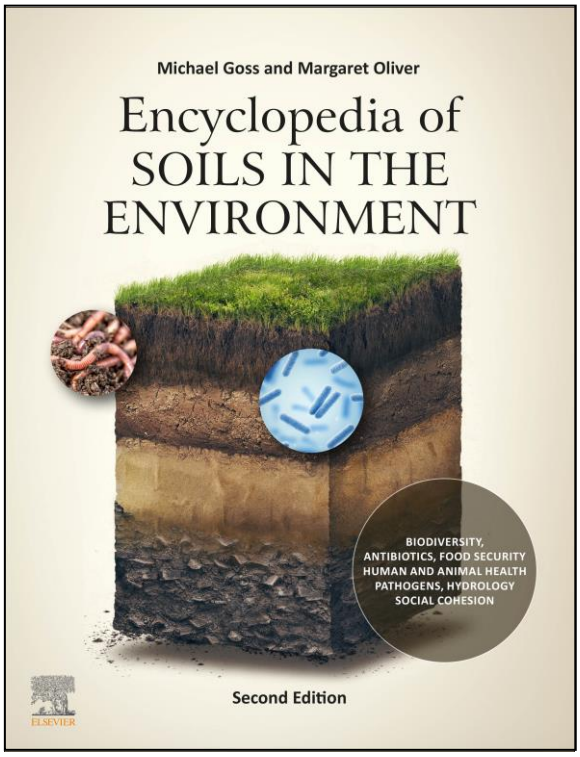

Malone Brendan, Arrouays Dominique, Poggio Laura, Minasny Budiman and McBratney Alex B. (2023) Digital soil mapping: Evolution, current state and future directions of the science. In: Goss, Michael and Oliver, Margaret (eds.) *Encyclopedia of Soils in the Environment Second Edition*. vol. 4, pp. 684-695. UK: Elsevier.

dx.doi.org/10.1016/B978-0-12-822974-3.00130-0

Digital soil mapping: Evolution, current state and future directions of the science

Brendan Malone[a], Dominique Arrouays[b], Laura Poggio[c], Budiman Minasny[d], and Alex B McBratney[d], [a]CSIRO Agriculture and Food, Butler Laboratory, Black Mountain, Acton, ACT, Australia; [b]INRAE, InfoSol, Orléans, France; [c]ISRIC World Soil Information, Wageningen, Netherlands; [d]Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia

Abstract

Digital soil mapping (DSM) entails the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge derived from related environmental variables. This chapter summarizes the state of the science of DSM and the mechanics of its various implementations. It has moved from research to practical application and is now widely used as a method for making soil maps. The outlook for continuing research into DSM and related efforts is positive, because objective spatio-temporal assessments of the soil resources are needed and continually require updating.

Key points

- This chapter describes some of the underpinning mechanics of digital soil mapping, celebrates some of its successes, and identifies areas of the science which need further thought and work.
- Digital soil mapping has been borne out and developed alongside technological developments which include advances in computers and computer sciences, together with unparalleled growth in associated modelling technologies and widespread availability of digital Earth observation data.
- Digital soil mapping augments rather than replaces conventional soil survey and is recognized as an additive technology that enables delivery of richer spatial soil information and has the potential to deliver products customized to the needs of end-users.

Introduction

In recent times, we have witnessed the advancement of the computer and information technology ages. With such advances, vast amounts of data and tools in all fields of endeavor have become available. This has motivated numerous initiatives around the world to build spatial data infrastructures aiming to facilitate the collection, maintenance, dissemination and use of spatial information.

The science of digital soil mapping has been borne out of and developed alongside these technological developments. Digital soil mapping (DSM) is a numerical and model-based framework for inferring the spatial and temporal variation of soil types and properties from soil observation and knowledge derived from related environmental variables. Digital soil mapping has never been considered a replacement to traditional forms of soil survey and mapping, rather as a complementary tool, albeit quantitative in principle, to characterize the spatial and temporal properties of soils in ways that legacy soil mapping could not achieve.

There are many aspects to DSM, including its generalizations given different forms of soil information, diverse modelling frameworks and accommodation for both 2- and 3-dimensional vertical support, and the quantification of uncertainties, all of which are discussed in this chapter.

Digital soil mapping has matured as a science and is used as an operational tool throughout the world for the assessment and characterization of soils within both government and privately funded organizations.

Globally, the need for comprehensive and specific digital soil information has increased, particularly amidst emerging existential crises of climate change, land degradation, and unsustainable land management practices in an increasingly populous world. Digital soil mapping will continue to evolve together with the broader technological advancements. More importantly, it will also branch out further into digital soil assessment, for example in the quantification of soil function and ecosystem services. It is this derived information that is required for assessing the status of ecosystems more generally, rather than just the creation of digital maps of soil characteristics and properties. To date, we are already down that path, and well positioned to confront these environmental challenges given the large and active network of practitioners throughout the world.

Digital soil mapping concepts

The various threads in research on soil spatial prediction began to converge in the early 2000s. The development of fuzzy logic and geostatistics and then soil–landscape quantification in the 1990s, brought about a strong theoretical foundation that enabled modern digital soil mapping to emerge rapidly at the turn of the millennium. Two seminal papers were published in 2003 that summarized historical developments, albeit from different perspectives. Scull et al. (2003) wrote about past research in soil spatial prediction from a physical geography perspective and coined the term predictive soil mapping to describe the general approach. About the same time, McBratney et al. (2003) generalized the diversity of quantitative mapping approaches into a framework they called 'scorpan.' This uses a Jenny (1941) clorpt-like framework not for mere explanation but rather empirical quantitative description of soil–landscape relationships for spatial prediction. They called these activities digital soil mapping (DSM).

The 'scorpan' framework is more formally known as scorpan-SSPFe, which includes soil spatial prediction functions (SSPF) and autocorrelated errors (McBratney et al., 2003). The scorpan factors are:

s: soil, its classes, or properties.
c: climate, including precipitation and temperature.
o: organisms, including vegetation, fauna, and other factors of the biotic environment.
r: relief, or topography.
p: parent material, including lithology.
a: age.
n: space, or spatial position.

The model is written as:

$$S_c = f(s, c, o, r, p, a, n) \text{ or } S_p = f(s, c, o, r, p, a, n)$$

where S_c is soil classes and S_p is soil properties or attributes. Considering soils are sampled at spatial coordinates x, y at an approximate point in time, $\sim t$, the model can be expressed explicitly as:

$$S[x, y, \sim t] = f(s[x, y, \sim t], c[x, y, \sim t], o[x, y, \sim t], r[x, y, \sim t], p[x, y, \sim t], a[x, y, \sim t], n[x, y, \sim t]).$$

Soil (s) is included as a factor because soil can be predicted from its properties, and soil properties from its class or other properties (McBratney et al., 2003). This additional soil information could be gathered from a prior soil map or from either remote or proximal soil sensing or even expert knowledge. The factor n means that soil can be predicted as a function of spatial position alone, as in the case of kriging, but it may also be predicted as a function of the distance from some landscape features such as streams, hilltops, roads or point sources of pollution, etc. Information that may represent the other scorpan factors is described in Table 1.

Table 1 Possible sources of information to represent the scorpan factors.

