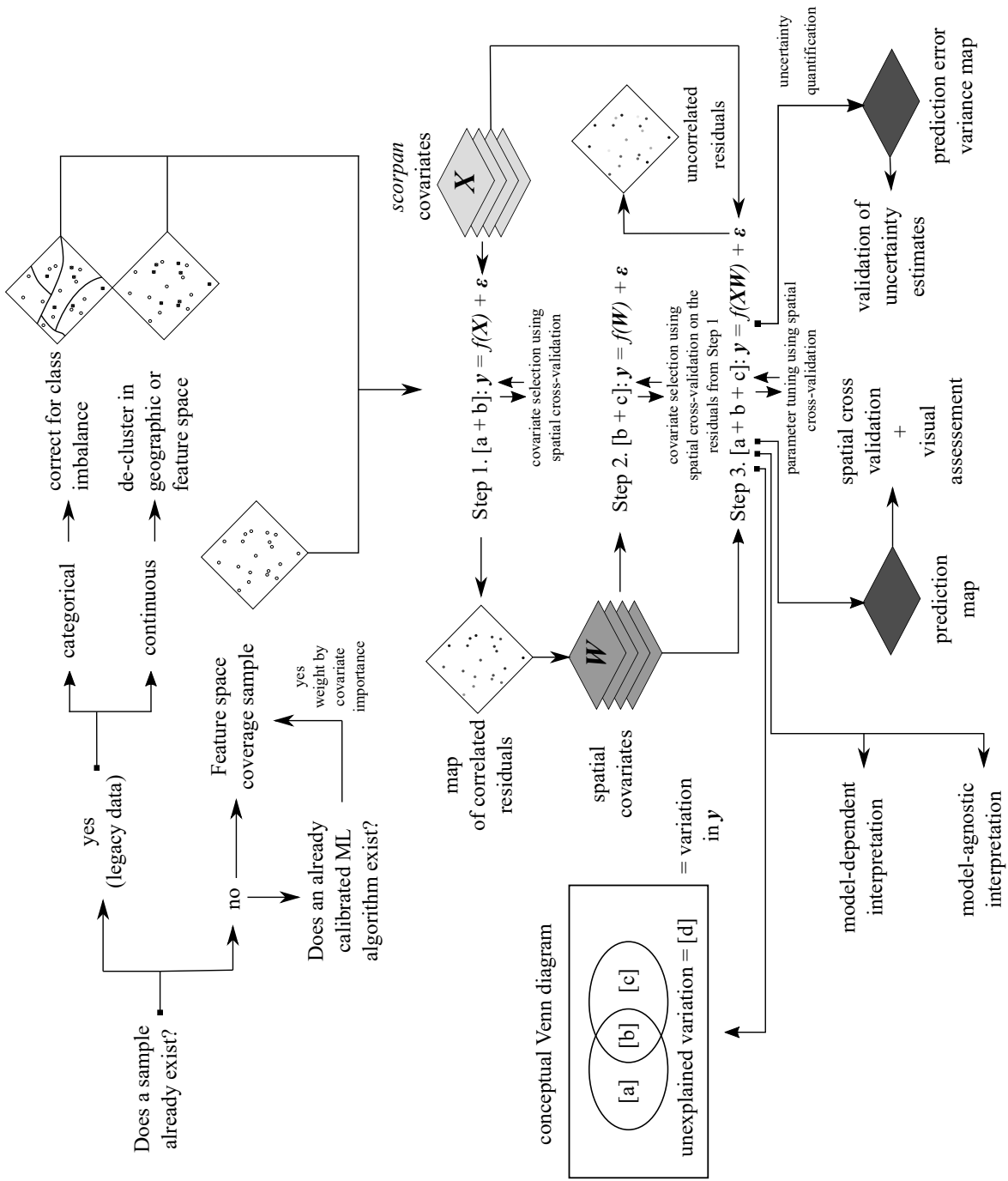


Figure 3: The recommended framework for digital soil mapping with machine learning. The modeller must first decide whether a legacy soil sample or a new sample is collected. He must also decide whether the objective is a categorical or quantitative map. The recommended framework enables the separation between the variation explained by the pedologically relevant covariates and by the spatial covariates. It is recommended to use a spatial cross-validation strategy for validation, but also for parameter tuning and covariate selection.

736 of our knowledge of soil forming processes, is the best option.

### 737 3.7. Machine learning and pedological knowledge

738 Accounting for existing expert soil knowledge in DSM with machine learning is a  
739 challenging exercise (Ma et al., 2019). ML algorithms do not build on any existing a  
740 priori conceptual model of the soil processes and only processes that are conveyed by  
741 the input data are represented in the map (Coveney et al., 2016; Koch et al., 2019).  
742 To prevent extrapolation, Hengl et al. (2014) do not provide soil maps in some under-  
743 sampled areas of the globe such as deserts and glaciers for global mapping of several  
744 soil properties. This stems from incomplete datasets of soil observations for these  
745 areas, despite that extensive expert knowledge exists. In Hengl et al. (2017a) this is  
746 solved by integrating the expert knowledge in the form of expert-based pseudo-points  
747 to guide the ML model in areas of evident extrapolation. In Koch et al. (2019), 600  
748 pseudo-points are also added in under-represented areas of the geographic space. The  
749 study stresses the importance of consulting an expert when building a ML model.  
750 In the same study, meaningful covariates are selected based on existing knowledge  
751 on the soil process, and plausibility of the predicted soil map is made in consid-  
752 eration of the knowledge of soil forming process. On many occasions, meaningful  
753 covariates are selected for mapping soil properties or classes. For example, Brungard  
754 et al. (2015) used a set of covariates selected a priori by an expert on the area under  
755 study. In Viscarra-Rossel & Chen (2011) a set of *scorpan* covariates is selected for  
756 mapping soil properties in Australia. These examples show that in the literature,  
757 adding expert-based pseudo-points and selecting meaningful covariates are, to date,  
758 two straightforward options to include existing knowledge into a ML algorithm for  
759 DSM.

760  
761 The above shows that little is known on how to account for existing knowledge  
762 in ML models. Unfortunately, this is the same order in which the complexity of the  
763 models increases and our understanding of the model functioning decreases. The  
764 increasing caution in the use of predictions made by a complex ML model that one  
765 should expect as a result is not evident. A ML model predicting a number based on  
766 relationships between covariates that are unknown in the view of existing knowledge,  
767 should not be taken with the same seriousness as a number predicted by mechanistic  
768 steps or an established theory. Improvement in this situation is made by ensuring  
769 that the calibrated ML algorithm matches the existing knowledge of the soil  
770 processes, for example by reflecting or confirming the current hypothesis or prior  
771 knowledge on the soil spatial variation for an area. If the model prediction does not  
772 agree with existing maps, this means that the model has instead modelled a different

773 process and is thus likely to be invalid. A model is invalid until it is validated, not  
774 only against data, but also against the researcher experience and validation of the  
775 model creation process (Gahegan, 2019). In short, pedological knowledge should be  
776 integrated to enforce results consistent with the existing scientific principles. This  
777 can be done at each step of the model building, calibration and validation. One  
778 can incorporate additional knowledge by selecting appropriate covariates or adding  
779 pseudo-points. In model building, knowledge takes the form of a hybrid model, a spe-  
780 cific model architecture or objective function (in neural networks models) constrain-  
781 ing the calibration process according to specific knowledge. For example, Wadoux  
782 (2019b) adds the constraint that the prediction of topsoil clay, silt and sand must  
783 sum to 100% in a neural network model. Finally, pedological knowledge is used to  
784 make *post-hoc* checks on the plausibility of the calibrated model and predicted maps.

785  
786 Gahegan (2019) stress that since ML models (the author used the term “predic-  
787 tive process model” in the sense in which “machine learning models” is used in this  
788 article) have no connection to established theory, one can never be sure that the  
789 outcome is realistic given the real-world processes involved. The problem is that a  
790 non-valid model is difficult to recognize and to reject since it is often not interpretable  
791 by a human. To ensure that models fit the existing knowledge, they must be opened  
792 and understood in their functioning. Opening the “black box” is then necessary but  
793 not straightforward (see next section on interpretability), and is often reduced to the  
794 analysis of which environmental covariates are the most often used by the model to  
795 make a prediction (see for example Mahmoudabadi et al. (2017) or McNicol et al.  
796 (2019)).

797  
798 Several authors, however, have warned against the use of accuracy metrics for  
799 pedological interpretation (e.g. Fourcade et al. (2018) or Wadoux et al. (2019c)).  
800 Wadoux et al. (2019c) use meaningless, pseudo-covariates to map soil organic car-  
801 bon over a hypothetical area. The authors obtain an accurate map, and conclude  
802 that ML algorithm should not be used for obtaining new soil knowledge because the  
803 ML algorithm aims at predicting a pattern rather than finding causal relationships.  
804 Wadoux et al. (2019c) suggest to use calibrated ML models as a “hypothesis dis-  
805 covery” tool, in which the mechanisms conveyed by the calibrated ML model are  
806 supplied to the researcher for possible explanations of the soil process, which can  
807 then be confronted to experiments and principles of soil genesis. The challenge that  
808 then arises, noticed by Gahegan (2019) is the conversion of the mechanisms of the  
809 ML model (the model “knowledge”) from a data language to a human one. The data  
810 language is typically parameters or metrics such as the “mean decrease of purity”

811 or “Gini importance index” of a covariate to assess its importance in the prediction  
812 of a soil property or class. Such metrics are not interpretable in terms of human  
813 explanation and they do not relate to soil processes. Translating the data language  
814 to the domain (the human language) requires some attention and further research.  
815 More discussion on this issue is found in [Gahegan et al. \(2001\)](#).

