European Journal of Soil Science

European Journal of Soil Science, March 2019, **70**, 216–235 doi: 10.1111/ejss.12790

Rusell Review

Pedology and digital soil mapping (DSM)

Yuxin Ma, Budiman Minasny, Brendan P. Malone & Alex B. McBratney
Sydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, Eveleigh, New South Wales, Australia

Summary

Pedology focuses on understanding soil genesis in the field and includes soil classification and mapping. Digital soil mapping (DSM) has evolved from traditional soil classification and mapping to the creation and population of spatial soil information systems by using field and laboratory observations coupled with environmental covariates. Pedological knowledge of soil distribution and processes can be useful for digital soil mapping. Conversely, digital soil mapping can bring new insights to pedogenesis, detailed information on vertical and lateral soil variation, and can generate research questions that were not considered in traditional pedology. This review highlights the relevance and synergy of pedology in soil spatial prediction through the expansion of pedological knowledge. We also discuss how DSM can support further advances in pedology through improved representation of spatial soil information. Some major findings of this review are as follows: (a) soil classes can be mapped accurately using DSM, (b) the occurrence and thickness of soil horizons, whole soil profiles and soil parent material can be predicted successfully with DSM techniques, (c) DSM can provide valuable information on pedogenic processes (e.g. addition, removal, transformation and translocation), (d) pedological knowledge can be incorporated into DSM, but DSM can also lead to the discovery of knowledge, and (e) there is the potential to use process-based soil–landscape evolution modelling in DSM. Based on these findings, the combination of data-driven and knowledge-based methods promotes even greater interactions between pedology and DSM.

Highlights

- Demonstrates relevance and synergy of pedology in soil spatial prediction, and links pedology and DSM.
- Indicates the successful application of DSM in mapping soil classes, profiles, pedological features and processes.
- Shows how DSM can help in forming new hypotheses and gaining new insights about soil and soil processes.
- Combination of data-driven and knowledge-based methods recommended to promote greater interactions between DSM and pedology.

Introduction

Pedology is the study of soils as they occur in their environment. This includes soil formation, genesis, classification and cartography (Bockheim *et al.*, 2005). These topics make pedology relevant for tackling global issues such as soil, food, energy and water security, and climate regulation and human health (McBratney *et al.*, 2014). Pedology is an integrative and extrapolative science (Singer, 2005). Pedologists integrate an understanding of landscapes, vegetation patterns, climate and human activity into knowledge about soils and their distribution, and extrapolate their knowledge into soil maps. Jenny's (1941) *clorpt* model provides the first framework in pedology by quantifying or linking the relation of soil with state factors controlling soil formation. It is also the theoretical backbone for digital soil mapping (DSM), where the spatial soil-forming factors are used in quantitative spatial prediction (McBratney *et al.*, 2003).

Lagacherie & McBratney (2007) defined DSM as "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 observations and knowledge and from related environmental variables." The explicit geographic nature of DSM makes it relevant to pedology (Lin *et al.*, 2006). For example,

Correspondence: Yuxin Ma. Email: yuxin.ma@sydney.edu.au
Received 23 April 2018; revised version accepted 17 January 2019

observed soil characteristics (soil horizons and soil properties) serve as both evidence of past processes and an indicator of present processes, which are useful for understanding and predicting soil variation. Conversely, spatial information on soils can provide vital input to pedologic models (Bui, 2016; Thompson *et al.*, 2012). Therefore, the effective links between pedology and DSM can improve the connection between soil spatial distribution and process-based models.

The aim of this review was to (a) document the application of DSM in pedology, (b) highlight the relevance and synergy of pedology in soil spatial prediction through the expansion of pedological knowledge and (c) discuss how DSM can support further advances in pedology through improved representation of spatial soil information. The structure of this review is as follows: first we discuss pedology models and DSM concepts; this is followed by mapping soil classes, mapping soil profiles, and pedological features and processes. Finally, we discuss the relation between pedological knowledge and DSM and apply mechanistic pedological models.

Pedology models and DSM

The Clorpt model

The scientific rationale for making soil maps has been the *clorpt* model of Jenny, who considered soil as a dynamic system and formalized quantitatively the soil-forming factors, previously discussed by Hilgard (1860) and Dokuchaev (1883), into the state-factor equation:

$$s = f\ (cl, o, r, p, t, \ldots), \tag{1}$$

where $s =$ soil, $cl =$ climate, $o =$ organisms, $r =$ relief, $p =$ parent material and $t =$ time. The ellipsis (…) is reserved for additional unique factors that might be locally significant, such as atmospheric deposition (Thompson *et al.*, 2012). "The factors are not formers, or creators, or forces; they are variables (state factors) that define the state of a soil system" (Jenny, 1961). The state factors are independent from the soil system and vary in space and time (Amundson & Jenny, 1997).

Pedologists generally apply the mental, qualitative version of the *clorpt* principle for conventional soil mapping. Based on coupling the factors of soil formation with soil–landscape relations, soil survey is a scientific strategy (Hudson, 1992) and a 'knowledge system' (Bui, 2004). When pedologists or soil surveyors map the soil, they take into account the purpose of the map, the scale of publication, the soil-forming factors and their relative prevalence across the region studied and observed soil–landscape relations. However, pedologists or soil surveyors often cannot fully transfer their knowledge into a map. Therefore, the soil map produced by a soil surveyor can be considered as a structured representation of knowledge of the soil's spatial distribution in the landscape that reflects the pedologists' mental model. It also reflects the operational constraints of the soil survey (e.g. costs, accessibility, and so on). These mental models are generally described as narratives, are difficult to replicate and can be unsuitable for quantitative studies (Hartemink *et al.*, 2010). Moreover, these models lack spatial detail and assessment of the accuracy of soil attributes (Adhikari *et al.*, 2014).

Jenny's *clorpt* model can also be used quantitatively (i.e. a single factor is defined while the others are held constant). This functional factorial model has been important because it changed the way that soils were studied, leading to the development of empirical models to describe pedogenesis in the form of mathematical relations. Climo-, bio-, topo-, litho- and chrono-functions, for example, are useful for quantifying the effects of climate, organisms, topography, parent material and time, respectively, on soil formation (Brantley *et al.*, 2007; McBratney *et al.*, 2003; Yaalon, 1975).