Scorpan factor	*Possible representatives*
s	Legacy soil maps, point observations, expert knowledge
c	Temperature and precipitation records, remote sensing derive climate data
o	Vegetation maps, species abundance maps, yield maps, land use maps, remote sensing derived vegetation indices
r	Digital elevation model, terrain attributes
p	Legacy geology maps, gamma radiometric information
a	Weathering indices, geology maps
n	Latitude and longitude or easting and northing, distance from landscape features, distance from roads, distance from point sources of pollution

Since the establishment of DSM in the early 2000s, there have also been developments and huge leaps forward in terms of Earth observation data from both remote and proximal sensing platforms. This has resulted in a proliferation of high-resolution environmental spatial data, and many of these can be used to represent various scorpan factors. Digital elevation models (DEMs) and satellite imagery data are two prominent examples. Subsequently, DSM techniques have been deployed to build or populate spatial soil information systems from relatively sparse datasets from the ground up. The DSM techniques have also been used to update or renew existing soil maps (e.g. Kempen et al., 2009). It is seen as a practicable framework for fulfilling the current and future demand for relevant soil information (Searle et al., 2021).

In practice, DSM starts by defining a mapping domain, and soil samples are taken based on an established sampling design. Several sampling designs for DSM have been developed, such as stratified sampling using *k*-means algorithm or conditioned Latin hypercube sampling. This is followed by intersecting the observations with a set of pedologically meaningful environmental layers. Then, fitting a mathematical or machine learning model to obtain a spatial soil prediction function. Once the model has been fitted at the observation points, it is then extended to all grid cell nodes of the raster layers giving a digital soil map. This three-component process is the hallmark of DSM (Minasny and McBratney, 2016), which entails:

1. The input which includes soil data and associated environmental covariates.
2. The modelling process, and
3. The output, i.e. the digital soil map with uncertainty.

This DSM approach is quite distinct from earlier notions of digital soil mapping that simply involved the digitization of conventional soil maps by electronic scanning. A more appropriate term for this would be digitized soil map.

Soil spatial prediction functions

The modelling step is crucial in the digital soil mapping process and fundamentally distinguishes it from digitized soil mapping. The form of the spatial soil prediction function $f()$ is usually determined at the outset. When deciding which mathematical model is the most appropriate for a given application, several factors are usually considered, including:

1. The operator's familiarity with the model.
2. The model's ease of application in the context of the project, the availability of covariates and the idiosyncrasies of the available soil information.
3. The model's complexity and its power to capture potentially complex soil–landscape relationships within the target mapping domain.

Many mathematical models are available to represent $f()$, and their development continues with advances in statistical and machine learning models. In general, some models, such as multiple linear regression or regression trees, are suited to modelling continuous variables whereas others, such as logistic regression or classification trees, are suited to modelling ordinal or nominal categorical data.

Some of the simplest models are simple linear models with either ordinary or generalized least-squares fitting. More complex models include generalized linear and additive models. Logistic regression models are a type of generalized linear model suited to modelling categorical variables. Recursive partition models such as classification and regression trees are particularly favorable because of their non-parametric structure and their capacity to deal with non-linear relationships between soil data and environmental covariates. For similar reasons machine learning algorithms (such as random forests and support vector machines to mention only two) and neural networks are often considered, as are deep convolutional neural networks, particularly in terms of incorporating spatial contextual information (Padarian et al., 2019) for which these algorithms are suited. Collectively, these model structures provide DSM practitioners with a rich and varied suite of options to explore complex soil spatial variation.

In terms of handling any spatial autocorrelation in the residual (e) that is likely to result from fitting a scorpan model, regression kriging (alternatively scorpan kriging; McBratney et al., 2003) or universal kriging could be used. Coupling a machine learning model with geostatistical modelling of the residuals is also a generalized regression kriging approach that is often used in digital soil mapping studies.

Lark et al. (2006) described a REML–EBLUP (residual maximum likelihood–empirical best linear unbiased predictor) model. The REML–EBLUP is intrinsically like regression-kriging in that both are mixed models where the observed data are modelled as the additive combination of fixed effects (the secondary environmental data), random effects (the spatially correlated residuals e), and independent random error. The difference is that REML estimates the parameters of the trend and covariance functions without bias. These parameters are then used in the EBLUP i.e. a general linear mixed model.

The application of formal geostatistical modelling approaches to categorical variables, such as soil classes, is relatively limited. Kempen et al. (2012) reminded us that the popular methods for categorical prediction of soil types to create a digital soil type map—multinomial logistic regression, classification tress etc.—are non-spatial models i.e. we do not consider spatial locational properties of the target variable as for continuous variables. Subsequently, they explored a generalized linear geostatistical model framework that addressed this issue. The method was theoretically appealing, but was computationally cumbersome and ultimately did not yield significant gains in accuracy when compared to a non-spatial multinomial logistic regression model.

Assessing digital soil map quality

In DSM, a model is created, and a map is produced, which then needs to be verified in terms of whether the resultant spatial patterns appear plausible across the landscape. To date this is a largely qualitive assessment, and usually requires expert elicitation or an assessment by the model creator who may be familiar with the area mapped and will be able to verify accordingly.

Quantitative assessments of map quality are usually done by assessing the fidelity between observations (observed data of soil phenomena) with corresponding co-located model predictions. For continuous variables, goodness of fit statistical measures are based on some form of deviation assessment such as bias and root mean square error between observations and predictions (Hastie et al., 2009). For categorical data, the measures include overall, user and producer accuracies together with indices such as the kappa coefficient.

For completely unbiased assessments of model quality, it is ideal to have an additional data set that is completely independent of the model data. When validating trained models with some sort of data sub-setting mechanism, keep in mind that the validation statistics will be biased. As Brus et al. (2011) explain, sampling from the target mapping area to be used for DSM is often from legacy soil survey. Consequently, these data are unlikely to have been collected using a probability sampling design. Therefore, the sample will be biased i.e. not a true representation of the total population. Thus, an independent probability sample is required. It is recommended that some random sampling from the target area be conducted, such as simple random sampling and stratified simple random sampling.

From an operational perspective it is usually difficult to arrange the additional costs of organizing and implementing some sort of probability sampling for determining unbiased model quality assessment. The alternative is to perform some sort of data sub-setting, such that with a data set split it into a set for model calibration and another set for validation. This type of procedure can take different forms: the two main ones being random-hold back and leave-one-out-cross-validation (LOCV). Random-hold back is where we may sample a data set by some pre-determined proportion (say 70%) which is used for model calibration. We then validate the model using the other 30% of the data. If the dataset is small, there could be a random chance that the model could overstate or falsely represent model confidence. To avoid such outcomes, this random split procedure needs to be performed many times, and the mean of such validation will provide a more representative assessment.

For *k*-fold validation we divide the dataset into equal sized partitions or folds, with all but one of the folds being used for the model calibration; the remaining fold is used for validation. We could repeat this *k*-fold process several times, each time using a different random sample from the data set for model calibration and validation. This allows one to derive distributions of the validation statistics efficiently to assess the stability and sensitivity of the models and parameters. A variant of *k*-fold cross-validation (CV) is spatial cross-validation where, instead of random subsets of data to act as the different folds, the data are clustered spatially into groups (which we can think of as folds to keep with the general idea).

Quantifications of uncertainty

It is relatively easy to quantify the accuracy of a digital soil map using a set of soil observations set apart from the spatial modelling process for this purpose. A range of statistical tests is available and provides a global appraisal of model performance: that is, they typically cannot tell us anything about the performance of a model at a specific location in the prediction area.