### 816 *3.8. Interpretation of the models*

817 Soil scientists rely on ML algorithms to gain insights into the modelled processes.  
818 Despite providing higher prediction accuracy than other conventional models, ML  
819 models are considered as a black box. Broadly speaking, we do not learn from the  
820 model how the input covariates are related to the output soil property or classes,  
821 and what are the underlying mechanisms behind the prediction. This is unfortunate  
822 for soil science because in many cases the model itself is considered as a source of  
823 knowledge in addition to the collected soil data. Scientific findings remain hidden  
824 when the model only gives a prediction without explanations. In this case, the inter-  
825 pretability of the model warrants the extraction of the knowledge captured by the  
826 calibrated model. [Miller \(2019\)](#) defines interpretability as the degree to which human  
827 can understand the cause of a decision. In general, the need for interpretability of a  
828 machine learning algorithm stems from a deficiency in problem formalization ([Doshi-  
829 Velez & Kim, 2017](#); [Molnar, 2019](#)). This means that for a given task (i.e. mapping  
830 the spatial distribution of soil organic carbon), the prediction itself does not fully  
831 solve the original problem. We suggest three reasons which drive the demand for  
832 interpretability in DSM (adapted from [Doshi-Velez & Kim \(2017\)](#)). The first and  
833 most obvious reason is to increase our scientific understanding of the soil system by  
834 extracting knowledge from the mechanisms captured by the model. Scientists wish  
835 to know which are the drivers of a soil process and, more importantly, whether the  
836 mechanisms captured by the model confirm our scientific understanding of the sys-  
837 tem (see Section 3.7). The second reason is to audit the calibrated ML algorithm. Is  
838 the ML algorithm predicting for the right reasons? If a scientist makes a model for  
839 mapping the topsoil nitrogen content of a field, the interpretation might reveal that  
840 the model is actually predicting soil clay, that is, a proxy of the initial objective.  
841 The third reason is to avoid financial loss or to prevent a safety issue. Take the  
842 example of the remediation of the soil due to radioactive fallout after the Fukushima  
843 nuclear accident. A map of contaminated soils made by a ML algorithm would typi-  
844 cally predict the dominant soil type characteristics, i.e. forest soil (about 75% of the  
845 area), for classification into contaminated or not contaminated areas. Interpretation  
846 of the model might then reveal that the important features learned by the model are  
847 unrealistic for agricultural landscapes and residential areas whose remediation is yet

848 critical to safely move back the population ([Evrard et al., 2019](#)).

849



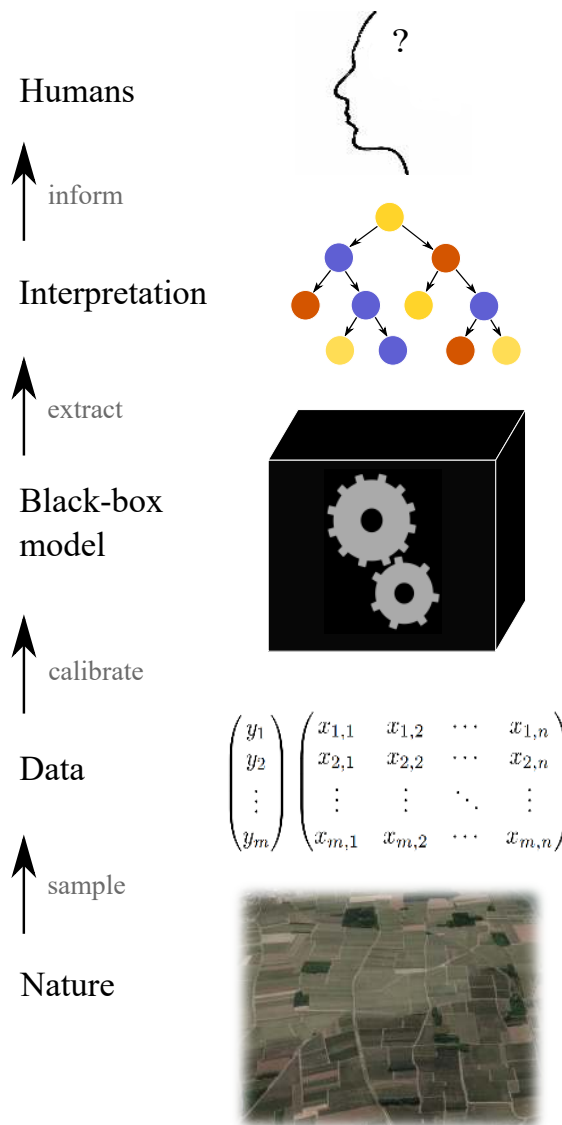


Figure 4: Summary framework for model-agnostic interpretable ML, adapted from Molnar (2019). The lowest level is the reality, the unknown real-world soil that one wants to predict. The second level is the dataset that is extracted from the reality. We collect a fraction of the reality, a sample, and link it to exhaustively known environmental covariates. The relationships between the covariates and the sample is learned by a black-box machine learning model (level 3), on top of which comes the interpretation level to extract some knowledge from the structure of the calibrated model. The structure of the model is converted to human understandable knowledge.

850 A straightforward way to increase interpretability is to decrease model complexity,  
851 for example by building a single decision tree instead of a random forest composed  
852 of several thousand trees. A simple model enables visualization of the important  
853 mechanisms of the model and resultant explanations. For DT algorithms, it is possi-  
854 ble to map the predicted values for specific rules (if the model is sufficiently simple).  
855 Decreasing complexity, however, is done at the expense of model prediction accu-  
856 racy. For more complex ML algorithms, built-in features allow the user to retrieve  
857 the variable importance. In decision-tree like algorithms, the variable importance  
858 is derived from the thresholds used for the splits. For neural networks, the output  
859 weights associated with the input layer neurons provides an indication of the impor-  
860 tant features (Gahegan, 2000). One drawback of these techniques is their inability to  
861 provide information on whether the covariates have a causal link to the modelled soil  
862 property or class, which leads several authors to warn against their use for knowledge  
863 discovery (e.g. Fourcade et al., 2018). More importantly, these variable importance  
864 metrics are summary statistics not always meaningful and they are model-specific,  
865 i.e. they preclude comparison between models or parts of the predicted map. Molnar  
866 (2019) reviews techniques to interpret ML algorithms and define two main categories  
867 of interpretation techniques. The first are the model specific ones. They are rou-  
868 tinely used in DSM activities (e.g. RF variable importance). The second category  
869 falls into model-free techniques, also called model-agnostics (Molnar, 2019). It en-  
870 ables the users to use any model, thus not restraining themselves to simple models or  
871 models with embedded features of interpretation. A summary of how model-agnostic  
872 techniques are employed is shown in Fig. 4. Examples of model-agnostic techniques  
873 are the partial dependence plot (Friedman, 2001) if the number of covariate is small  
874 (two maximum), individual conditional expectation (Goldstein et al., 2015), and  
875 global or local model-agnostic explanation (LIME, Ribeiro et al., 2016). Finally,  
876 sensitivity analysis is also a straightforward means of *post-hoc* interpretation of how  
877 the model output depends upon the different covariates.

#### 878 4. The way forward

879 Machine learning algorithms are now extensively used in soil mapping for re-  
880 gression and classification purposes, much in the same way as routinely employed  
881 in other fields of science. There is no doubt that prediction accuracy benefits from  
882 these data-driven models because ML algorithms are not constrained by a pre-defined  
883 conceptual model of the soil spatial variation, in comparison to mechanistic or even  
884 geostatistical models. The question now is how to increase our scientific understand-  
885 ing of the soil and how to adapt and guide the use of ML to the challenges pertaining

886 to soil mapping and soil science in general. Future research on soil mapping with  
887 machine learning should incorporate the three core elements proposed by Roscher  
888 et al. (2019) and Lipton (2018) which we adapted, as follows:

889

890 *Plausibility:* Models should not only be accurate but also valid in light of the  
891 current knowledge and scientific theories. A model should predict for the right rea-  
892 sons. The plausibility is the solution path taken by the ML algorithm to link the  
893 input to the output, and does not depend directly on the data (Lipton, 2018). In  
894 practical terms, it starts with the model building step, by feeding the model with  
895 credible covariates and by accounting for the spatial particularities of soil data. Spa-  
896 tial or temporal correlation among data should be modelled, either by using a specific  
897 model (e.g. a convolutional neural network), or by using a model architecture that  
898 accounts for this particularity (see Section 3.3). Plausibility also takes the form of  
899 model constraints, to avoid the prediction of unrealistic proportions or ratios. The  
900 plausibility can be further tested in terms of model simulatability (Lipton, 2018).  
901 Since ML algorithms can model arbitrary patterns, there should be some attempts  
902 to test the model with synthetic data or data from a calibrated mechanistic model  
903 representing a large range of dynamics (Reichstein et al., 2019). Increasing model  
904 plausibility will facilitate the acceptance of ML to a large range of scenarios in soil  
905 science.