Jenny *et al.* (1968) further proposed an "integrated clorpt" model, where all factors can be modelled simultaneously in the form of a multivariate linear regression:

$$S = a + k1\ \text{MAP} + k2\ \text{MAT} + k3\ \text{Parent material} + k4\ \text{Slope} + k5\ \text{Vegetation} + k6\ \text{Latitude}, \tag{2}$$

where S indicates soil properties (texture, C, N, cations and the first principal component of clay minerals), k are empirical coefficients, and MAP and MAT correspond to mean annual precipitation and temperature. This approach considered the combined influence of multiple variables and tried to explain the controlling factors of soil properties. The development of regression methods, coupled with computing power, has made the *clorpt* model a viable tool for quantitative soil science.

Scorpan model

McBratney *et al*. (2003) reformulated Jenny's state-factor model into the following equation for mapping soil:

$$S = f\ (s, c, o, r, p, a, n) + e, \tag{3}$$

where S represents soil attributes or classes that can be predicted from s, soil, other or previously measured attributes of the soil, c is climate, o are the organisms (including land cover and natural vegetation), r is the topography, p is the parent material, a is age or time, n is spatial location or position and e is spatially correlated residuals. Rooted in the *clorpt* model, the functional relations of the *scorpan* model between soil attributes or classes and environmental covariates can now be readily formulated using mathematical or statistical models. *Scorpan* factors are more than just environmental covariates because the model includes soil and spatial position. The *scorpan* equation is different from Jenny's factorial model or the Sergey–Zakharaov equation (Florinsky, 2012). Basically, the *clorpt* model was designed to produce knowledge, whereas *scorpan* was not intended for explaining soil formation, but rather a pragmatic empirical model for predicting soil properties and soil classes (production of maps). Nevertheless, we can still put pedological knowledge into the mapping process, or extract it from the map, as will be discussed in the 'Pedological knowledge and DSM' section.

The *scorpan* approach can also be defined as a partially dynamic method (McBratney *et al*., 2003) by taking the partial differentials of the *scorpan* factor over time (e.g. d*o*/d*t* or d*c*/d*t*), which means that we can project the existing soil map forward in time. Compared with a fully dynamic simulation model, this approach has limitations, such as lack of feedback and possible extrapolation problems. However, it is a quick and relatively simple way of estimating soil changes, and it has attracted considerable application in projecting the fate of soil carbon under future climate change scenarios, with examples from Gray and Bishop (2016) and Yigini & Panagos (2016).

STEP–AWBH model

To account explicitly for the importance of anthropogenic factors in soil formation, Grunwald *et al.* (2011) proposed another conceptual model for understanding soil properties for a pixel of size x (p_x) at a specific location on Earth, at a given soil depth (z) and at the current time(t_c):

$$\mathrm{SA}\left(z, p_x, t_c\right) = f\left\{\sum_{j}^{n}\left(S_j\left(z, p_x, t_c\right), T_j\left(p_x, t_c\right), E_j\left(p_x, t_c\right), P_j\left(p_x, t_c\right)\right)\right\},$$

$$\int_{i=0}^{m}\left\{\sum_{j}^{n}\left(A_j\left(p_x, t_i\right), W_j\left(p_x, t_i\right), B_j\left(p_x, t_i\right), H_j\left(p_x, t_i\right)\right)\right\}, \quad (4)$$

where SA is the soil property of interest, which is a function of a number ($j = 0,1, \dots n$) of relatively static environmental factors (only at t_c) and dynamic environmental conditions (with values representing dynamics through time t_i with $i = 0,1, \dots m$), S represents ancillary soil properties, T is topographic properties (e.g. elevation, slope gradient, slope curvature and compound topographic index), E represents ecological properties (e.g. physiographic region and ecoregion), P is the parent material and geologic properties (e.g. geologic formation), A represents atmospheric properties (e.g. precipitation, temperature and solar radiation), W is water properties (e.g. soil moisture and surface runoff), B represents biotic properties (e.g. vegetation or land cover, spectral indices derived from remote sensing and organisms) and H is human-induced forcings (e.g. land use and land-use change, contamination and disturbances).

In the *scorpan* model, topography (r) indirectly expresses the effects of hydrology on soil, whereas the *STEP–AWBH* model attempts to separate hydrologic properties (W) from topography (T) and climate (A) factors. Similar to the *scorpan* model, the *STEP–AWBH* model is spatially explicit by constraining soil properties to a specific pixel location (Thompson *et al*., 2012). The *STEP–AWBH* model is temporally explicit with the inclusion of time (Zhang *et al.*, 2016). Although the *STEP–AWBH* model seems to be more explicit, the real application is still minimal because of the difficulty of fulfilling and differentiating all the factors.

Digital mapping of soil classes

Mapping the spatial distribution of soil taxonomic classes is important for characterizing soil spatial variation (Lagacherie, 2005) and informing soil use and management decisions. Unsupervised, supervised and knowledge-based approaches have been used to produce maps of soil classes (MacMillan, 2008). As discussed in the previous section, the DSM approach starts with an input of soil class observations, selection of covariates that can represent the spatial distribution of soil classes, selecting prediction function or model, and assessing the accuracy of the model.

Top-down and bottom-up soil class

Here, we identify approaches to soil classification as top-down and bottom-up (Figure 1), also known as downward and upward approaches or classification from above and below (Odgers, 2010). Mapping soil taxonomic classes digitally is usually done by top-down classification (Minasny & McBratney, 2007).

Traditional top-down classification involves the division of the soil universe into mutually exclusive, non-overlapping subclasses based on a single soil property or several (Odgers *et al.*, 2011a). For example, Manil (1959) stated, “it starts from general facts and principles and goes down to more and more detailed categories as observations proceeds.” Most widely used national and international soil classification systems such as the US Soil Taxonomy and World Reference Base (WRB) are examples of top-down systems.