On the other hand, the local appraisal of a model's performance can be done through examination of uncertainties if they are available. Such uncertainties can be quantified at each grid cell across the prediction area and are often expressed in the form of a prediction variance or prediction interval for soil attributes, or even probability estimate for soil classes or exceedance thresholds (Fig. 1).

Uncertainties may be computed with a range of methods. For example, it is relatively straightforward to compute the kriging variance with geostatistics in the case of purely spatial models of soil variation. Minasny et al. (2011) demonstrated a model-based Bayesian approach, while Bayesian networks (Taalab et al., 2015) offer a more expert-driven approach.

Machine-learning methods, and particularly those based on iterative resampling and boosted model fitting, can provide empirical estimates of uncertainty. An example here is the quantile regression forest algorithm that was used in France by Vaysse and Lagacherie (2017).

Methods based on model perturbation through data resampling or bootstrapping have been shown to be effective for quantifying model uncertainties, particularly where the number of model parameters is high, or for very large mapping extents the use of model-based geostatistical approaches is computationally prohibitive. Malone et al. (2011) demonstrated another empirical approach which is based on data resampling and fuzzy *k*-means with extragrades.

In quantifying prediction uncertainties, we explicitly acknowledge that a digital soil map is not free from error. A major source of error is the sparseness of soil data, both in the landscape space and the attribute space. Often the uncertainty created by this error is of a magnitude that would preclude the use of a digital soil map in many situations where fine precision is a requirement. Consequently, decisions or policies developed based on mapping need to be made with a certain amount of risk, although the risk can often be quantified. To reduce this risk, the uncertainties may be used to prioritize resources for data collection or direct the application of alternative modelling approaches to improve digital soil maps and reduce uncertainty.

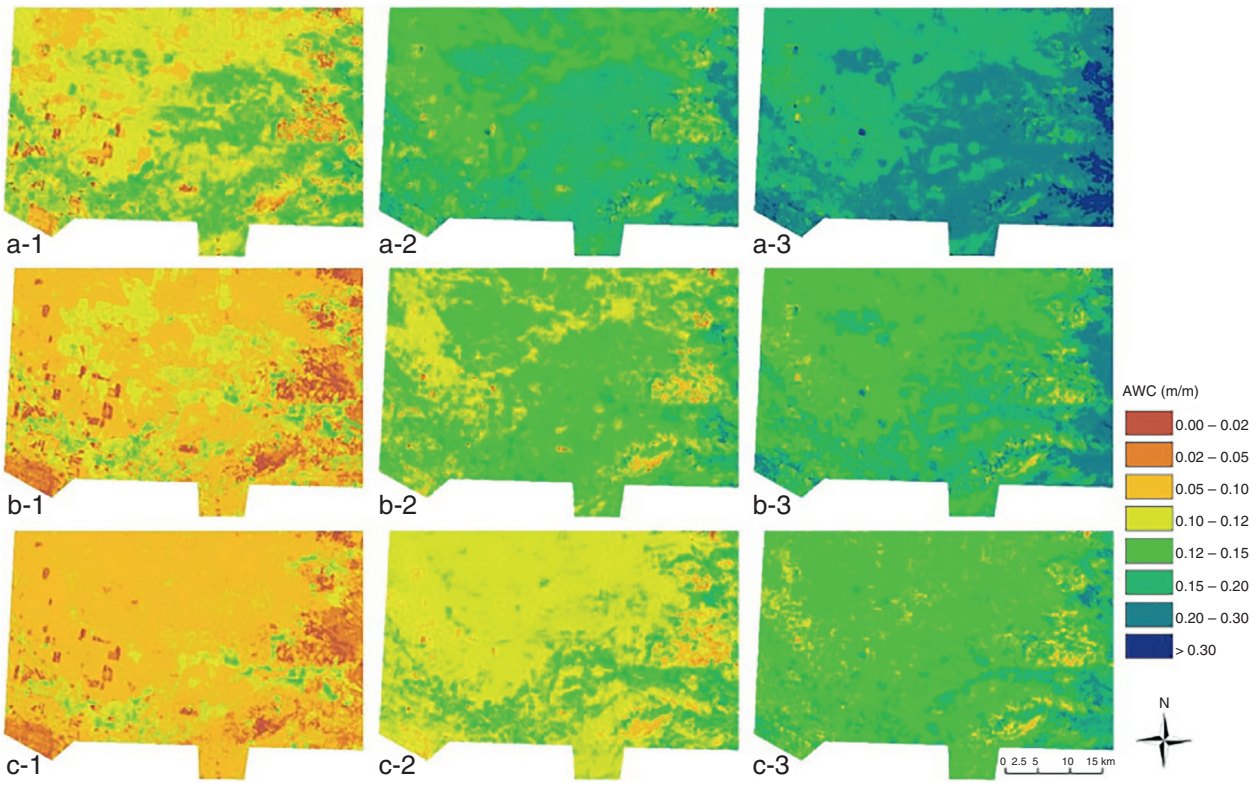

Fig. 1 Variability of AWC at 0–10 cm, 30–40 cm and 80–100 cm across the Edgeroi study area. Lower prediction limit (1), DSM final prediction (2), and upper prediction limit (3). This figure is an example of presenting digital soil map uncertainty in terms of a prediction interval with an expressed level of confidence (here 95%). These estimates are derived for every mapping pixel where one can relatively easily distinguish areas of high prediction certainty (lower prediction interval range) from low prediction certainty (high prediction interval range). From Malone BP, McBratney AB, and Minasny B (2011) Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma* 160: 614–626 and from https://www.sciencedirect.com/science/article/pii/S0016706110003666, Fig. 9.

Nuances of digital soil mapping

Knowledge-based inference and mechanistic modelling for digital soil mapping

An alternative to purely empirical modelling is knowledge-based inference. Knowledge-based inference enables the digital soil mapper to integrate expert knowledge into the mapping process. One of the best known knowledge-based tools is the Soil Land Inference Model or SoLIM (Zhu et al., 1997). SoLIM allows an expert to create membership functions manually that describe the presumed relationship between specific soil types and a range of environmental and topographic variables. With these membership functions, the expert can make predictions of soil type or properties at unobserved locations by a weighted estimate that is based on an environmental similarity score for each soil member. Subsequently, the appealing concept of fuzziness between soil objects is maintained, as well as some quantitative means for assessing the uncertainty of mapped predictions.

In similar work, Bui (2004) pointed out that the soil map legend, which is a representation of the distilled knowledge of the soil surveyors' mental model of soil variation across a mapping domain, contains valuable and important structured language that can be used for automating soil mapping if spatial cover of environmental predictors are available. Probably the closest empirical relative to this approach is the decision tree type model, where data are recursively partitioned to minimize some predictive variance.

The idea of utilizing existing soil maps in new, quantitative ways has also found application in the spatial disaggregation of soil map units, where the information contained in these map units can be downscaled to make spatial predictions of the map units' constituent soil types. This can be further enhanced by the incorporation of expert elicited rules about specific soil–landscape relationships that occur across the study area (e.g. Vincent et al., 2016).

Bayesian networks, described by Taalab et al. (2015), also enable the operator to include expert knowledge explicitly. They are also able to be tuned empirically. The methodology first requires the need to express prior knowledge, in the form of probabilities, about the interaction between a target variable and a set of environmental covariates. With these prior probabilities together with Bayesian inference, one can estimate either continuous or categorical variables and then use the derived posterior probabilities as a quantitative measure of uncertainty.