906

907 *Interpretability:* Interpretability is the translation of an abstract model or model  
908 output into terms understandable by humans (Montavon et al., 2018). Model in-  
909 terpretability pairs with model plausibility and hypothesis discovery. Complex and  
910 arbitrary patterns extracted from the data by an algorithm can be understood only  
911 by the transparency of the model. Interpretation is obtained by model-specific and  
912 model-agnostic methods, described in Section 3.8. Visual examination of the maps is  
913 also a means of interpretability. While complex ML models are potentially harmful  
914 because they often do not model any real-world process, there is an opportunity to  
915 challenge existing knowledge by *post-hoc* comparison of existing maps produced by  
916 expert knowledge with the maps predicted by a ML model, and by analysis of the  
917 striking differences. This is possible only if the model is interpretable by humans  
918 and the physical relationships between variables are realistic (the model is *plausible*).  
919 Model interpretation is also an opportunity to generate new hypotheses, by inter-  
920 preting the relationships found by the ML algorithm in the stores of soil data. The  
921 new hypotheses derived by these interpretations may challenge existing knowledge  
922 on the soil spatial variation and genesis.

923

924 *Explainability*: Modellers should shy away from mindless model fitting and pre-  
925 diction and intensify research on models that both predict and explain. Explanations  
926 aim to answer the three questions: *what* is the modelled process?, *how* has it been  
927 modelled?, and *why* has this process been modelled? (Miller, 2019). In this sense,  
928 explaining a process is an interpretation of a ML model plus expert knowledge and  
929 contextual information. For example, a different explanation is warranted when one  
930 wants to explain the pattern of a predicted soil map or the reason for two close  
931 predicted soil classes to be different. To explain, the modeller uses the data, the  
932 plausibility of the model and its interpretation using expert knowledge (see Fig. 5).  
933 Explainability is helped by model structure providing algorithmic explanations in  
934 the form of graphs or equations.

935  
936 An example of model structure providing algorithmic explanations in DSM is  
937 found with the use of Bayesian belief networks (BBN, Cooper, 1990) in Mayr et al.  
938 (2010) and later Taalab et al. (2015). BBN is a probabilistic graphical model pre-  
939 dicting the likely value of a soil property or class given conditional dependencies  
940 between covariates. Recent advances in ML have made a step further by discovering  
941 the graph structure directly from the data. However, while BBN is an interpretable  
942 ML model of conditional dependence between variables, the process that generated  
943 these dependencies remains hidden. To discover new processes from data, inductive  
944 process modelling (Asgharbeygi et al., 2006) and genetic algorithms (Goldberg &  
945 Holland, 1988) are the way forward. Both are automated model discovery process,  
946 in which equations describing a process are inductively (i.e. using the data) assem-  
947 bled into a single predictive model by heuristic search methods (Bridewell et al., 2008;  
948 Gahegan, 2019). The calibrated model is a set of equations constrained by existing,  
949 verified equations (e.g. differential equations of the water flow) representing causal  
950 relationships between variables. The model can be refined using expert knowledge  
951 and additional data (Dale et al., 1989). More importantly, these models produce  
952 explanations, which can be refuted or approved in light of scientific principles.

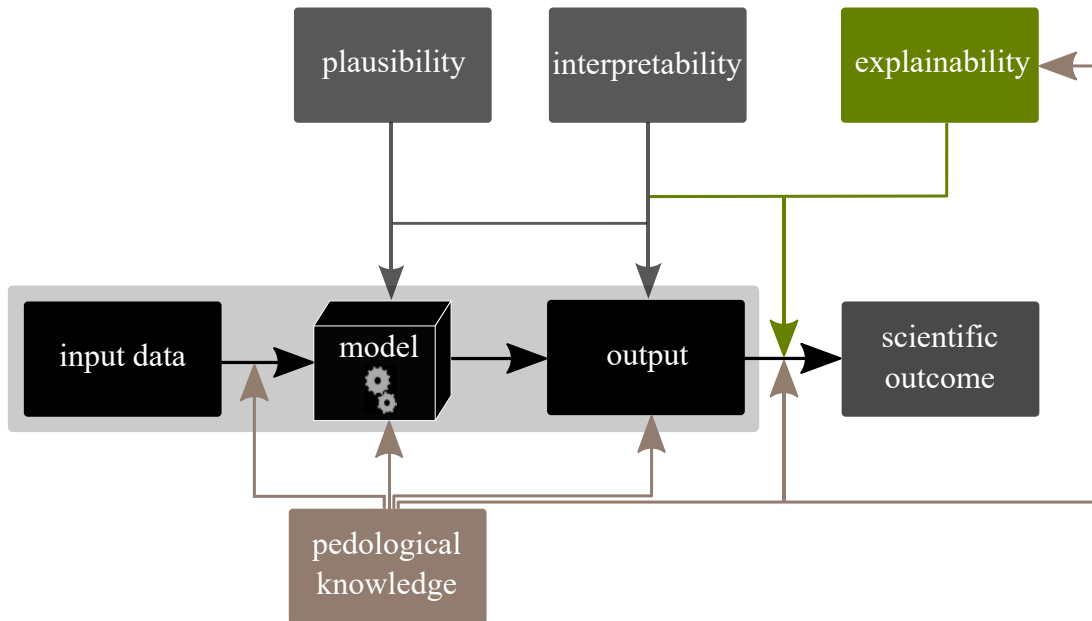


Figure 5: Conceptual framework for the derivation of a scientific outcome from a ML model, adapted from Roscher et al. (2019). The light grey box represents the conventional use of ML algorithms in digital soil mapping, in which an output is derived from a calibrated ML model given a set of input data. A scientific outcome is obtained by explaining the output of a model using pedological knowledge, but also by ensuring scientific consistency at each link of the chain. Alternatively, a plausible and interpretable model can be explained using pedological knowledge.

953 Figure 5 illustrates the central role played by the three elements *plausibility*,  
 954 *interpretability* and *explainability* in obtaining a scientific outcome from machine  
 955 learning. Fig. 5 shows that the three core elements are conditioned to the use of  
 956 pedological knowledge at each link of the chain. Enforcing pedological knowledge  
 957 during modelling restricts the solution space to scientifically consistent results and  
 958 may decrease the overall prediction accuracy. For digital soil mapping purposes,  
 959 it is not obvious whether an increase of predictive accuracy worth the substantial  
 960 decrease of model consistency. For this reason, recent studies (e.g. Bennett et al.,  
 961 2013; Lapuschkin et al., 2019) advocate the use of other criteria to measure the  
 962 overall performance, such as model complexity or consistency (Karpatne et al., 2017).  
 963 Including other criteria to assess the overall performance of a ML model would  
 964 certainly make one step towards “conscious” digital soil mapping, and participate to  
 965 the uptake of knowledge discovery via machine learning in soil science.

## 966 5. Conclusion

967 In this contribution, we have reviewed the current and prospective use of ML al-  
968 gorithms for digital soil mapping. From the existing use of ML in DSM, we identified  
969 key challenges and provided partial solutions We draw the following conclusions.

- 970 • There has been a large number of studies mapping soil properties or classes  
971 using ML algorithms. A wide range of soil properties, attributes and types have  
972 been predicted. Likewise, an increasing number of machine learning algorithms  
973 have been tested. Case studies are dominated by the use of legacy samples for  
974 local to regional scale (about  $10^4$  km<sup>2</sup>) areas. Ensemble of different algorithms  
975 to improve prediction are gaining more attention. All studies reported at least  
976 one validation statistics but few reported the uncertainty associated with the  
977 prediction.
- 978 • The configuration of a good sampling design for mapping with machine learning  
979 is largely unknown. The impact of the sampling design on model calibration  
980 and prediction has generally been disregarded. More research is needed in this  
981 direction.
- 982 • A large number of studies have focused solely on achieving a high mapping  
983 accuracy. Comparison between models and other studies are made based on  
984 validation statistics, while ignoring model complexity or consistency with re-  
985 spect to the existing pedological knowledge.
- 986 • The benefit of using a large number of covariates, or pseudo-covariates ac-  
987 counting for residual spatial autocorrelation for mapping using ML algorithms  
988 should be avoided. To build consistent models, we suggested to select a set of  
989 pedologically relevant covariates, and to model the potential residual spatial  
990 autocorrelation with *post-hoc* fitting of another model using spatial surrogate  
991 covariates. This procedure also enables a separate analysis of the variation  
992 explained by environmental or spatial covariates.

993 Overall, our review of the literature suggested that in recent studies inference is  
994 relegated to the background with the emergence of the mapping accuracy as the sole  
995 standard by which progress is measured. While the mapping accuracy is valuable, it  
996 should not be the only objective one should pursue. To date, ML is applied to digital  
997 soil mapping the same way as other fields such as image detection or pattern recog-  
998 nition do. Any prediction can become a soil map, whether it contains soil knowledge  
999 or not, and without any assessment on whether the fitted relationships relate to a

1000 real-world soil process.