The top-down soil mapping approach (Figure 2) associates the estimation of soil properties (e.g. soil depth, texture, structure, moisture, temperature, cation exchange capacity, base saturation, clay mineralogy, organic matter content and salt content) with an existing soil classification system, allocates soil profiles into different hierarchical categories, such as order, suborder, great group, subgroup, family and series, and then make maps for such classes. One of the limitations of the top-down approach is that it is not formulated to characterize the continuous nature of soil variation by using mutually exclusive and non-overlapping soil classes. The top-down classification has generally been qualitative and too subjective when deciding which diagnostic soil properties to use, as well as where to place taxonomic divisions in the chosen properties (Odgers, 2010).

Bottom-up classification involves the aggregation of similar individuals into classes, and similar classes into higher-level classes, based on the assumption that we can identify the individuals and the overall similarity between individuals can be described quantitatively (Odgers, 2010). This approach is more objective and quantitative than the top-down method.

Odgers *et al*. (2011a) created a set of continuous soil layer classes based on soil properties measured in the laboratory and estimated from soil mid-infrared spectra using a fuzzy k-means clustering algorithm. These soil layer classes are ‘composite objects,’ that is, entities that are composed of sub-entities arranged in a specific sequence, for example soil profiles comprise sequences of soil layers. To create a soil profile, Odgers *et al.* (2011b) used the ‘Outil statistique d’aide à la cartogénèse automatique’ (OSACA)

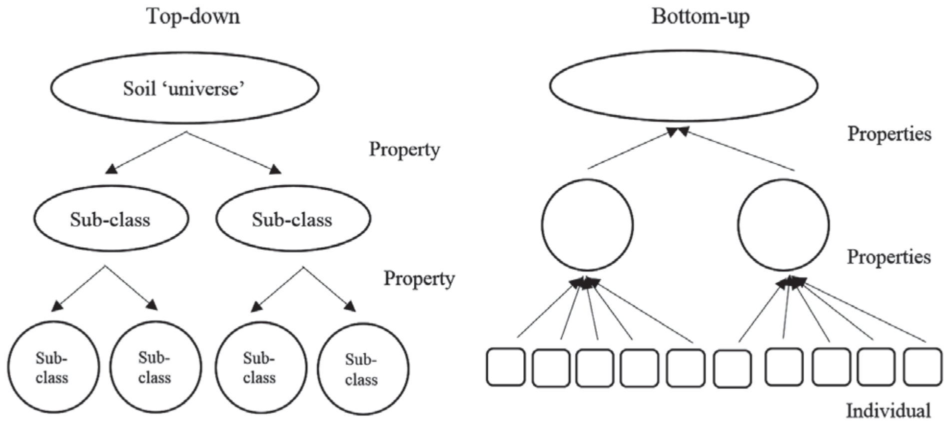

Figure 1 Schematic representation of the difference between top-down and bottom-up approaches to classification (from Odgers, 2010).

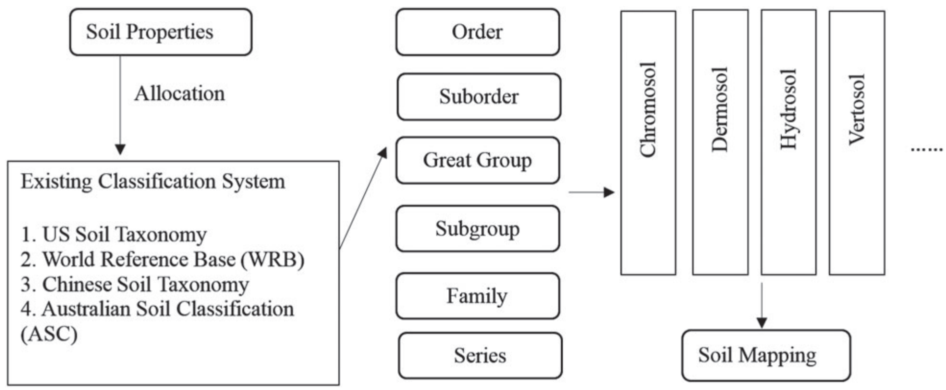

Figure 2 Top-down soil mapping.

algorithm of Carré & Jacobson (2009) to create continuous 'soil series classes' based on a set of continuous soil layer classes.

Bottom-up classification starts at the fundamental level, the soil layer, and builds upwards to soil series classes and mapping unit classes (Figure 3). This approach provides a greater degree of objectivity by using taxonomic distance and membership functions. It is a first step towards the development of a functional numerical soil classification system. One drawback of the approach is that it requires a large number of observations, soil analyses and descriptions, which can be time consuming and expensive.

Covariate selection: Expert versus machine

Various covariates representing soil state factors (e.g. climate, terrain factors and remotely sensed imagery) have been widely used in the predictive models (Heung *et al*., 2016; Marchetti *et al*., 2011; Pahlavan-Rad *et al*., 2014). The most easily quantified and directly correlated state factors are terrain with DEM derivatives as the main predictor variables (McBratney *et al*., 2003). However, not all soil state factors have representative covariates that are directly related to a particular factor; some of the covariates have indirect or multiple-factor relations. For example, direct estimates of time are absent from predictive models unless incorporated manually (Noller, 2010). However, indirect estimates of time can be inferred from relative landscape position (e.g. on the basis of principle of superposition), surface reflectance (e.g. red surfaces indicate a highly weathered soil), weathering indices based on gamma radiometry (Wilford, 2012), or parent material and geological maps (Thompson *et al*., 2012).

In addition to continuous variables, categorical maps such as geology, parent materials, geomorphology and legacy soil survey maps, which were derived from manual interpretation, can also be used as covariates (Pahlavan-Rad *et al*., 2014; Taghizadeh-Mehrjardi *et al.*, 2015). In particular, geomorphological maps (created manually) have been found to be a useful source of information for assessing soil parent material and soil genesis, and dominant in determining the spatial distribution of soil classes (Scull *et al.*, 2005). One issue of using such legacy information is that it can often be quite coarsely resolved and general. Such information has the potential to thwart rather than progress the application of DSM in some areas, meaning that the value of such data in a given mapping domain would need to be evaluated on a case by case basis.