Structural equation modelling (SEM) has its roots in social sciences research and Angelini et al. (2016) have demonstrated its application to digital soil mapping. It enables the integration of empirical information with mechanistic knowledge by deriving the model equations from known causal relationships, while estimating the model parameters using the available data. This approach necessitates a thorough understanding of soil processes. Moreover, these models can predict many soil properties simultaneously, while preserving the relationships between them. The recognition that soil characteristics are not independent variables, which is how they are treated in most DSM model applications is a welcome advancement to DSM. In data science and machine learning, however, advances are continually being made and have been demonstrated by Wadoux (2019) who used a convolutional neural network model for simultaneous prediction of soil texture, organic carbon, pH, and nitrogen across France. This purely empirical model, had the advantage of quantifying prediction uncertainties, which is a general requirement for DSM and to date has not been investigated for SEM.

The future directions of DSM are likely to entail a deeper integration with soil pedology either by pure- or semi-mechanistic approaches like those described above. Ma et al. (2019) reviews the literature on this and provide insight on how both pedology and DSM can be enriched with knowledge and ideas from each science domain.

Space–time modelling

The spatio-temporal modelling of dynamic soil properties such as soil carbon and soil moisture (plant available water) necessitates an integration of process-based knowledge and empiricism and is at the frontier of DSM research. To date most spatio-temporal models of soil carbon are based around space-for-time substitution models (Gray and Bishop, 2019), where a projection of soil change because of climate or landuse change can be calculated readily. Dynamic modelling such as through RothC (soil carbon model) and biomass growth models for more explicit accounting of the biogeochemical processes underpinning soil organic carbon spatio-temporal variability have been trialled (Lee et al., 2020).

Similarly, a better understanding of the spatio-temporal controls of soil moisture variation for real time estimation and then its forecasting, and mechanistic understanding of soil–water dynamics is paramount. Ensemble Kalman filtering is one of the favored approaches for this type of work and its application in an agricultural setting has been demonstrated by Huang et al. (2017).

Soil depth functions and digital soil mapping

Soil survey, soil classification and conventional mapping of soils consider soil as a three-dimensional (3-D) entity or body. Most initial DSM research tended to focus only on the 2-D sense of this general concept where predictions of soil property variation were made for single depth intervals or horizons (and predominantly only from the topsoil).

It is important to attempt to map the variation of soil properties in both the lateral and vertical dimensions. For carbon accounting and understanding the carbon sequestration potential of soil, determining the amount of water soils can hold across a field or watershed, determining the depth to an impeding layer for crop growth across a farm, and investigations of soil acidity, will require some understanding of how soil properties vary with depth.

Ponce-Hernandez et al. (1986) in Malone et al. (2009) suggested that soil properties vary more-or-less continuously with depth. The variation is often anisotropic for certain properties such as carbon and soil texture which may be the result of land use activity or the gravitational vector of profile weathering and development or both. Exceptions to continuous soil property variation with depth is where there is strong anthropogenic (cultivation, removal, and replacement of soils), geologic (contrasting parent materials), and pedological (the development of clear and abrupt soil horizons) forcing for which sharp discontinuities in the depth distribution of soil properties will occur.

Empirical functions describing the depth distribution of soil properties include linear and polynomial functions, exponential and logarithmic functions. Asymmetric peak functions (Myers et al., 2011) and smoothing splines are further depth functions. A special case of the smoothing spline is the pycnophylactic (mass preserving) type as investigated by Bishop et al. (1999) in Malone et al. (2009). Mass preserving splines model the continuous variation of soil properties with depth while maintaining the average of the observed property through the observed horizons or layers. The proviso for a 'good fit' is that numerous observations at regular depths are required. For attributes such as soil organic carbon, it is necessary to have observations close to the surface to avoid over-estimating the contents in the very top few centimeters. Conversely, for some contaminated undisturbed soils with retained elements such as lead (Pb) observations close to the surface are necessary to avoid 'diluting' the effect of the large concentration at the surface.

The coupling of soil depth functions with DSM seems an intuitive advance toward understanding soil variation in all its spatial dimensions. A study by Minasny et al. (2006) reported in Malone et al. (2009) used the negative exponential depth function to describe variation in soil carbon concentration with depth in the Edgeroi area, Australia. The authors modelled the parameters of the exponential function using a modified neural network approach, then predicted parameters of the exponential function over the whole area, which enabled them to calculate the carbon distribution within the profile and the storage of carbon at any depth.

One of the limitations of using equivalents of the negative exponential depth function is that the function is only useful for certain soil properties, such as soil organic carbon, that naturally have that type of anisotropic variation down the soil profile. Polynomial soil depth functions also need to be considered carefully because the value at one depth will also affect the fit of the curve at other depths.

Malone et al. (2009) described a regression kriging approach using neural networks coupled with mass-preserving splines for mapping available water capacity and soil carbon stocks in Australia. The approach is performed in two steps with the first being the standardization of depth intervals given a collection of soil profile data with the spline. Integrating the spline for the specified depth intervals, for example 0–5 cm, 5–15 cm, 30–60 etc., an average value was defined for each depth for each soil profile. Not only are these values the predicted means for a specified depth interval, but they are also parameters of the spline function that can be used for subsequent refitting and soil information interrogation. After the splines are fitted, for each specified depth interval a spatial model is fitted which is then used for mapping. With such a coupling of depth function and DSM, Malone et al. (2009) were able to generate scenarios such as the depth at which the cumulative total of soil carbon was equal to 5 kg m^{-2} among other queries (Fig. 2).