1001

1002 We also found, however, that there is opportunity to include pedological knowl-  
1003 edge at each step of the modelling chain, to improve or correct the existing dataset,  
1004 to design the model architecture, to constrain the model calibration, or to analyse  
1005 the output using *post-hoc* checks on the predicted soil maps. Future studies on DSM  
1006 should use *plausible*, *interpretable* and *explainable* ML models to extract novel sci-  
1007 entific results from soil data. One step towards achieving this goal is to integrate  
1008 model consistency in addition to model prediction accuracy to evaluate the overall  
1009 performance of the mapping approach. This will ensure that future studies use mod-  
1010 els that are not only accurate but also valid in light of the current knowledge and  
1011 scientific theories.

## 1012 **Acknowledgement**

1013 Budiman Minasny is member of a consortium supported by LE STUDIUM Loire  
1014 Valley Institute for Advanced Studies through its LE STUDIUM Research Consor-  
1015 tium Programme.

## 1016 **References**

1017 Adhikari, K., Hartemink, A. E., Minasny, B., Kheir, R. B., Greve, M. B., & Greve,  
1018 M. H. (2014). Digital mapping of soil organic carbon contents and stocks in den-  
1019 mark. *PLOS ONE*, *9*, e105519.

1020 Aitkenhead, M. J., & Coull, M. C. (2016). Mapping soil carbon stocks across Scotland  
1021 using a neural network model. *Geoderma*, *262*, 187–198.

1022 Akpa, S. I. C., Odeh, I. O. A., Bishop, T. F. A., & Hartemink, A. E. (2014). Digital  
1023 mapping of soil particle-size fractions for Nigeria. *Soil Science Society of America*  
1024 *Journal*, *78*, 1953–1966.

1025 Arrouays, D., McKenzie, N., Hempel, J., de Forges, A. R., & McBratney, A. B.  
1026 (2014). *GlobalSoilMap: Basis of the Global Spatial Soil Information System*. CRC  
1027 press, Boca Raton, USA.

1028 Asgharbeygi, N., Langley, P., Bay, S., & Arrigo, K. (2006). Inductive revision of  
1029 quantitative process models. *Ecological Modelling*, *194*, 70–79.

- 1030 Batjes, N. H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl, T., & Mendes de  
1031 Jesus, J. (2017). WoSIS: providing standardised soil profile data for the world.  
1032 *Earth System Science Data*, *9*, 1–14.
- 1033 Beguin, J., Fuglstad, G.-A., Mansuy, N., & Paré, D. (2017). Predicting soil properties  
1034 in the Canadian boreal forest with limited data: Comparison of spatial and non-  
1035 spatial statistical approaches. *Geoderma*, *306*, 195–205.
- 1036 Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., & Goldschmitt,  
1037 M. (2005). Digital soil mapping using artificial neural networks. *Journal of Plant*  
1038 *Nutrition and Soil Science*, *168*, 21–33.
- 1039 Behrens, T., Schmidt, K., MacMillan, R. A., & Viscarra-Rossel, R. A. (2018a).  
1040 Multi-scale digital soil mapping with deep learning. *Scientific Reports*, *8*, 15244.
- 1041 Behrens, T., Schmidt, K., Viscarra-Rossel, R. A., Gries, P., Scholten, T., & MacMil-  
1042 lan, R. A. (2018b). Spatial modelling with Euclidean distance fields and machine  
1043 learning. *European Journal of Soil Science*, *69*, 757–770.
- 1044 Behrens, T., Zhu, A.-X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital  
1045 terrain analysis and feature selection for digital soil mapping. *Geoderma*, *155*,  
1046 175–185.
- 1047 Bel, L., Allard, D., Laurent, J. M., Cheddadi, R., & Bar-Hen, A. (2009). CART al-  
1048 gorithm for spatial data: Application to environmental and ecological data. *Com-*  
1049 *putational Statistics & Data Analysis*, *53*, 3082–3093.
- 1050 Bel, L., Laurent, J. M., Bar-Hen, A., Allard, D., & Cheddadi, R. (2005). A spatial  
1051 extension of CART: application to classification of ecological data. In P. Renard,  
1052 H. Demougeot-Renard, & R. Froidevaux (Eds.), *Geostatistics for Environmental*  
1053 *Applications* (pp. 99–109). Springer, Berlin, Heidelberg.
- 1054 Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H.,  
1055 Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C.  
1056 et al. (2013). Characterising performance of environmental models. *Environmental*  
1057 *Modelling & Software*, *40*, 1–20.
- 1058 Blanco, C. M. G., Gomez, V. M. B., Crespo, P., & Ließ, M. (2018). Spatial prediction  
1059 of soil water retention in a Páramo landscape: Methodological insight into machine  
1060 learning using random forest. *Geoderma*, *316*, 100–114.
- 1061 Braben, D. W. (1985). Innovation and academic research. *Nature*, *316*, 401–402.



- 1062 Breiman, L. (2017). *Classification and Regression Trees*. Routledge, New York, USA.
- 1063 Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of  
1064 prediction rules in remote sensing: The R package sperrorest. In *2012 International*  
1065 *Geoscience and Remote Sensing Symposium* (pp. 5372–5375). IEEE.
- 1066 Bridewell, W., Langley, P., Todorovski, L., & Džeroski, S. (2008). Inductive process  
1067 modeling. *Machine learning*, *71*, 1–32.
- 1068 Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards Jr, T. C.  
1069 (2015). Machine learning for predicting soil classes in three semi-arid landscapes.  
1070 *Geoderma*, *239*, 68–83.
- 1071 Brus, D. J. (2019). Sampling for digital soil mapping: A tutorial supported by R  
1072 scripts. *Geoderma*, *338*, 464–480.
- 1073 Bui, E., Henderson, B., & Viergever, K. (2009). Using knowledge discovery with data  
1074 mining from the Australian Soil Resource Information System database to inform  
1075 soil carbon mapping in Australia. *Global Biogeochemical Cycles*, *23*, GB4033.
- 1076 Bui, E. N., Loughhead, A., & Corner, R. (1999). Extracting soil-landscape rules  
1077 from previous soil surveys. *Soil Research*, *37*, 495–508.
- 1078 Carré, F., McBratney, A. B., & Minasny, B. (2007). Estimation and potential im-  
1079 provement of the quality of legacy soil samples for digital soil mapping. *Geoderma*,  
1080 *141*, 1–14.
- 1081 Chen, S., Mulder, V. L., Martin, M. P., Walter, C., Lacoste, M., Richer-de Forges,  
1082 A. C., Saby, N. P. A., Loiseau, T., Hu, B., & Arrouays, D. (2019). Probability  
1083 mapping of soil thickness by random survival forest at a national scale. *Geoderma*,  
1084 *344*, 184–194.
- 1085 Cialella, A. T., Dubayah, R., Lawrence, W., & Levine, E. (1997). Predicting soil  
1086 drainage class using remotely sensed and digital elevation data. *Photogrammetric*  
1087 *Engineering and Remote Sensing*, *63*, 171–177.
- 1088 Cooper, G. F. (1990). The computational complexity of probabilistic inference using  
1089 bayesian belief networks. *Artificial intelligence*, *42*, 393–405.
- 1090 Coveney, P. V., Dougherty, E. R., & Highfield, R. R. (2016). Big data need big theory  
1091 too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and*  
1092 *Engineering Sciences*, *374*, 20160153.

- 1093 Cressie, N., & Johannesson, G. (2008). Fixed rank kriging for very large spatial data  
1094 sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,  
1095 *70*, 209–226.
- 1096 Dai, F., Zhou, Q., Lv, Z., Wang, X., & Liu, G. (2014). Spatial prediction of soil  
1097 organic matter content integrating artificial neural network and ordinary kriging  
1098 in Tibetan Plateau. *Ecological Indicators*, *45*, 184–194.
- 1099 Dale, M. B., McBratney, A. B., & Russell, J. S. (1989). On the role of expert systems  
1100 and numerical taxonomy in soil classification. *Journal of Soil Science*, *40*, 223–234.
- 1101 De Gruijter, J. J., Brus, D. J., Bierkens, M. F. P., & Knotters, M. (2006). *Sampling*  
1102 *for Natural Resource Monitoring*. Springer Science & Business Media, Dordrecht,  
1103 NL.
- 1104 Dharumarajan, S., Hegde, R., & Singh, S. K. (2017). Spatial prediction of major  
1105 soil properties using Random Forest techniques-A case study in semi-arid tropics  
1106 of South India. *Geoderma Regional*, *10*, 154–162.
- 1107 Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable  
1108 machine learning. [arXiv:1702.08608](https://arxiv.org/abs/1702.08608).
- 1109 Ellili, Y., Malone, B. P., Michot, D., Minasny, B., Vincent, S., Walter, C., &  
1110 Lemercier, B. (2019). Comparing three approaches of spatial disaggregation of  
1111 legacy soil maps based on DSMART algorithm. *SOIL Discussions*, .
- 1112 Evrard, O., Laceby, J. P., & Nakao, A. (2019). Effectiveness of landscape decontam-  
1113 ination following the Fukushima nuclear accident: a review. *SOIL*, *5*, 333–350.
- 1114 Fick, S. E., & Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution  
1115 climate surfaces for global land areas. *International Journal of Climatology*, *37*,  
1116 4302–4315.
- 1117 Forkuor, G., Hounkpatin, O. K. L., Welp, G., & Thiel, M. (2017). High resolution  
1118 mapping of soil properties using remote sensing variables in south-western Burkina  
1119 Faso: a comparison of machine learning and multiple linear regression models.  
1120 *PLOS ONE*, *12*, e0170478.
- 1121 Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution  
1122 of species, or the challenge of selecting environmental predictors and evaluation  
1123 statistics. *Global Ecology and Biogeography*, *27*, 245–256.