Brungard *et al.* (2015) derived 113 covariates from DEM and Landsat imagery at several resolutions. The question that naturally arises with many covariates is which should be used as predictors in a DSM model. The selection of relevant covariates for mapping

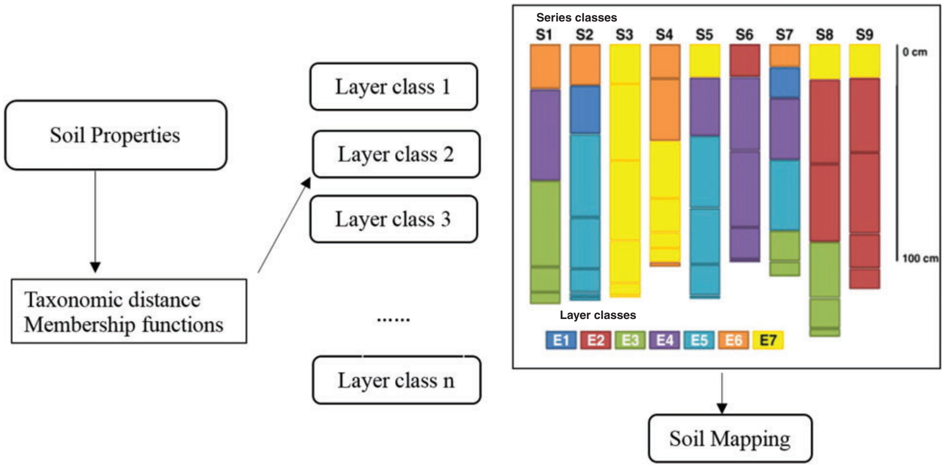

Figure 3 Bottom-up soil mapping (adapted from Odgers *et al*., 2011b). Layer classes are labelled as E1–E7 and soil series classes as S1–S9.

soil classes in an area can be based on the researchers' expertise in the domains of soil–environment processes. However, the decision can be biased or even fail in regions where knowledge on process is insufficient (Vasques *et al.*, 2012). Brungard *et al*. (2015) compared how covariates are selected based on the accuracy of the derived soil maps.

1. Covariates selected *a priori* by soil scientists familiar with each area, anticipating how soil–landscape relations would best be represented for modelling.
2. The covariates in set 1 plus 113 additional covariates derived from DEM and Landsat imagery at several resolutions that represented a large suite of potentially useful covariates.
3. Covariates selected using recursive feature elimination, a machine learning algorithm, from covariate sets 1 and 2.

The study found that for soil class prediction, covariates selected by recursive feature elimination consistently gave the most accurate predictions. Models using covariates selected by expert knowledge consistently performed the worst. This finding appears counterintuitive because expert judgement from pedologists would seem to be more reliable, given they would have a better understanding of the soil variation and potential reasons and controlling factors for this variation, relative to the tools of a data modeller. However, this was also not surprising because in clinical studies it has been shown that statistical prediction consistently outperforms expert judgement (Grove *et al.*, 2000). When faced with a large number of variables (covariates), a pedologist might not be able to identify optimal covariates *a priori* because of the complexity of soil-forming processes (Brungard *et al*., 2015).

In most DSM studies, covariate values are extracted at the location where soil records have been sampled. Soil properties in larger neighbourhoods might be different because of local geomorphic processes even if all other state factors are identical (Behrens *et al.*, 2010). It is necessary therefore to incorporate local and regional geomorphic context by incorporating information not only from the specific point, but also from its surrounding environment. Some studies tried to include a 'window size' to calculate terrain attributes and set covariates at various resolutions to resolve local and broader information together (Smith *et al.*, 2006). Behrens *et al*. (2010) presented the contextual elevation mapping (ConMap) approach using differences in contextual elevation from the centre pixel to each pixel across multiple neighbourhoods instead of common terrain attributes. Instead of trying different neighbourhood sizes for calculating terrain attributes, recent development of a deep learning approach for image classification, such as the convolution neural network (CNN) (Padarian *et al.*, 2018), could be useful for extracting information on covariates from images. The CNN model can extract features from images and then use these as input to the model. This method has started to be applied in DSM studies (Padarian *et al*., 2018).

Models and model selection

Following covariate selection, there is now a large array of mathematical, statistical and numerical models that can be used to analyse the direct or indirect relations between soil classes and environmental factors. According to Hengl *et al.* (2007), there are five types of models for soil class mapping.

1. Covariate classification techniques, which are mainly unsupervised classification of the *scorpan* variables.
2. Taxonomic distance techniques, which transform soil classes into continuous variables.
3. Multinominal logistic regression techniques.
4. Geostatistical techniques based on kriging.
5. Expert systems based on qualitative assessment and auxiliary data.

The above methods can be grouped into three categories: unsupervised (1), supervised (2–4) and knowledge-based approaches (5).

Selection of the above models can be based on the amount of input data. When there are few data only, unsupervised classification or experts' existing knowledge of the domains are the optimal options. When the soil classes can be defined exactly, expert knowledge can be used to create locally appropriate classification rules (MacMillan, 2008). When more input data are collected, a data-driven approach is complementary, such as regression techniques, neural networks or the use of pedological knowledge (e.g. taxonomic distance methods). Geostatistical methods could be used where there is a large amount of input data. Because of the difficulty in computing variograms for less frequent classes that occur at isolated locations, interpolating class memberships in geostatistics can be an alternative but can also be cumbersome because the universal kriging algorithm can become unstable or be excessive in computation time (Hengl *et al*., 2007).

Most DSM studies focus on comparing various numerical models for soil classes with observations from pedons and quantitative substitutes of the soil state factors. Brungard *et al*. (2015) compared 11 machine learning algorithms for predicting subgroup classes of the US Soil Taxonomy using pedon observations at three geographically distinct areas in the semiarid western USA. They divided the models into three groups based on model interpretability and the number of parameters required: simple, moderate and complex. They concluded that complex models (e.g. support vector machines and neural networks) were consistently more accurate than simple (e.g. multinominal logistic regression) or moderately complex models (e.g. classification tree). Heung *et al*. (2016) compared 10 machine learning techniques for mapping soil great groups and orders, and further confirmed that complex models such as the support vector machine produced more accurate results than simpler models (e.g. logistic model or classification trees). Rather than selecting the best model, there is an option to combine output from all models and thus combine their strengths. Ensemble models of continuous soil attributes have been applied in DSM (Malone *et al.*, 2014), but have not been applied to soil class mapping.