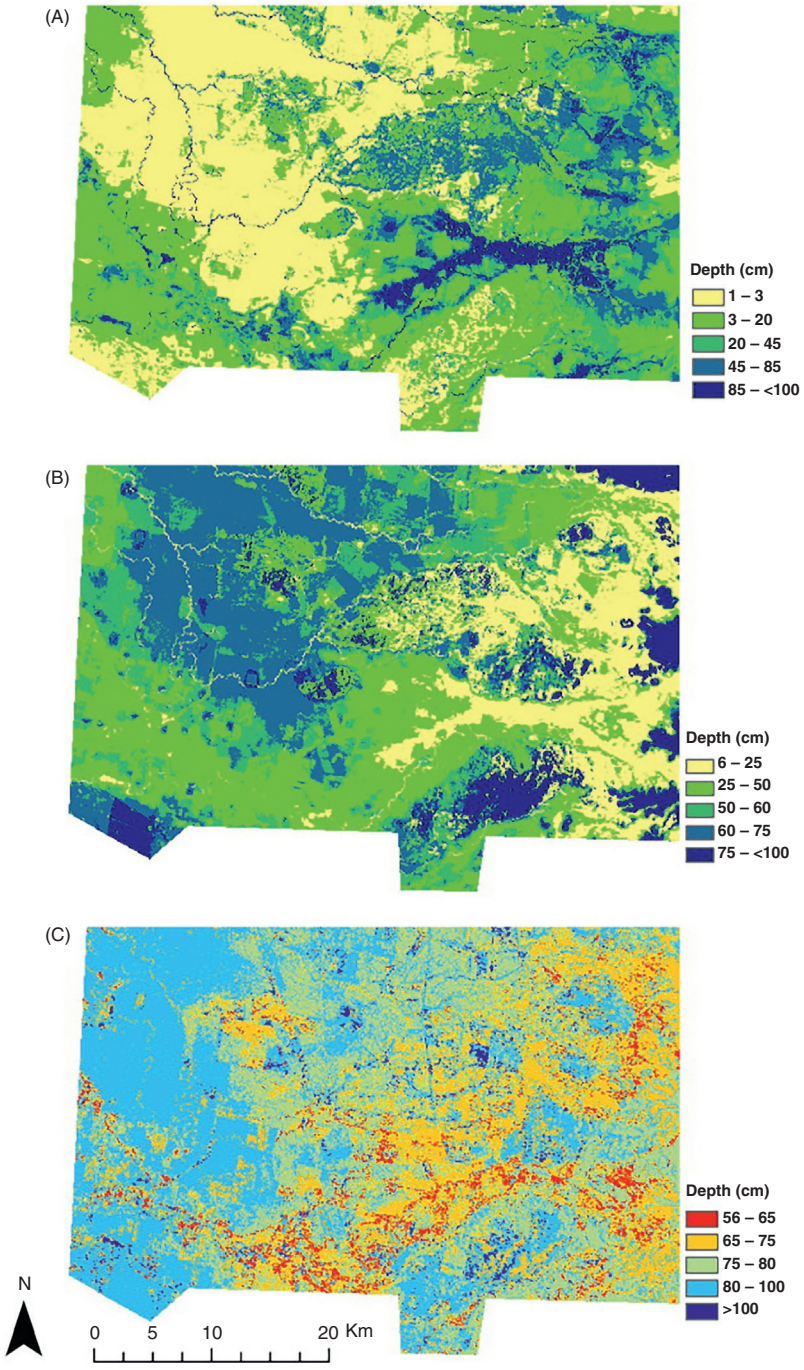

Fig. 2 Maps of scenario-based queries by Malone et al. (2009) in the Edgeroi district in northern New South Wales, Australia. (A) Depth at which soil carbon decreases to below 1%. (B) Depth at which cumulative total of soil carbon equals 5 kg m^{-2}. (C) Depth at which cumulative sum of AWC is 100 mm. From https://www.sciencedirect.com/science/article/pii/S0016706109003255, Fig. 9.

Poggio and Gimona (2014) describe a one-step approach to 3-D digital soil mapping using a hybrid generalized additive model (GAM) for carbon stocks in Scotland. One-step approaches are appealing because they reduce the complexity of workflow. With the Poggio and Gimona (2014) approach they avoided standardizing soil prediction depth intervals i.e. they used the observed values, and then modelled the 3-D trend in soil variation with a GAM and an associated 3-D smoother with related covariates. The one-step approach was also appealing to Orton et al. (2016), who introduced an area-to-point kriging methodology. This model could include all the data in one statistical analysis, and importantly maintained the integrity of the sample support of the soil profile data as well as providing an explicit methodology for quantifying the uncertainty of the fitted soil depth function.

Kempen et al. (2011) developed a depth function that combines general pedological knowledge with geostatistical modelling. They modelled the distribution of soil organic matter content based on typical horizons from 10 soil types. Five depth function building blocks were defined, and for each soil type the depth function structure was obtained by stacking a subset of modelled horizons. The parameters of the depth function for each of the horizons were interpolated using a geostatistical procedure that combined environmental information. While pedologically the most appealing approach, the horizon depth function method of Kempen et al. (2011) could be limiting because its application is limited to the spatial extent in which it was developed, or in similar landscape contexts or soil types.

Generalizing the practice of digital soil mapping

While most applications of DSM are based on using observed soil point data coupled with environmental covariates and a spatial predictive model function, it is certainly not restricted to this process. For example, some previous discussion of soil map disaggregation has been made where mapping units are disaggregated into their constituent classes or series whose spatial pattern is determined by a spatial prediction function.

Before exploring that concept in further detail, it is worth pointing out the various possible scenarios that could be encountered for DSM. They are described by Minasny and McBratney (2010) who presented a decision tree for DSM methodology based on the nature of available legacy soil data (Fig. 3). This tree was proposed as a general framework to aid the delivery of a digital global soil map. It can be viewed as guidance to practitioners in "what do I do?" situations. Once the practitioner has defined an area of interest and assembled a suite of environmental covariates for that area, depending on what available data there are to use, there are suggested approaches that could be implemented.. The most common approach is scorpan kriging which can be performed when there are only point data, but can also be used when there are both point and map data available. The idea here is that scorpan

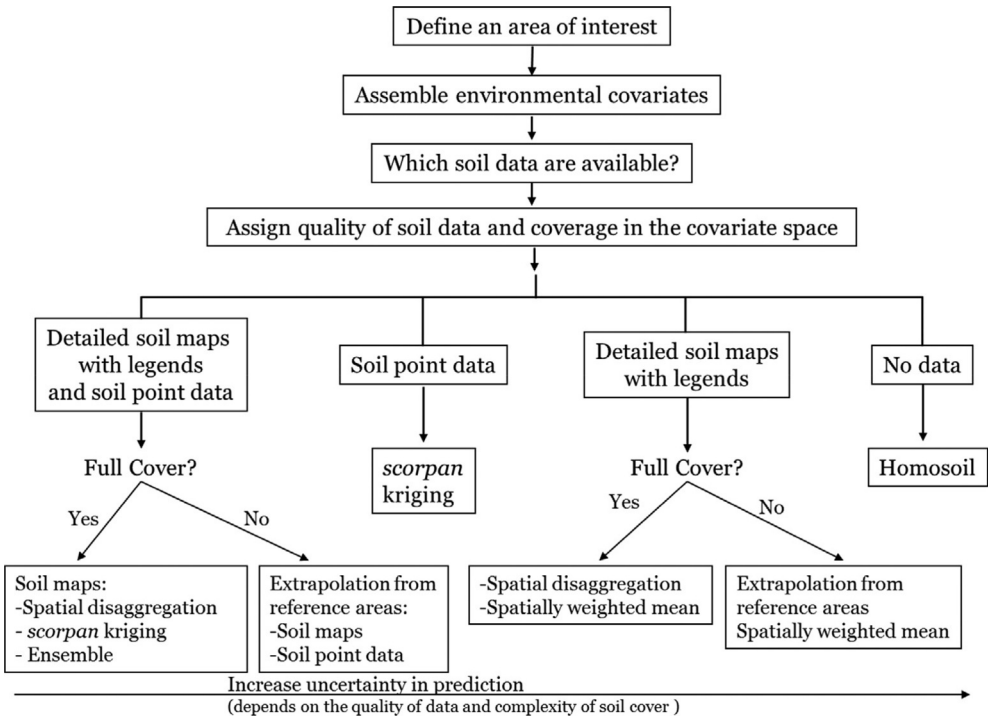

Fig. 3 Decision tree of digital soil mapping approaches. Adapted from Minasny B, and McBratney AB (2010) Methodologies for global soil mapping. In: Boettinger JL, Howell DW, More AC, Hartemink AE, and Kienast-Brown S (eds.) *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation.* London: Springer.