- 1124 Friedman, J. H. (2001). Greedy function approximation: a gradient boosting ma-  
1125 chine. *Annals of statistics*, *25*, 1189–1232.
- 1126 Gahegan, M. (2000). On the application of inductive machine learning tools to  
1127 geographical analysis. *Geographical Analysis*, *32*, 113–139.
- 1128 Gahegan, M. (2019). Fourth paradigm GIScience? prospects for automated discovery  
1129 and explanation from data. *International Journal of Geographical Information*  
1130 *Science*, *34*, 1–21.
- 1131 Gahegan, M., Wachowicz, M., Harrower, M., & Rhyne, T.-M. (2001). The integration  
1132 of geographic visualization with knowledge discovery in databases and geocompu-  
1133 tation. *Cartography and Geographic Information Science*, *28*, 29–44.
- 1134 Gasch, C. K., Hengl, T., Gräler, B., Meyer, H., Magney, T. S., & Brown, D. J.  
1135 (2015). Spatio-temporal interpolation of soil water, temperature, and electrical  
1136 conductivity in 3D+ T: The Cook Agronomy Farm data set. *Spatial Statistics*,  
1137 *14*, 70–90.
- 1138 Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J., Lonjou,  
1139 V., Lafrance, B., Massera, S., Gaudel-Vacaresse, A. et al. (2017). Copernicus  
1140 Sentinel-2A calibration and products validation status. *Remote Sensing*, *9*, 584.
- 1141 Georganos, S., Grippa, T., Gadiaga, A. N., Linard, C., Lennert, M., Vanhuysse, S.,  
1142 Mboga, N. O., Wolff, E., & Kalogirou, S. (2019). Geographical random forests: A  
1143 spatial extension of the random forest algorithm to address spatial heterogeneity  
1144 in remote sensing and population modelling. *Geocarto International*, *1*, 1–12.
- 1145 Goldberg, D. E., & Holland, J. H. (1988). *Genetic algorithms and machine learning*.  
1146 Kluwer Academic Publishers-Plenum Publishers. Kluwer Academic Publishers.
- 1147 Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black  
1148 box: Visualizing statistical learning with plots of individual conditional expecta-  
1149 tion. *Journal of Computational and Graphical Statistics*, *24*, 44–65.
- 1150 Gomes, L. C., Faria, R. M., de Souza, E., Veloso, G. V., Schaefer, C. E. G. R., &  
1151 Fernandes Filho, E. I. (2019). Modelling and mapping soil organic carbon stocks  
1152 in Brazil. *Geoderma*, *340*, 337–350.
- 1153 Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. (2008). Soil organic carbon  
1154 concentrations and stocks on Barro Colorado Island—Digital soil mapping using  
1155 Random Forests analysis. *Geoderma*, *146*, 102–113.

- 1156 Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arel-  
1157 lano Hernández, C., Arévalo, G. E., Arroyo-Cruz, C. E., Bolivar, A., Bunning,  
1158 S. et al. (2018). No silver bullet for digital soil mapping: country-specific soil  
1159 organic carbon estimates across Latin America. *SOIL*, *4*, 173–193.
- 1160 Hamzhepour, N., Shafizadeh-Moghadam, H., & Valavi, R. (2019). Exploring the  
1161 driving forces and digital mapping of soil organic carbon using remote sensing and  
1162 soil texture. *CATENA*, *182*, 104141.
- 1163 Hansen, M. K., Brown, D. J., Dennison, P. E., Graves, S. A., & Brickleyer, R. S.  
1164 (2009). Inductively mapping expert-derived soil-landscape units within dambo  
1165 wetland catenae using multispectral and topographic data. *Geoderma*, *150*, 72–  
1166 84.
- 1167 Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., & Schröder, B. (2012). Spa-  
1168 tial disaggregation of complex soil map units: A decision-tree based approach in  
1169 Bavarian forest soils. *Geoderma*, *185*, 37–47.
- 1170 Hartmann, J., & Moosdorf, N. (2012). The new global lithological map database  
1171 GLiM: A representation of rock properties at the Earth surface. *Geochemistry,*  
1172 *Geophysics, Geosystems*, *13*, 1–37.
- 1173 Hawkins, B. A. (2012). Eight (and a half) deadly sins of spatial analysis. *Journal of*  
1174 *Biogeography*, *39*, 1–9.
- 1175 He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions*  
1176 *on Knowledge and Data Engineering*, *21*, 1263–1284.
- 1177 Henderson, B. L., Bui, E. N., Moran, C. J., & Simon, D. A. P. (2005). Australia-wide  
1178 predictions of soil properties using decision trees. *Geoderma*, *124*, 383–398.
- 1179 Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M.,  
1180 Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B.  
1181 et al. (2017a). SoilGrids250m: Global gridded soil information based on machine  
1182 learning. *PLOS ONE*, *12*, e0169748.
- 1183 Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M.,  
1184 Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G., Walsh, M. G. et al.  
1185 (2014). SoilGrids1km—global soil information based on automated mapping.  
1186 *PLOS ONE*, *9*, e105992.

- 1187 Hengl, T., Leenaars, J. G. B., Shepherd, K. D., Walsh, M. G., Heuvelink, G. B. M.,  
1188 Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E. et al. (2017b). Soil  
1189 nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250  
1190 m spatial resolution using machine learning. *Nutrient Cycling in Agroecosystems*,  
1191 *109*, 77–102.
- 1192 Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018).  
1193 Random forest as a generic framework for predictive modeling of spatial and spatio-  
1194 temporal variables. *PeerJ*, *6*, e5518.
- 1195 Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G.  
1196 (2016). An overview and comparison of machine-learning techniques for classifica-  
1197 tion purposes in digital soil mapping. *Geoderma*, *265*, 62–77.
- 1198 Heuvelink, G. B. M., & Webster, R. (2001). Modelling soil variation: past, present,  
1199 and future. *Geoderma*, *100*, 269–301.
- 1200 Holmes, K. W., Odgers, N. P., Griffin, E. A., & van Gool, D. (2014). Spatial disag-  
1201 gregation of conventional soil mapping across Western Australia using DSMART.  
1202 In D. Arrouays, N. McKenzie, J. Hempel, A. Richer de Forges, & A. B. McBratney  
1203 (Eds.), *GlobalSoilMap: Basis of the Global Spatial Soil Information System* (pp.  
1204 273–279). Taylor & Francis, London, UK.
- 1205 Hounkpatin, K. O. L., Schmidt, K., Stumpf, F., Forkuor, G., Behrens, T., Scholten,  
1206 T., Amelung, W., & Welp, G. (2018). Predicting reference soil groups using legacy  
1207 data: A data pruning and Random Forest approach for tropical environment (Dano  
1208 catchment, Burkina Faso). *Scientific Reports*, *8*, 9959.
- 1209 Illés, G., Sutikno, S., Szatmári, G., Sandhyavetri, A., Pásztor, L., Kristijono, A.,  
1210 Molnár, G., Yusa, M., & Székely, B. (2019). Facing the peat CO<sub>2</sub> threat: digital  
1211 mapping of Indonesian peatlands—a proposed methodology and its application.  
1212 *Journal of Soils and Sediments*, (pp. 1–16).
- 1213 Jenny, H. (1941). *Factors of Soil Formation: A System of Quantitative Pedology*.  
1214 McGrawHill, New York, USA.
- 1215 Jiang, Z., Li, Y., Shekhar, S., Rampi, L., & Knight, J. (2017). Spatial ensemble  
1216 learning for heterogeneous geographic data with class ambiguity: A summary of  
1217 results. In *Proceedings of the 25th ACM SIGSPATIAL International Conference*  
1218 *on Advances in Geographic Information Systems* 23 (pp. 1–10). ACM.