Classification accuracy and mapping infrequent soil classes

The overall accuracy of classification in each study area depends largely on the number of soil taxonomic classes and the frequency distribution of pedon observations between taxonomic classes (many classes with few observations = poor model performance). The number of soil classes appears related to the inherent variability of a given landscape (Brungard *et al*., 2015). Taghizadeh-Mehrjard *et al*. (2012) showed that in general classes with lower sampling frequencies were predicted less accurately because of the difficulty of separating soil classes in feature space with limited observations. This problem in machine learning is well known as 'class imbalance learning' (Liu *et al.*, 2006), where some classes are very underrepresented compared to others.

To address such imbalanced data, the following approaches can be applied. (a) Increase the number of observations in classes with few. A soil surveyor could manually identify potential locations of rare soil types with a combination of conditioned Latin hypercube sampling (cLHS) (Minasny & McBratney, 2006) and targeted sampling or case-based reasoning (Shi *et al.*, 2009). (b) Decrease the number of taxonomic classes by combining similar classes and modelling separate sub-areas. Combining similar subgroup classes could be achieved by using higher taxonomic levels such as great group or suborder (Jafari *et al.*, 2013) or those based on a particular soil property (e.g. bedrock contact) (Pahlavan-Rad *et al.*, 2014). Modelling separate sub-areas is theoretically appealing because different pedo-geomorphic sub-areas are likely to have different relations between subgroup classes and environmental covariates (McBratney *et al*., 2003). (c) Apply a weighting scheme to soil classes with few observations during model construction. However, Stum *et al.* (2010) found that this method sacrificed overall accuracy to improve the classification of minority classes. Although weighting may be intuitively desirable for imbalanced datasets, for very imbalanced datasets the method does not appear to be appropriate.

Taxonomic distance

All of the statistical and data-mining models treat the soil class as a 'label' and try to minimize the misclassification error. There should be a taxonomic relation between soil classes at any taxonomic level; so far, no spatial model has been used to account for these relations (Minasny & McBratney, 2007). The idea of calculating taxonomic distances to express the level of similarity and dissimilarity between different soil taxonomic units was first applied in the 1960s (Hole & Hironaka, 1960), but only with local data and was of limited scope. Taxonomic distance was resurrected in the 21st century by Minasny & McBratney (2007), who incorporated it between soil classes in a supervised classification routine (such as the decision tree) into spatial prediction and digital mapping of soil classes. Minasny *et al*. (2010) derived taxonomic distances for the WRB Reference Soil Groups (RSGs) based on the main key soil environmental properties. They concluded that the mean taxonomic distance showed a good relation between climate and soil classes, and appeared to be a useful index of pedodiversity, which combined the abundance and taxonomic relation between soil groups. Rossiter, Zeng, and Zhang (2017) showed how adjustments to measures of conventional classification accuracy can be made, taking into account taxonomic distance between classes, expressed as class similarities. These similarities can be weighted by experts, class hierarchy, numerical taxonomy or a loss function in accuracy assessments.

DSM versus conventional soil mapping

Some studies have compared the accuracy of soil class mapping by DSM with conventional methods of soil mapping in terms of accuracy (Lorenzetti *et al.*, 2015), cost, efficiency (Kempen *et al.*, 2012; Zeraatpisheh *et al.*, 2017), spatial correspondence (Bazaglia Filho *et al*., 2013) and spatial detail (Roecker *et al.*, 2010).

Lorenzetti *et al*. (2015) compared a traditional pedological approach and data mining techniques (neural networks, random forests, boosted tree, classification and regression tree, and support vector machine (SVM)) to assess the frequency of WRB reference soil groups (RSGs) in the 1:5 000 000 map of Italian soil regions. They showed that the SVM method was more accurate than the deterministic pedological approach. Both approaches were more successful in predicting absence rather than presence of a soil type.

Kempen *et al*. (2012) compared a universal kriging method and conventional soil maps created with the representative profile description and map unit means methods. Similarly, Zeraatpisheh *et al*. (2017) compared multinomial logistic regression and random forest methods with conventional soil survey for producing soil maps at four taxonomic levels. Both studies concluded that despite the differences in accuracy being small, DSM produced more informative and cost-efficient maps than conventional soil mapping.

Bazaglia Filho *et al*. (2013) compared four maps drawn independently by different soil scientists with the same set of information and a DSM obtained by fuzzy *k*-means clustering. They concluded that the average spatial correspondence between the conventional and DSM maps was similar. However, DSM can eliminate the subjectivity of soil surveyors by using quantifiable parameters.

Because of the reproducibility, easy-to-update workflow pattern, greater accuracy and cost-effectiveness, coupled with ability to quantify prediction uncertainties, DSM has become a useful and practicable approach to soil mapping.

Extracting soil information from soil maps, learning from the surveyors

Research has indicated that valuable knowledge about the relations between soil classes and underlying environmental conditions was embedded in the legacy soil map (Grunwald, 2009; Lagacherie *et al.*, 2001). For many environmental applications, such as soil erosion, soil carbon or biodiversity auditing, the level of detail and accuracy of existing maps is insufficient and funds are lacking to refine and update these maps by traditional soil survey. Kempen *et al*. (2012) showed that DSM can be an efficient alternative to conventional soil survey for updating a soil map in terms of accuracy and costs. Collard *et al*. (2014) showed that accuracy of the 1:250 000 reconnaissance soil map could be improved just by utilizing the relation between soil type and covariates calibrated on the existing legacy soil map. That is because soil mapping units already correspond to well-identified landscape units, which makes adding more precise and up-to-date covariates useful for producing a better map. This study suggested that when good quality maps are available, this method can be used in particular parts of the world where reconnaissance soil maps only are available.