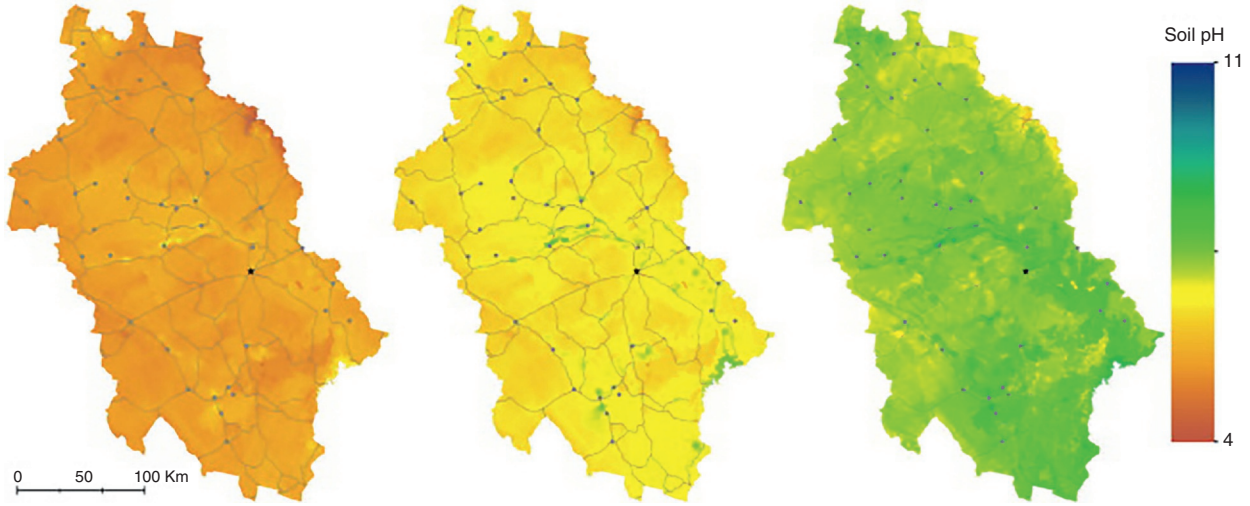

Fig. 4 Model averaging outputs from Malone et al. (2014). Model averaged digital soil maps of soil pH for the 0–5-cm depth interval across the Dalrymple Shire, QLD, Australia. These maps were produced using Granger–Ramanathan (GRA) model averaging, that both corrects bias and preferentially weights inputs based on comparative model accuracies. Inputs were a soil pH map with estimates from using point-based digital soil mapping (scorpan-kriging) and from disaggregated conventional map approaches (DSMART + PROPR) described in Odgers et al. (2014, 2015). The center map represents the combined prediction, while the left and right maps respectively are the 90% lower and upper prediction limits. From https://www.sciencedirect.com/science/article/pii/S0016706114001906, Fig. 5.

kriging and soil map disaggregation are fused through a combinatorial approach such as model averaging. Malone et al. (2014) exemplified such a procedure in an area of Queensland, Australia (Fig. 4).

The options are quite different when only soil map information is available. Consider that the quality of the soil maps depends on the scale and subsequent variation in soil cover; such that smaller scale maps e.g., 1:100,000 would be considered better and more detailed than large scale maps e.g. 1:500,000. The elemental basis for extracting soil properties from legacy soil maps comes from the central and distributional concepts of soil mapping units. For example, modal soil profile data of soil classes can be used to create soil property maps quickly. An early example of this is the multilayer soil characteristics dataset for the conterminous United States (CONUS–SOIL), which has been and continues to be used in many climate, hydrology and land surface models (Miller and White, 1998 in Odgers et al., 2014). Where mapping units consist of more than one component, we can use a spatially weighted means type method, i.e. estimation of the soil properties is based on the modal profile of the components and the proportional area of the mapping unit each component covers. As a pre-processing step prior to creating soil attribute maps, it may be necessary to harmonize soil mapping units (in the case of adjacent soil maps) and or perform some type of disaggregation technique to retrieve the map unit's component information.

More recently, research into the spatial disaggregation of soil map units has emerged as a means of downscaling choropleth soil mapping. For example, the DSMART algorithm (Odgers et al., 2014) spatially disaggregates a choropleth map by iteratively resampling it to create a series of classification trees that create realizations of the potential soil class distribution. The realizations of soil class distribution are merged to estimate the probabilities of occurrence for all the soil classes in the choropleth map area. These probabilities together with modal soil profile characterizations associated with each soil class can then be integrated to derive soil attribute maps (such as pH, soil texture fractions etc.) as shown by Odgers et al. (2015). Modifications to DSMART by explicit inclusion of soil–landscape relationships, which assist and or constrain the allocation of likely soil types within a soil mapping unit, has been demonstrated with some success in the French region of Brittany (Vincent et al., 2016).

How do we do digital soil mapping in an area that does not have any soil data? In such areas ('recipient' areas) it may be possible to extrapolate using a model that was constructed in an area with available soil data (a 'donor' area). The rationale, called homosoil (Mallavan et al., 2010), is that if the recipient area is sufficiently homologous to the donor area with respect to the strength and expression of soil forming factors then extrapolation is feasible without an unacceptable degree of uncertainty. Identification of potential donor areas is based on computation of the similarity to the recipient area in terms of the relevant soil forming factors.

Operationalization of digital soil mapping

Since the end of the first decade of the 21st century, DSM has matured as a science and in some places worldwide is used as an operational tool for soil mapping agencies. In Australia for example, Kidd et al. (2020) describe the evolution of DSM from research undertakings within governmental agencies to a nationally coordinated program to deliver a freely available, comprehensive National digital soil information system and an active community of practitioners in both public and private funded organizations. Similar experiential pathways have been encountered elsewhere and new national, continental and global digital soils information facilities have also emerged similarly (Chen et al., 2022).

Not all operational work has focused on the development of digital soil maps ab initio. For example, government-sponsored work has led to the updating of legacy maps using digital techniques (Kempen et al., 2009). Agencies in other jurisdictions are working to integrate digital soil mapping methods with existing survey procedures. The relatively recent phenomena of national and global scale DSMs have been motivated through coordinated efforts by a series of international conferences and the GlobalSoilMap initiative (Arrouays et al., 2020a).

The GlobalSoilMap initiative

Soil mapping at the global extent is not a recent or new endeavor. The FAO–UNESCO soil map of the world was the first world soil map published (1981) that used a single map legend which accounted for the global diversity of soils. In around 2006 a proposal for a new global grid of the most important soil functional properties was made. Later that year, the GlobalSoilMap.net consortium was formed to create a new, high-resolution digital soil map of the world using state-of-the-art and emerging technologies. The consortium was inspired in part because of "policy-maker's frustrations" about not being able to get quantitative answers to questions such as how much carbon is sequestered or emitted by soils in a particular region? Or, what is its impact on biomass production and human health? Or how do such estimates change over time?

Technical specifications of the GlobalSoilMap are given in the specifications document (Arrouays et al., 2014). This publication articulates not how digital soil maps should be created, but the standard to which they should conform to permit collation for the assemblage of a global product. This specification promotes innovation in digital soil mapping practice and recognizes that practitioners may prefer certain methods, and that different methods may perform better in different environments.

Key aspects of the specifications include the spatial entity such as the data support and resolution at which maps should be created, the soil properties to be predicted, the date associated with their prediction, and an explicit communication of the prediction uncertainty and accuracy. Other aspects include documentation standards and reproducibility, and the data release policy.