- 1219 Kalambukattu, J. G., Kumar, S., & Raj, R. A. (2018). Digital soil mapping in  
1220 a Himalayan watershed using remote sensing and terrain parameters employing  
1221 artificial neural network model. *Environmental Earth Sciences*, *77*, 203.
- 1222 Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A.,  
1223 Shekhar, S., Samatova, N., & Kumar, V. (2017). Theory-guided data science: A  
1224 new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge  
1225 and Data Engineering*, *29*, 2318–2331.
- 1226 Keskin, H., Grunwald, S., & Harris, W. G. (2019). Digital mapping of soil carbon  
1227 fractions with machine learning. *Geoderma*, *339*, 40–58.
- 1228 Kheir, R. B., Greve, M. H., Abdallah, C., & Dalgaard, T. (2010a). Spatial soil zinc  
1229 content distribution from terrain parameters: A GIS-based decision-tree model in  
1230 Lebanon. *Environmental Pollution*, *158*, 520–528.
- 1231 Kheir, R. B., Greve, M. H., Bøcher, P. K., Greve, M. B., Larsen, R., & McCloy,  
1232 K. (2010b). Predictive mapping of soil organic carbon in wet cultivated lands  
1233 using classification-tree based models: The case study of Denmark. *Journal of  
1234 Environmental Management*, *91*, 1150–1160.
- 1235 Kirkwood, C., Cave, M., Beamish, D., Grebby, S., & Ferreira, A. (2016). A machine  
1236 learning approach to geochemical mapping. *Journal of Geochemical Exploration*,  
1237 *167*, 49–61.
- 1238 Koch, J., Stisen, S., Refsgaard, J. C., Ernstsén, V., Jakobsen, P. R., & Højberg, A. L.  
1239 (2019). Modeling Depth of the Redox Interface at High Resolution at National  
1240 Scale Using Random Forest and Residual Gaussian Simulation. *Water Resources  
1241 Research*, *55*, 1451–1469.
- 1242 Kovačević, M., Bajat, B., & Gajić, B. (2010). Soil type classification and estimation  
1243 of soil properties using support vector machines. *Geoderma*, *154*, 340–347.
- 1244 Kühn, I., & Dormann, C. F. (2012). Less than eight (and a half) misconceptions of  
1245 spatial analysis. *Journal of Biogeography*, *39*, 995–998.
- 1246 Kühn, I., Nobis, M. P., & Durka, W. (2009). Combining spatial and phylogenetic  
1247 eigenvector filtering in trait analysis. *Global Ecology and Biogeography*, *18*, 745–  
1248 758.

- 1249 Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., & Walter, C.  
1250 (2014). High resolution 3d mapping of soil organic carbon in a heterogeneous  
1251 agricultural landscape. *Geoderma*, *213*, 296–311.
- 1252 Lagacherie, P. (2008). Digital soil mapping: a state of the art. In A. E. Hartemink,  
1253 A. McBratney, & M. de Lourdes Mendonça-Santos (Eds.), *Digital Soil Mapping*  
1254 *with Limited Data* (pp. 3–14). Springer, Dordrecht, Netherlands.
- 1255 Lagacherie, P., & Holmes, S. (1997). Addressing geographical data errors in a classi-  
1256 fication tree for soil unit prediction. *International Journal of Geographical Infor-*  
1257 *mation Science*, *11*, 183–198.
- 1258 Lagacherie, P., & McBratney, A. (2006). Spatial soil information systems and spatial  
1259 soil inference systems: perspectives for digital soil mapping. *Developments in Soil*  
1260 *Science*, *31*, 3–22.
- 1261 Lamichhane, S., Kumar, L., & Wilson, B. (2019). Digital soil mapping algorithms  
1262 and covariates for soil organic carbon mapping and their implications: A review.  
1263 *Geoderma*, (pp. 395–413).
- 1264 Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-  
1265 R. (2019). Unmasking Clever Hans predictors and assessing what machines really  
1266 learn. *Nature Communications*, *10*, 1096.
- 1267 Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. (2014). Spa-  
1268 tial leave-one-out cross-validation for variable selection in the presence of spatial  
1269 autocorrelation. *Global Ecology and Biogeography*, *23*, 811–820.
- 1270 Ließ, M., Glaser, B., & Huwe, B. (2012). Uncertainty in the spatial prediction of  
1271 soil texture: comparison of regression tree and random forest models. *Geoderma*,  
1272 *170*, 70–79.
- 1273 Ließ, M., Schmidt, J., & Glaser, B. (2016). Improving the spatial prediction of soil  
1274 organic carbon stocks in a complex tropical mountain landscape by methodological  
1275 specifications in machine learning approaches. *PLOS ONE*, *11*, e0153673.
- 1276 Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*, 31–57.
- 1277 Liu, F., Zhang, G.-L., Song, X., Li, D., Zhao, Y., Yang, J., Wu, H., & Yang, F.  
1278 (2019). High-resolution and three-dimensional mapping of soil texture of China.  
1279 *Geoderma*, *361*, 114061.

- 1280 Lorenzetti, R., Barbetti, R., Fantappiè, M., L'Abate, G., & Costantini, E. A. (2015).  
1281 Comparing data mining and deterministic pedology to assess the frequency of wrb  
1282 reference soil groups in the legend of small scale maps. *Geoderma*, *237*, 237–245.
- 1283 Ma, Y., Minasny, B., Malone, B. P., & Mcbratney, A. B. (2019). Pedology and digital  
1284 soil mapping (DSM). *European Journal of Soil Science*, *70*, 216–235.
- 1285 MacKay, D. J. C. (1992). Information-based objective functions for active data  
1286 selection. *Neural Computation*, *4*, 590–604.
- 1287 Mahmoudabadi, E., Karimi, A., Haghnia, G. H., & Sepehr, A. (2017). Digital soil  
1288 mapping using remote sensing indices, terrain attributes, and vegetation features in  
1289 the rangelands of northeastern Iran. *Environmental Monitoring and Assessment*,  
1290 *189*, 500.
- 1291 Malone, B. P., McBratney, A. B., Minasny, B., & Laslett, G. M. (2009). Mapping  
1292 continuous depth functions of soil carbon storage and available water capacity.  
1293 *Geoderma*, *154*, 138–152.
- 1294 Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier,  
1295 V., & Beaudoin, A. (2014). Digital mapping of soil properties in Canadian managed  
1296 forests at 250 m of resolution using the k-nearest neighbor method. *Geoderma*,  
1297 *235*, 59–73.
- 1298 Marchant, B. P., Viscarra Rossel, R. A., & Webster, R. (2013). Fluctuations in  
1299 method-of-moments variograms caused by clustered sampling and their elimination  
1300 by declustering and residual maximum likelihood estimation. *European Journal of*  
1301 *Soil Science*, *64*, 401–409.
- 1302 Massawe, B. H. J., Subburayalu, S. K., Kaaya, A. K., Winowiecki, L., & Slater, B. K.  
1303 (2018). Mapping numerically classified soil taxa in Kilombero valley, Tanzania  
1304 using machine learning. *Geoderma*, *311*, 143–148.
- 1305 Mayr, T., Rivas-Casado, M., Bellamy, P., Palmer, R., Zawadzka, J., & Corstanje, R.  
1306 (2010). Two methods for using legacy data in digital soil mapping. In M. A. H. A.  
1307 K.-B. S. Boettinger J.L., Howell D.W. (Ed.), *Digital Soil Mapping* (pp. 191–202).  
1308 Springer, Dordrecht, NL.
- 1309 McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping.  
1310 *Geoderma*, *117*, 3–52.



- 1311 McNicol, G., Bulmer, C., D'Amore, D., Sanborn, P., Saunders, S., Giesbrecht, I.,  
1312 Arriola, S. G., Bidlack, A., Butman, D., & Buma, B. (2019). Large, climate-  
1313 sensitive soil carbon stocks mapped with pedology-informed machine learning in  
1314 the North Pacific coastal temperate rainforest. *Environmental Research Letters*,  
1315 *14*, 014004.
- 1316 Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving  
1317 performance of spatio-temporal machine learning models using forward feature  
1318 selection and target-oriented validation. *Environmental Modelling & Software*,  
1319 *101*, 1–9.
- 1320 Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial  
1321 predictor variable selection in machine learning applications—moving from data  
1322 reproduction to spatial prediction. *Ecological Modelling*, *411*, 108815.
- 1323 Micheletti, N., Foresti, L., Robert, S., Leuenberger, M., Pedrazzini, A., Jaboyed-  
1324 off, M., & Kanevski, M. (2014). Machine learning feature selection methods for  
1325 landslide susceptibility mapping. *Mathematical Geosciences*, *46*, 33–57.
- 1326 Miller, B. A., Koszinski, S., Wehrhan, M., & Sommer, M. (2015a). Comparison of  
1327 spatial association approaches for landscape mapping of soil organic carbon stocks.  
1328 *Soil*, *1*, 217—233.
- 1329 Miller, B. A., Koszinski, S., Wehrhan, M., & Sommer, M. (2015b). Impact of multi-  
1330 scale predictor selection for modeling soil properties. *Geoderma*, *239*, 97–106.
- 1331 Miller, T. (2019). Explanation in artificial intelligence: Insights from the social  
1332 sciences. *Artificial Intelligence*, *267*, 1–38.
- 1333 Mira, M., Weiss, M., Baret, F., Courault, D., Hagolle, O., Gallego-Elvira, B., &  
1334 Olioso, A. (2015). The MODIS (collection V006) BRDF/albedo product MCD43D:  
1335 Temporal course evaluated over agricultural landscape. *Remote Sensing of Envi-  
1336 ronment*, *170*, 216–228.
- 1337 Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. [arXiv:  
1338 1411.1784](https://arxiv.org/abs/1411.1784).
- 1339 Molnar, C. (2019). *Interpretable machine learning*. Lulu, Morrisville, USA.
- 1340 Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and  
1341 understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