The knowledge embedded within soil units delineated by experts can be partially retrieved as a covariate or a source of calibration data. However, the extent to which the models can be extrapolated to yield valid predictions can be contentious. An attempt to identify a relevant area digitally for extrapolation used taxonomic distance between the local soilscapes and those in a reference area (Lagacherie *et al*., 2001). Another modelling method involved the rule induction process of Bui & Moran (2001), where decision tree rules were created in training areas where detailed soil maps were available, and the rules were extrapolated to larger areas where detailed mapping was unavailable. However, both methods were used for extrapolation to areas within a given region and have not been tested on areas that are not geographically continuous (Minasny & McBratney, 2010). Grinand *et al.* (2008) tested extrapolation with the boosted classification tree and extrapolated regional soil landscapes from an existing soil map. They observed strong differences in accuracy between the training and extrapolated areas. They also found that sampling intensity did not greatly affect the accuracy of classification, and spatial context integration with a mean filtering algorithm increased the accuracy of prediction across the extrapolated area. However, the predictive capacity of models remained quite weak when extrapolated to an independent validation area. This study reiterates that knowledge from soil maps can be retrieved for extrapolation; however, the uncertainty needs to be accounted for.

Legacy soil maps typically encompass an assemblage of soil classes within a single soil polygon (Liu *et al*., 2016). The geographic locations of individual soil classes within the polygon are not specified. More recently, soil map disaggregation studies for determining the spatial configuration of individual soil classes include manually based approaches (Bui & Moran, 2001; Nauman *et al.*, 2012; Thompson *et al*., 2010) and model-based or data mining procedures (Häring *et al.*, 2012; Nauman & Thompson, 2014). Odgers *et al.* (2014) extracted individual soil series or soil class information from convolved soil map units by the DSMART (disaggregation and harmonization of soil map units through resampled classification trees) algorithm. The DSMART algorithm can best be explained as a data mining with repeated resampling algorithm. It can be quite a powerful algorithm for disaggregating legacy soil mapping, as demonstrated by Chaney *et al*. (2016), who disaggregated whole soil maps of the USA at a fine resolution (30 m).

In many places around the world, legacy soil information is difficult to obtain or can be non-existent. For these areas without detailed maps or soil observations, it is necessary to extrapolate from other parts of the world. Bui & Moran (2003) first used existing soil maps as 'reference' areas to represent a range of lithology, topography and climate, and developed rules for soil distribution and applied such rules over the corresponding physiographic domain for mapping the soils of the Murray-Darling Basin. Mallavan *et al.* (2010) further presented the method Homosoil, which assumes homology of soil-forming factors between a reference area and the region of interest. This includes climate, physiography and parent materials. They showed the similarity index for homoclime, homolith and homotop (areas or regions in the world with similar climate, lithology and topography) for Harare. A combination of these factors gives the homosoil similarity index for the area. This novel approach involves seeking the smallest taxonomic distance of *scorpan* factors between the region of interest and other reference areas (with soil data) in the world. Malone *et al.* (2016) provided a further overview of soil homologue together with a real-world application, which compared different extrapolating functions. Overall, development and application of the Homosoil approach or other analogues will become increasingly important for the operational advancement of

DSM in areas where soil information is non-existent or difficult to obtain.

Output assessment

There is always a deviation between prediction using DSM and observation in the real world, and pedologists must know the quality of predictions to judge their benefit for specific purposes. From this point, an important step in DSM workflow has been to quantify and summarize the performance of soil prediction models by calibration or validation, or both. Most researches have estimated the accuracy of soil class prediction models statistically using the confusion index or Kappa analysis (Brungard *et al*., 2015; Odgers *et al*., 2011a, 2011b), but have failed to interpret models in terms of expert knowledge. Balancing pedological interpretation and pure statistical accuracy would be a way that DSM and pedology could benefit each other.

Mapping soil profiles and pedological features

A soil profile is characterized by its horizons and pedological features. A soil map can be regarded as a planar projection of the spatial distribution of complexes of soil horizons. Then the map of the presence and absence of certain soil horizons is superimposed to carry out soil mapping (Sidorova & Krasilnikov, 2008). There are several examples of the successful application of DSM to predict the spatial distribution of soil horizons, such as albic, argillic, calcic, salic, spodic and fragipan horizons (Thompson *et al*., 2012). Soil horizons generally form within parent materials on stable surfaces (Schaetzl & Anderson, 2005) and soil parent materials have a considerable effect on pedological and geomorphological processes; therefore, there has been interest in mapping soil parent materials using the DSM method (Heung *et al.*, 2014; Lacoste *et al.*, 2011).

Mapping the thickness of soil horizons

Soil horizon thickness and its variation across a region is an essential factor for plant growth, environmental issues (soil erosion) and agricultural production, and plays a crucial role in farm management, land use or environmental protection (Gastaldi *et al.*, 2012). The depth and thickness of horizons can be influenced by geomorphologic positions and topographic attributes, such as elevation, slope, aspect, and hydrological and erosion processes (Odeh *et al.*, 1991).

Several studies have mapped the thickness or depth of individual soil horizons. Tsai *et al.* (2001) predicted the depth to A, B and BC horizons using linear soil–landscape regression models with limited information about the landscape properties, such as elevation, slope or surface stone content. Sidorova & Krasilnikov (2008) used an indicator kriging approach to study the spatial variation of the thickness of O, A, E and B horizons at three sites in southern and central Karelia, Russia. In peatlands, the interest is in mapping peat thickness to delineate areas for conservation and calculating carbon stocks (Rudiyanto *et al.*, 2018).

The topsoil thickness above the argillic horizon is an important factor in soil quality and productivity. In past research, topsoil thickness was estimated by fitting empirical regression equations to single-sensor apparent electrical conductivity data (Saey *et al.*, 2008). Sudduth *et al.* (2010) improved topsoil thickness estimates by combining data from multiple apparent electrical conductivity sensors, and by inverting a two-layer soil model incorporating instrument response functions to determine topsoil thickness.

Mapping regolith thickness to bedrock is important for environmental modelling in general and for seismic hazard assessment in particular. Shafique *et al.* (2011) developed a generic remote sensing and geophysics-based approach to model regolith thickness for areas with limited possibility of direct field observation. Wilford & Thomas (2013) described a method that maps the depth of regolith (with estimates of mapping uncertainty) in a complex geomorphic and weathering landscape in South Australia and to the moderately weathered or saprolite boundary for the whole of Australia (Wilford *et al.*, 2016), which is a considerable advance over mapping regolith depth by traditionally based regolith–landscape mapping methods.