The initial organizational and collaborative structures brought about by the GlobalSoilMap initiative no longer exist, but it has left a lasting impact as exemplified by national soil mapping efforts, together with a large and active international research community which supports new, upcoming and seasoned scientists through training and education in the 'how' and how 'to' of doing DSM (Kidd et al., 2020). In 2016, the International Union of Soil Sciences created a Working Group (WP) named "Global Soil Map" to maintain this active international community. The Pillar 4 "Soil Information" of the UN–FAO Global Soil Partnership actively promotes country-based DSM activities and now includes the above mentioned WG in its scientific committee (Arrouays et al., 2020a).

Uses and users of digital soil mapping

As outlined earlier, the GlobalSoilMap effort aimed to provide high-resolution soil attribute maps that can ultimately be used in a variety of applications. These maps can provide soil inputs (e.g. texture, organic carbon and soil-depth parameters) to models predicting land-cover changes in response to global climatic and human disturbances. Here, we will discuss a few examples.

Maps of more readily available soil properties can be used to estimate other more costly, or difficult to measure soil attributes. For example, Ballabio et al. (2016) used soil texture maps (%clay, %silt and %sand) created using topsoil data from the European Land Use and Coverage Area frame Survey (LUCAS) to derive maps of bulk density, USDA soil texture classes and available water capacity. These soil attributes may then be used to assess the C sequestration potential of topsoils (bulk density), assess the soil water holding potential of agricultural soils for informed soil management, or estimate soil compaction hazards (texture classes).

Similar research has been conducted in France to estimate potential C sequestration and C storage (Chen et al., 2018) and available water capacity (Román Dobarco et al., 2019). More recently, French GlobalSoilMap products were used in combination with modelling to assess the feasibility of reaching the 4 per Mille target of increasing C stocks in soils (Martin et al., 2021).

High resolution soil attribute maps can also be used for monitoring and forecasting biophysical properties. More explicitly, these soil attribute maps may be used as input soil information to drive physical systems simulation models (e.g. crop simulation models such as APSIM) or for land suitability assessment, or more broadly defined as digital soil assessment.

An example of the use of fine-resolution 3-D soil-attribute maps is land suitability assessment that contributed to the GlobalSoilMap effort in Australia (Kidd et al., 2015), and was derived for Tasmania, Australia. Here, digital soil assessment was used to provide information on the agricultural suitability of land for 20 crops (perennial horticultural, cereal and vegetable crops) in a new irrigation scheme 'Water for Profit Program' commissioned by the Tasmanian state government. Together with 3-D soil-attribute maps (80 m × 80 m), climate grids (e.g. growing degree-days and frost risk) were applied to a range of defined enterprise-suitability rulesets to produce enterprise suitability maps. All maps created are stored on a publicly available spatial internet portal (Land Information Services Tasmania, LISTmap https://www.thelist.tas.gov.au/app/content/data#) and can be used interactively to perform a range of land management based assessments, such as identifying limiting soil and climate conditions. Outputs also include an enterprise versatility index; and the highest-valued agricultural land with the highest earning potential for individual commodities can also be identified.

Soil security to maintain soil function in supporting food and fiber production together with ecosystem balance is widely reported to be under threat due to human impacts and population increase. Consequently, there is growing activity throughout the world to measure and monitor soil resources (Arrouays et al., 2021), or more specifically the functions and services that soils provide directly and indirectly (Adhikari and Hartemink, 2016 in Aitkenhead and Coull, 2019). Digital soil mapping and

assessment can provide comprehensive and objective assessments of attributes including soil carbon stocks, and assessment of the potential of soils to store more carbon, through to explicit assessment of soil ecosystem services (Aitkenhead and Coull, 2019). It is expected that practitioners of DSM will soon respond in better and more creative ways to end-user requests for soil and soil-related information for these more general assessments of ecosystem function, rather than concentrated practice of creating estimated soil property maps as an end in itself.

Conclusions

Digital soil mapping is an effective tool for producing nuanced soil information. This information can be used in various ways, from understanding the status and condition of soils in a highly granular way, to aid in the planning and deployment of strategic land management practices and policies.

Scientifically, the incorporation of mechanistic soil processes into digital soil modelling is an innovation which should be given greater focus. Recent work with soil carbon, water and texture have been important contributions, together with structural equation modelling which provide explicit characterization of soil processing into the model framework. Digital soil mapping should not just become a machine learning exercise, and there are clear examples and experiences of the pitfalls of relying solely on machine learning algorithms to do the process in a thoughtless manner, and without rudimentary quality control mechanisms (for example, does the spatial pattern of the predictions make sense in a soil–landscape pedological context) of the outputs that are generated (Arrouays et al., 2020b).

There will be a widening and deepening role of public–private partnerships in the development and application of digital soil mapping to address serious environmental issues. For this, the process of digital soil mapping should not be considered a static task, but rather viewed as a dynamic process where efforts are continually revisited, updated, and improved where possible. Efficient strategies need to be developed to ensure the best use of available funding resources for the most gain.

Ultimately, soils are our most valuable natural resource. Digital soil mapping and assessment provide the tools and objective abilities to characterize and monitor the status of this resource better. Digital soil mapping is playing its part and will continue to be engaged in efforts throughout the world to ensure that we live on this planet in a more sustainable way.

References

Adhikari K and Hartemink AE (2016) Linking soils to ecosystem services — A global review. *Geoderma* 262: 101–111.

Aitkenhead MJ and Coull MC (2019) Digital mapping of soil ecosystem services in Scotland using neural networks and relationship modelling. Part 2: Mapping of soil ecosystem services. *Soil Use and Management* 35: 217–231.

Angelini ME, Heuvelink GBM, Kempen B, and Morrás HJM (2016) Mapping the soils of an Argentine Pampas region using structural equation modelling. *Geoderma* 281: 102–118.

Arrouays D, McBratney A, Minasny B, Hempel J, Heuvelink G, Macmillan R, Hartemink A, Lagacherie P, and McKenzie N (2014) The GlobalSoilMap project specifications. In: Arrouays D, McKenzie NJ, Hempel J, De Forges AR, and McBratney AB (eds.) *GlobalSoilMap: Basis of the Global Spatial Soil Information System*. London: CRC Press.

Arrouays D, Poggio L, Salazar Guerrero OA, and Mulder VL (2020a) Digital soil mapping and GlobalSoilMap. Main advances and ways forward. *Geoderma Regional* 21: e00265.

Arrouays D, McBratney A, Bouma J, Libohova Z, Richer-De-Forges AC, Morgan CLS, Roudier P, Poggio L, and Mulder VL (2020b) Impressions of digital soil maps: The good, the not so good, and making them ever better. *Geoderma Regional* 20: e00255.

Arrouays D, Mulder VL, and Richer-De-Forges AC (2021) Soil mapping, digital soil mapping and soil monitoring over large areas and the dimensions of soil security—A review. *Soil Security* 5: 100018.

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

Bishop TFA, McBratney AB, and Laslett GM (1999) Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma* 91: 27–45.

Brus DJ, Kempen B, and Heuvelink GBM (2011) Sampling for validation of digital soil maps. *European Journal of Soil Science* 62: 394–407.

Bui EN (2004) Soil survey as a knowledge system. *Geoderma* 120: 17–26.

Chen S, Martin MP, Saby NPA, Walter C, Angers DA, and Arrouays D (2018) Fine resolution map of top- and subsoil carbon sequestration potential in France. *Science of the Total Environment* 630: 389–400.