- 1342 Moran, C. J., & Bui, E. N. (2002). Spatial data mining for enhanced soil map  
1343 modelling. *International Journal of Geographical Information Science*, *16*, 533–  
1344 549.
- 1345 Mosleh, Z., Salehi, M. H., Jafari, A., Borujeni, I. E., & Mehnatkesh, A. (2016). The  
1346 effectiveness of digital soil mapping to predict soil properties over low-relief areas.  
1347 *Environmental Monitoring and Assessment*, *188*, 195.
- 1348 Mulder, V. L., Lacoste, M., Richer-de Forges, A. C., Martin, M. P., & Arrouays, D.  
1349 (2016). National versus global modelling the 3D distribution of soil organic carbon  
1350 in mainland France. *Geoderma*, *263*, 16–34.
- 1351 Nauman, T. W., & Duniway, M. C. (2019). Relative prediction intervals reveal larger  
1352 uncertainty in 3D approaches to predictive digital soil mapping of soil properties  
1353 with legacy data. *Geoderma*, *347*, 170–184.
- 1354 Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L.,  
1355 Schaepman, M. E., & Papritz, A. (2018). Evaluation of digital soil mapping ap-  
1356 proaches with large sets of environmental covariates. *Soil*, *4*, 1–22.
- 1357 Odgers, N. P., Sun, W., McBratney, A. B., Minasny, B., & Clifford, D. (2014).  
1358 Disaggregating and harmonising soil map units through resampled classification  
1359 trees. *Geoderma*, *214*, 91–100.
- 1360 Oldeman, L. R., & Van Engelen, V. W. P. (1993). A world soils and terrain digital  
1361 database (SOTER)—An improved assessment of land resources. *Geoderma*, *60*,  
1362 309–325.
- 1363 Oliver, M. A. (1987). Geostatistics and its application to soil science. *Soil Use and*  
1364 *Management*, *3*, 8–20.
- 1365 Ottoy, S., De Vos, B., Sindayihebura, A., Hermy, M., & Van Orshoven, J. (2017).  
1366 Assessing soil organic carbon stocks under current and potential forest cover using  
1367 digital soil mapping and spatial generalisation. *Ecological Indicators*, *77*, 139–150.
- 1368 Padarian, J., Minasny, B., & McBratney, A. B. (2019). Using deep learning for  
1369 digital soil mapping. *Soil*, *5*, 79–89.
- 1370 Pahlavan-Rad, M. R., & Akbarimoghaddam, A. (2018). Spatial variability of soil  
1371 texture fractions and pH in a flood plain (case study from eastern Iran). *CATENA*,  
1372 *160*, 275–281.

- 1373 Peres-Neto, P. R., Legendre, P., Dray, S., & Borcard, D. (2006). Variation partition-  
1374 ing of species data matrices: estimation and comparison of fractions. *Ecology*, *87*,  
1375 2614–2625.
- 1376 Poggio, L., Lassauce, A., & Gimona, A. (2019). Modelling the extent of northern  
1377 peat soil and its uncertainty with sentinel: Scotland as example of highly cloudy  
1378 region. *Geoderma*, *346*, 63–74.
- 1379 Pohjankukka, J., Pahikkala, T., Nevalainen, P., & Heikkonen, J. (2017). Estimating  
1380 the prediction performance of spatial models via spatial k-fold cross validation.  
1381 *International Journal of Geographical Information Science*, *31*, 2001–2019.
- 1382 Pouladi, N., Møller, A. B., Tabatabai, S., & Greve, M. H. (2019). Mapping soil  
1383 organic matter contents at field level with cubist, random forest and kriging. *Geo-*  
1384 *derma*, *342*, 85–92.
- 1385 Pozdnoukhov, A., & Kanevski, M. (2006). Monitoring network optimisation for  
1386 spatial data classification using support vector machines. *International Journal of*  
1387 *Environment and Pollution*, *28*, 465–484.
- 1388 Rahman, R., Otridge, J., & Pal, R. (2017). IntegratedMRF: random forest-based  
1389 framework for integrating prediction from different data types. *Bioinformatics*,  
1390 *33*, 1407–1410.
- 1391 Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., &  
1392 Thompson, J. (2018). Soil property and class maps of the conterminous United  
1393 States at 100-meter spatial resolution. *Soil Science Society of America Journal*,  
1394 *82*, 186–201.
- 1395 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais,  
1396 N. et al. (2019). Deep learning and process understanding for data-driven earth  
1397 system science. *Nature*, *566*, 195–204.
- 1398 Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explain-  
1399 ing the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*  
1400 *International Conference on Knowledge Discovery and Data Mining* (pp. 1135–  
1401 1144). ACM.
- 1402 Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G.,  
1403 Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W. et al. (2017).  
1404 Cross-validation strategies for data with temporal, spatial, hierarchical, or phylo-  
1405 genetic structure. *Ecography*, *40*, 913–929.

- 1406 Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2019). Explainable machine  
1407 learning for scientific insights and discoveries. [arXiv:1905.08883](https://arxiv.org/abs/1905.08883).
- 1408 Rudiyanto, B., Minasny, Setiawan, B. I., Saptomo, S. K., McBratney, A. B. et al.  
1409 (2018). Open digital mapping as a cost-effective method for mapping peat thickness  
1410 and assessing the carbon stock of tropical peatlands. *Geoderma*, *313*, 25–40.
- 1411 Ruß, G., & Brenning, A. (2010). Data mining in precision agriculture: management  
1412 of spatial information. In *International Conference on Information Processing and  
1413 Management of Uncertainty in Knowledge-Based Systems* (pp. 350–359). Springer.
- 1414 Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyper-  
1415 parameter tuning and performance assessment of statistical and machine-learning  
1416 algorithms using spatial data. *Ecological Modelling*, *406*, 109–120.
- 1417 Scull, P., Franklin, J., & Chadwick, O. A. (2005). The application of classification  
1418 tree analysis to soil type prediction in a desert landscape. *Ecological Modelling*,  
1419 *181*, 1–15.
- 1420 Segal, M., & Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary  
1421 Reviews: Data Mining and Knowledge Discovery*, *1*, 80–87.
- 1422 Sergeev, A., Buevich, A., Baglaeva, E., & Shichkin, A. (2019). Combining spatial  
1423 autocorrelation with machine learning increases prediction accuracy of soil heavy  
1424 metals. *CATENA*, *174*, 425–435.
- 1425 Sharififar, A., Sarmadian, F., Malone, B. P., & Minasny, B. (2019). Addressing  
1426 the issue of digital mapping of soil classes with imbalanced class observations.  
1427 *Geoderma*, *350*, 84–92.
- 1428 Shi, J., Yang, L., Zhu, A., Qin, C., Liang, P., Zeng, C., Pei, T. et al. (2018). Machine-  
1429 learning variables at different scales vs. knowledge-based variables for mapping  
1430 multiple soil properties. *Soil Science Society of America Journal*, *82*, 645–656.
- 1431 Siewert, M. B. (2018). High-resolution digital mapping of soil organic carbon in  
1432 permafrost terrain using machine learning: a case study in a sub-Arctic peatland  
1433 environment. *Biogeosciences*, *15*, 1663–1682.
- 1434 da Silva Chagas, C., de Carvalho Junior, W., Bhering, S. B., & Calderano Filho, B.  
1435 (2016). Spatial prediction of soil surface texture in a semiarid region using random  
1436 forest and multiple linear regressions. *CATENA*, *139*, 232–240.

- 1437 Sinha, P., Gaughan, A. E., Stevens, F. R., Nieves, J. J., Sorichetta, A., & Tatem, A. J.  
1438 (2019). Assessing the spatial sensitivity of a random forest model: Application in  
1439 gridded population modeling. *Computers, Environment and Urban Systems*, *75*,  
1440 132–145.
- 1441 Somarathna, P. D. S. N., Malone, B. P., & Minasny, B. (2016). Mapping soil organic  
1442 carbon content over New South Wales, Australia using local regression kriging.  
1443 *Geoderma regional*, *7*, 38–48.
- 1444 Somarathna, P. D. S. N., Minasny, B., & Malone, B. P. (2017). More data or a better  
1445 model? figuring out what matters most for the spatial prediction of soil carbon.  
1446 *Soil Science Society of America Journal*, *81*, 1413–1426.
- 1447 Song, X.-D., Wu, H.-Y., Ju, B., Liu, F., Yang, F., Li, D.-C., Zhao, Y.-G., Yang, J.-  
1448 L., & Zhang, G.-L. (2020). Pedoclimatic zone-based three-dimensional soil organic  
1449 carbon mapping in China. *Geoderma*, *363*, 114145.
- 1450 Song, Y.-Q., Zhao, X., Su, H.-Y., Li, B., Hu, Y.-M., & Cui, X.-S. (2018). Predicting  
1451 spatial variations in soil nutrients with hyperspectral remote sensing at regional  
1452 scale. *Sensors*, *18*, 3086.
- 1453 Stojanova, D., Ceci, M., Appice, A., Malerba, D., & Džeroski, S. (2013). Dealing  
1454 with spatial autocorrelation when learning predictive clustering trees. *Ecological  
1455 Informatics*, *13*, 22–39.
- 1456 Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-Stitt, S., Buzzo, G., Dumperth,  
1457 C., Wadoux, A., Xiang, W., & Scholten, T. (2016). Incorporating limited field  
1458 operability and legacy soil samples in a hypercube sampling design for digital soil  
1459 mapping. *Journal of Plant Nutrition and Soil Science*, *179*, 499–509.
- 1460 Subburayalu, S. K., & Slater, B. K. (2013). Soil series mapping by knowledge dis-  
1461 covery from an Ohio county soil map. *Soil Science Society of America Journal*,  
1462 *77*, 1254–1268.
- 1463 Szatmári, G., & Pásztor, L. (2019). Comparison of various uncertainty modelling  
1464 approaches based on geostatistics and machine learning algorithms. *Geoderma*,  
1465 *337*, 1329–1340.
- 1466 Szatmári, G., Pirkó, B., Koós, S., Laborczi, A., Bakacsi, Z., Szabó, J., & Pásztor,  
1467 L. (2019). Spatio-temporal assessment of topsoil organic carbon stock change in  
1468 Hungary. *Soil and Tillage Research*, *195*, 104410.