However, few have mapped the ensemble horizons as a whole soil profile. Gastaldi *et al*. (2012) developed soil–landscape regression models using terrain attributes, land use and geology to describe and predict the occurrence and the thickness of several soil horizons to 1-m depth. They used a combination of logistic regression with linear regression to model the occurrence of each of the horizons first and then their thickness, respectively. They found that prediction quality for individual horizons was not large, which might be a result of short-scale variation and observation error. This model revealed good relations between the soil attributes and the prediction of the occurrence and thickness of each of the soil horizons. For example, the occurrence and thickness of A horizons are mostly governed by land use, whereas the B horizons are more related to soil–landscape processes.

Mapping pedological features

Diagnostic horizons are the combination of specific soil characteristics that are diagnostic of certain soil classes. Some studies have focused on mapping diagnostic horizons.

Jafari *et al.* (2012) compared binary logistic regression as an indirect approach and multinomial logistic regression as a direct approach to producing soil class maps in the Zarand region of southeast Iran. In the indirect prediction method, the occurrence of relevant diagnostic horizons was mapped first, and subsequently the indicator maps were combined on a pixel basis by the presence or absence of diagnostic horizons. With direct prediction, the probability distribution of the great soil groups was predicted directly because the dependent variable was the great group itself. The results showed that soils with better predictions were those strongly influenced by topographical and geomorphological characteristics, and *vice versa*. The indirect method gives insight into the causes of errors in prediction at the scale of diagnostic horizons, which helps in the selection of better covariates.

For fine scale mapping, proximal sensing of the soil's electrical conductivity appears to be an efficient way to map calcic horizons (Priori *et al.*, 2013). Priori *et al*. (2013) concluded that the depth to the calcic horizon in a vineyard showed a strong correlation with soil electrical conductivity by combining data from the geoelectrical sensor with a limited number of soil cores.

Hydromorphic features (Curmi *et al*., 1998) are the results of hydrologic processes within the soil and provide evidence of the magnitude and direction of water movement within the soil. The presence of hydromorphic soils influences soil–water–plant interactions and partly controls the hydrological response of catchments (Chaplot *et al.*, 2004) by restricting root growth, storing more (or less) plant available water and promoting lateral water flow (Thompson *et al*., 2012). Chaplot *et al*. (2004) demonstrated that the soil hydromorphic index can be mapped accurately using topographic indices.

Bell *et al.* (1992) used multivariate discriminant analysis with topography and geological information to predict drainage classes spatially. Kravchenko *et al.* (2002) and Campling *et al.* (2002) constructed soil spatial inference models of soil drainage class using topographical variables. These examples demonstrated that soil drainage classes can be related to topographic indices, vegetation indices and soil electrical conductivity.

Mapping soil parent material

Soil parent material is the initial state of the soil system and the material from which soils are derived (Jenny, 1941). It is responsible for soil development and soil type; physical and chemical properties of soils are influenced by parent material. Data for parent materials are generally derived from existing geological or lithological maps, and less often from soil maps, soil surveys or airborne gamma spectrometry (Lacoste *et al*., 2011). The lack of direct observations from the land surface and over a large spatial extent makes the performance of models that predict parent material quite poor. In terms of information content, maps of parent material probably contain the least spatially detailed information among all soil-forming factors (Zhang, Liu, & Song, 2017).

Decision trees have been used for predictive soil parent material mapping. Lacoste *et al*. (2011) used multiple additive and regression tree algorithms to generate a regional soil parent material map. They concluded that the resultant predictive map had greater accuracy than the existing geological map generated using conventional mapping techniques. Heung *et al*. (2014) applied the random forest model to predict the spatial distribution of soil parent material using topographic indices and conventional soil survey maps. They showed that maps produced by the random forest model conformed strongly overall with soil surveys and highlighted the need for reliable training data for the disaggregation of multi-component parent material polygons.

With the advance of remote sensing techniques, there is progress in that shortwave infrared surface reflectance or hyperspectral data can provide direct information on the mineralogy or bare surface materials (Zhang *et al*., 2017). Indirect relations can also be established from other factors such as climate and land surface temperature. For example, in arid and agricultural environments where the soil surface is typically exposed, information on parent material and the soil can be inferred.

Digital mapping of pedological processes

Simonson (1959) developed a process-systems model, describing the evolution of soil types as a function as follows:

$$s = f\,(\text{addition, removal, translocation, transformation})\,. \qquad (4)$$

1. Addition: new materials deposited by wind or water add to the soil, such as decomposing vegetation, organic matter and dust.
2. Removal: soil particles (sand, silt, clay and organic matter) or chemical compounds can be eroded, leached or harvested from the soil through the movement of wind or water.
3. Transformation: primary minerals into secondary minerals, for example, the formation of clay minerals or transformation of coarse organic matter into decay-resistant organic compounds (humus).
4. Translocation: movement of soil organic or mineral constituents within the profile or between horizons (Figure 4).

Many DSM studies have generated useful information about pedogenic processes, addition (e.g. aeolian dust deposition), removal (e.g. soil erosion), transformation (e.g. weathering of primary minerals, formation of clay minerals and calcification) and translocation (e.g. clay illuviation) by remote sensing, proximal sensing and machine leaning techniques. For example, Jang *et al.* (2016) used a portable X-ray fluorescence spectrometer (pXRF) to measure simultaneously the geochemical composition (Ca, Fe, Ti and Zr) of soils that could be useful for pedological studies. They identified Ca-rich parent materials (limestone) based on Ca concentration and areas of texture-contrast soils or soils with accumulation of clays in the B horizon based on Fe content. In addition, they calculated a soil weathering index using elemental concentrations (i.e. Ti and Zr) to explore the history of soil formation.

Gamma-ray spectrometry has proved to be a useful tool for soil mapping, given a sound understanding of the process to infer the detected signals (Stahr *et al*., 2013). It is a passive remote sensing technique that measures the concentration of three radioelements, potassium (K), thorium (Th) and uranium (U) at the Earth's surface. Soil-forming processes have a different effect on the dilution or accumulation of these radioelements, resulting in a differentiation in the gamma-ray signal. The following sections explain the application of gamma-ray spectrometry during the pedological processes.