Chen S, Arrouays D, Mulder VL, Poggio L, Minasny B, Roudier P, Libohova Z, Lagacherie P, Shi Z, Hannam J, Meersmans J, Richer-De-Forges AC, and Walter C (2022) Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. *Geoderma* 409: 115567.

Gray JM and Bishop TFA (2019) Mapping change in key soil properties due to climate change over south-eastern Australia. *Soil Research* 57: 467–481.

Hastie T, Tibshirani R, and Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Huang J, McBratney A, Minasny B, and Triantafilis J (2017) Monitoring and modelling soil water dynamics using electromagnetic conductivity imaging and the ensemble Kalman filter. *Geoderma* 285: 76–93.

Jenny H (1941) *Factors of Soil Formation, a System of Quantitative Pedology*. New York: McGraw-Hill.

Kempen B, Brus DJ, Heuvelink GBM, and Stoorvogel JJ (2009) Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma* 151: 311–326.

Kempen B, Brus DJ, and Stoorvogel JJ (2011) Three-dimensional mapping of soil organic matter content using soil type-specific depth functions. *Geoderma* 162: 107–123.

Kempen B, Brus DJ, and Heuvelink GBM (2012) Soil type mapping using the generalised linear geostatistical model: A case study in a Dutch cultivated peatland. *Geoderma* 189-190: 540–553.

Kidd D, Webb M, Malone B, Minasny B, and McBratney A (2015) Digital soil assessment of agricultural suitability, versatility and capital in Tasmania, Australia. *Geoderma Regional* 6: 7–21.

Kidd D, Searle R, Grundy M, McBratney A, Robinson N, O'brien L, Zund P, Arrouays D, Thomas M, Padarian J, Jones E, Bennett JM, Minasny B, Holmes K, Malone BP, Liddicoat C, Meier EA, Stockmann U, Wilson P, Wilford J, Payne J, Ringrose-Voase A, Slater B, Odgers N, Gray J, Van Gool D, Andrews K, Harms B, Stower L, and Triantafilis J (2020) Operationalising digital soil mapping—Lessons from Australia. *Geoderma Regional* 23: e00335.

Lark RM, Cullis BR, and Welham SJ (2006) On spatial prediction of soil properties in the presence of a spatial trend: The empirical best linear unbiased predictor (E-BLUP) with REML. *European Journal of Soil Science* 57: 787–799.

Lee J, Viscarra Rossel RA, Luo Z, and Wang YP (2020) Simulation of soil carbon dynamics in Australia under a framework that better connects spatially explicit data with Roth C. *Biogeosciences Discussions* 2020: 1–24.
Ma Y, Minasny B, Malone BP, and McBratney AB (2019) Pedology and digital soil mapping (DSM). *European Journal of Soil Science* 70: 216–235.
Mallavan BP, Minasny B, and McBratney AB (2010) Homosoil, a methodology for quantitative extrapolation of soil information across the globe. In: Boettinger JL, Howell DW, Moore AC, Hartemink AE, and Kienast-Brown S (eds.) *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*. Dordrecht: Springer Netherlands.
Malone BP, McBratney AB, Minasny B, and Laslett GM (2009) Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma* 154: 138–152.
Malone BP, McBratney AB, and Minasny B (2011) Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma* 160: 614–626.
Malone B, Minasny B, Odgers N, and McBratney A (2014) Using model averaging to combine soil property rasters from legacy soil maps and from point data. *Geoderma* 232-234: 34–44.
Martin MP, Dimassi B, Román Dobarco M, Guenet B, Arrouays D, Angers DA, Blache F, Huard F, Soussana J-F, and Pellerin S (2021) Feasibility of the 4 per 1000 aspirational target for soil carbon: A case study for France. *Global Change Biology* 27: 2458–2477.
McBratney AB, Mendonca-Santos ML, and Minasny B (2003) On digital soil mapping. *Geoderma* 117: 3–52.
Miller DA and White RA (1998) A conterminous United States multilayer soil characteristics dataset for regional climate and hydrological modeling. *Earth Interactions* 2: 1–26.
Minasny B and McBratney AB (2010) Methodologies for global soil mapping. In: Boettinger JL, Howell DW, More AC, Hartemink AE, and Kienast-Brown S (eds.) *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*. London: Springer.
Minasny B and McBratney AB (2016) Digital soil mapping: A brief history and some lessons. *Geoderma* 264(Part B): 301–311.
Minasny B, McBratney AB, Mendonça-Santos ML, Odeh IOA, and Guyon B (2006) Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. *Australian Journal of Soil Research* 44: 233–244.
Minasny B, Vrugt JA, and McBratney AB (2011) Confronting uncertainty in model-based geostatistics using Markov Chain Monte Carlo simulation. *Geoderma* 163: 150–162.
Myers DB, Kitchen NR, Sudduth KA, Miles RJ, Sadler EJ, and Grunwald S (2011) Peak functions for modeling high resolution soil profile data. *Geoderma* 166: 74–83.
Odgers NP, Sun W, McBratney AB, Minasny B, and Clifford D (2014) Disaggregating and harmonising soil map units through resampled classification trees. *Geoderma* 214–215: 91–100.
Odgers NP, McBratney AB, and Minasny B (2015) Digital soil property mapping and uncertainty estimation using soil class probability rasters. *Geoderma* 237–238: 190–198.
Orton TG, Pringle MJ, and Bishop TFA (2016) A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals. *Geoderma* 262: 174–186.
Padarian J, Minasny B, and McBratney AB (2019) Using deep learning for digital soil mapping. *The Soil* 5: 79–89.
Poggio L and Gimona A (2014) National scale 3D modelling of soil organic carbon stocks with uncertainty propagation—An example from Scotland. *Geoderma* 232–234: 284–299.
Ponce-Hernandez R, Marriott FHC, and Beckett PHT (1986) An improved method for reconstructing a soil-profile from analysis of a small number of samples. *Journal of Soil Science* 37: 455–467.
Román Dobarco M, Bourennane H, Arrouays D, Saby NPA, Cousin I, and Martin MP (2019) Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. *Geoderma* 344: 14–30.
Scull P, Franklin J, Chadwick OA, and McArthur D (2003) Predictive soil mapping: A review. *Progress in Physical Geography* 27: 171–197.
Searle R, McBratney A, Grundy M, Kidd D, Malone B, Arrouays D, Stockman U, Zund P, Wilson P, Wilford J, Van Gool D, Triantafilis J, Thomas M, Stower L, Slater B, Robinson N, Ringrose-Voase A, Padarian J, Payne J, Orton T, Odgers N, O'brien L, Minasny B, Bennett JM, Liddicoat C, Jones E, Holmes K, Harms B, Gray J, Bui E, and Andrews K (2021) Digital soil mapping and assessment for Australia and beyond: A propitious future. *Geoderma Regional* 24: e00359.
Taalab K, Corstanje R, Zawadzka J, Mayr T, Whelan MJ, Hannam JA, and Creamer R (2015) On the application of Bayesian networks in digital soil mapping. *Geoderma* 259–260: 134–148.
Vaysse K and Lagacherie P (2017) Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma* 291: 55–64.
Vincent S, Lemercier B, Berthier L, and Walter C (2016) Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. *Geoderma* 311: 130–142.
Wadoux AMJC (2019) Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma* 351: 59–70.
Zhu AX, Band L, Vertessy R, and Dutton B (1997) Derivation of soil properties using a soil land inference model (SoLIM). *Soil Science Society of America Journal* 61: 523–533.