- 1469 Taalab, K., Corstanje, R., Mayr, T., Whelan, M., & Creamer, R. (2015). The  
1470 application of expert knowledge in bayesian networks to predict soil bulk density  
1471 at the landscape scale. *European Journal of Soil Science*, *66*, 930–941.
- 1472 Taghizadeh-Mehrjardi, R., Minasny, B., Sarmadian, F., & Malone, B. P. (2014).  
1473 Digital mapping of soil salinity in Ardakan region, central Iran. *Geoderma*, *213*,  
1474 15–28.
- 1475 Taghizadeh-Mehrjardi, R., Minasny, B., Toomanian, N., Zeraatpisheh, M., Amirian-  
1476 Chakan, A., & Triantafilis, J. (2019a). Digital mapping of soil classes using en-  
1477 semble of models in Isfahan region, Iran. *Soil Systems*, *3*, 37.
- 1478 Taghizadeh-Mehrjardi, R., Nabiollahi, K., & Kerry, R. (2016). Digital mapping of  
1479 soil organic carbon at multiple depths using different data mining techniques in  
1480 Baneh region, Iran. *Geoderma*, *266*, 98–110.
- 1481 Taghizadeh-Mehrjardi, R., Schmidt, K., Eftekhari, K., Behrens, T., Jamshidi, M.,  
1482 Davatgaar, N., Toomanian, N., & Scholten, T. (2019b). Synthetic resampling  
1483 strategies and machine learning for digital soil mapping in Iran. *European Journal*  
1484 *of Soil Science*, .
- 1485 Taghizadeh-mehrjardi, R., Toomanian, N., Khavaninzadeh, A. R., Jafari, A., & Tri-  
1486 antafilis, J. (2016). Predicting and mapping of soil particle-size fractions with  
1487 adaptive neuro-fuzzy inference and ant colony optimization in central Iran. *Euro-*  
1488 *pean Journal of Soil Science*, *67*, 707–725.
- 1489 Tajik, S., Ayoubi, S., Shirani, H., & Zeraatpisheh, M. (2019). Digital mapping of  
1490 soil invertebrates using environmental attributes in a deciduous forest ecosystem.  
1491 *Geoderma*, *353*, 252–263.
- 1492 Tuia, D., Pozdnoukhov, A., Foresti, L., & Kanevski, M. (2013). Active learning for  
1493 monitoring network optimization. In J. Mateu, & W. G. Müller (Eds.), *Spatio-*  
1494 *Temporal Design: Advances in Efficient Data Acquisition* (pp. 285–318). Wiley  
1495 Online Library, Chichester, UK.
- 1496 Tziachris, P., Aschonitis, V., Chatzistathis, T., & Papadopoulou, M. (2019). As-  
1497 sessment of spatial hybrid methods for predicting soil organic matter using DEM  
1498 derivatives and soil parameters. *CATENA*, *174*, 206–216.
- 1499 Uria, B., Côté, M.-A., Gregor, K., Murray, I., & Larochelle, H. (2016). Neural  
1500 autoregressive distribution estimation. *The Journal of Machine Learning Research*,  
1501 *17*, 7184–7220.

- 1502 Vaysse, K., & Lagacherie, P. (2015). Evaluating digital soil mapping approaches for  
1503 mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon  
1504 (France). *Geoderma Regional*, *4*, 20–30.
- 1505 Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate  
1506 uncertainty of digital soil mapping products. *Geoderma*, *291*, 55–64.
- 1507 Vermeulen, D., & Van Niekerk, A. (2017). Machine learning performance for pre-  
1508 dicting soil salinity using different combinations of geomorphometric covariates.  
1509 *Geoderma*, *299*, 1–12.
- 1510 Vincent, S., Lemercier, B., Berthier, L., & Walter, C. (2018). Spatial disaggregation  
1511 of complex soil map units at the regional scale based on soil-landscape relation-  
1512 ships. *Geoderma*, *311*, 130–142.
- 1513 Viscarra-Rossel, R. A., & Chen, C. (2011). Digitally mapping the information content  
1514 of visible–near infrared spectra of surficial Australian soils. *Remote Sensing of*  
1515 *Environment*, *115*, 1443–1455.
- 1516 Viscarra-Rossel, R. A., Chen, C., Grundy, M. J., Searle, R., Clifford, D., & Campbell,  
1517 P. H. (2015). The Australian three-dimensional soil grid: Australia’s contribution  
1518 to the GlobalSoilMap project. *Soil Research*, *53*, 845–864.
- 1519 Wadoux, A. M. J.-C. (2019a). *Sampling design optimization for geostatistical mod-*  
1520 *elling and prediction*. Ph.D. thesis Wageningen University & Research.
- 1521 Wadoux, A. M. J.-C. (2019b). Using deep learning for multivariate mapping of soil  
1522 with quantified uncertainty. *Geoderma*, *351*, 59–70.
- 1523 Wadoux, A. M. J.-C., Brus, D. J., & Heuvelink, G. B. M. (2019a). Sampling design  
1524 optimization for soil mapping with random forest. *Geoderma*, *355*, 113913.
- 1525 Wadoux, A. M. J.-C., Padarian, J., & Minasny, B. (2019b). Multi-source data  
1526 integration for soil mapping using deep learning. *Soil*, *5*, 107–119.
- 1527 Wadoux, A. M. J.-C., Samuel-Rosa, A., Poggio, L., & Mulder, V. L. (2019c). A note  
1528 on knowledge discovery and machine learning in digital soil mapping. *European*  
1529 *Journal of Soil Science*, .
- 1530 Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., & Li Liu, D.  
1531 (2018). High resolution mapping of soil organic carbon stocks using remote sensing  
1532 variables in the semi-arid rangelands of eastern Australia. *Science of The Total*  
1533 *Environment*, *630*, 367–378.

- 1534 Wang, S., Zhuang, Q., Wang, Q., Jin, X., & Han, C. (2017). Mapping stocks of soil  
1535 organic carbon and soil total nitrogen in Liaoning Province of China. *Geoderma*,  
1536 *305*, 250–263.
- 1537 Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment  
1538 of support vector regression, artificial neural networks, and random forests for pre-  
1539 dicting and mapping soil organic carbon stocks across an Afromontane landscape.  
1540 *Ecological Indicators*, *52*, 394–403.
- 1541 Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011). Digital mapping  
1542 of soil organic matter stocks using random forest modeling in a semi-arid steppe  
1543 ecosystem. *Plant and Soil*, *340*, 7–24.
- 1544 Wu, J., Teng, Y., Chen, H., & Li, J. (2016). Machine-learning models for on-site  
1545 estimation of background concentrations of arsenic in soils using soil formation  
1546 factors. *Journal of Soils and Sediments*, *16*, 1787–1797.
- 1547 Xu, S., An, X., Qiao, X., Zhu, L., & Li, L. (2013). Multi-output least-squares support  
1548 vector regression machines. *Pattern Recognition Letters*, *34*, 1078–1084.
- 1549 Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O’Loughlin, F., Neal,  
1550 J. C., Sampson, C. C., Kanae, S., & Bates, P. D. (2017). A high-accuracy map of  
1551 global terrain elevations. *Geophysical Research Letters*, *44*, 5844–5853.
- 1552 Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao,  
1553 Y.-G., & Li, D.-C. (2016). Comparison of boosted regression tree and random forest  
1554 models for mapping topsoil organic carbon concentration in an alpine ecosystem.  
1555 *Ecological Indicators*, *60*, 870–878.
- 1556 Zeraatpisheh, M., Ayoubi, S., Jafari, A., & Finke, P. (2017). Comparing the efficiency  
1557 of digital and conventional soil mapping to predict soil types in a semi-arid region  
1558 in Iran. *Geomorphology*, *285*, 186–204.
- 1559 Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., & Finke, P. (2019). Digital  
1560 mapping of soil properties using multiple machine learning in a semi-arid region,  
1561 central Iran. *Geoderma*, *338*, 445–452.
- 1562 Zhu, M., Feng, Q., Zhang, M., Liu, W., Deo, R. C., Zhang, C., & Yang, L. (2019).  
1563 Soil organic carbon in semiarid alpine regions: the spatial distribution, stock esti-  
1564 mation, and environmental controls. *Journal of Soils and Sediments*, *19*, 1–15.