Addition

The deposition and accumulation of aeolian dust is one type of pedogenetic addition and can be a critical component of soil

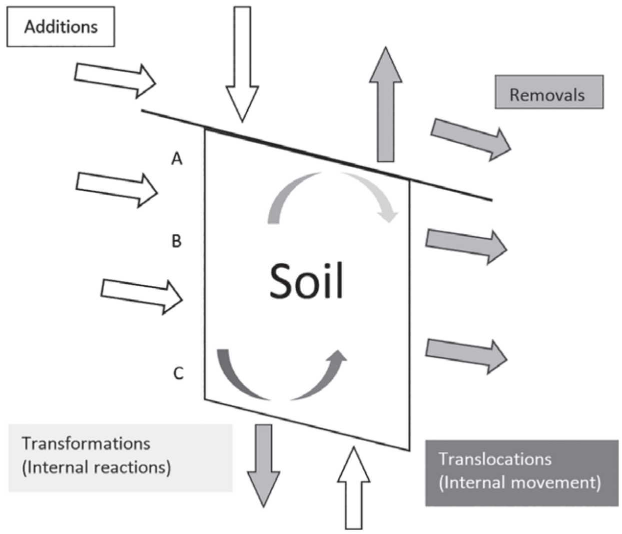

Figure 4 The removal, addition, transformation and translocation processes operating in soils (redrawn from Wilford & Minty, 2007).

development. A wide range of pedological, geochemical and geophysical analyses have been applied to identify and characterize aeolian dust inputs (Gatehouse *et al.*, 2001). However, most of these contributions often rely on specialized equipment and operator expertise, but do not indicate the spatial distribution of dust deposits across a landscape (Cattle *et al.*, 2003).

From different K, U and Th-bearing mineral suites, airborne gamma-ray spectrometry allows one to distinguish aeolian sediments from the underlying rock or a soil profile that has developed *in situ*. Cattle *et al*. (2003) identified aeolian dust in the topsoils of the Hillston district in western New South Wales using airborne with follow-up ground-based radiometric data and established a variable relation between large K signatures and apparent topsoil dust accumulations. However, in areas influenced by fluvial sediments, the radiometric signature was unable to indicate topsoil dust content.

Organic matter accumulation is another form of addition. Organic matter can absorb K; therefore, additional organic matter in the soil will dilute the overall gamma-ray signal. Whether this dilution is detectable depends on the degree of accumulation. According to Stahr *et al*. (2013), terrestrial arable soils with small organic matter concentrations (i.e. <2 mass%) cannot be detected, whereas thick organic layers on topsoil (O horizon) can be.

Removal

Soil erosion has been a major issue around the world. Gully erosion can have devastating effects on catchments, particularly those involved in agriculture (Valentin *et al.*, 2005). Kuhnert *et al.* (2010) used the random forest model to predict gully density, rate of gully erosion and its associated prediction error across a catchment using a suite of environmental variables as input to the model. However, their results cast doubt on the predictive ability of models of sediment transport that use gully erosion, where the error is estimated to be large. The consequence of this is that better modelling results will come only from improving the on-ground information on gully density. Rahmati *et al.* (2017) compared the performance of machine learning models to predict the occurrence of gully erosion in a watershed in Iran. They found that these models can be used in other gully erosion studies because they performed well both in the degree of fitting and in predictive performance. Hughes *et al*. (2001) found that land use, soil texture, temperature seasonality and relief are the most useful predictors for gullying across the continent, whereas lithology, mean annual rainfall, slope and hillslope length are also important regionally.

Transformation

Wilford & Minty (2007) illustrated that airborne gamma-ray spectrometry relates to the mineralogy and geochemistry of the bedrock and weathered materials (e.g. soils, saprolite, alluvial and colluvial sediments). Thorium and U concentrations often increase as K decreases during bedrock weathering and soil formation. They identified highly leached soils and deeply weathered bedrock and thin soils over fresh bedrock using a residual modelling technique that combined geological map units with the gamma-ray imagery based on the preferential loss of K-bearing minerals as the rock weathered. Wilford (2012) generated a weathering intensity map of the Australian continent based on the multivariate analysis of gamma-ray and terrain attributes. The map, therefore, has broad application in understanding weathering and geomorphic processes across a range of spatial and temporal landscape scales.

Calcium carbonate content is a key component of the regolith, particularly in arid and semiarid regions. It influences soil properties, is an important terrestrial carbon store and is used in mineral exploration. Wilford *et al.* (2015) developed a decision tree approach to map soil calcium carbonate abundance at the continental scale with inductive insight into the calcification process. They produced a prediction of soil calcium carbonate concentration in the upper regolith with associated degrees of model uncertainty and resolved the first-order controlling factors of soil calcium carbonate distribution over both depositional and erosional landscapes. This approach provides a better understanding of the environmental controls on soil carbonate formation and preservation in the landscape, which address inconsistencies and gaps in the existing national maps of soil carbonate and can be used to model and map geochemical and mineralogical properties in the upper regolith.

Translocation

The clay illuviation process represents the mechanical migration of clay from the surface horizons to the profile's deep horizons. Translocation of high-activity clay with a substantial concentration of K most probably leads to a decrease in the K signal in the topsoil (A and E horizons) and an increase in the enriched horizon (Bt). Stahr *et al*. (2013) illustrated different potassium surface and soil profile gamma-ray signals of soils with clay illuviation in Bor Krai village, northern Thailand, and concluded that the reference soil groups for high-activity clays (Alisol and Luvisol) showed a clear depth gradient. The incremental increase with depth was linear. In this case, gamma-ray spectrometry can be useful for mapping clay eluviation and illuviation, especially based on K signals.

Pedological knowledge and DSM

Walter *et al.* (2006) distinguished five broad domains where knowledge on soil properties and processes can be useful for spatial predictions or digital soil mapping.

1. Relative distribution of soil entities (profiles, horizons) within the landscape.
2. Identification of soil-forming factors.
3. Correlation between soil properties.
4. Spatial structures of soil properties.
5. Temporal dynamic of physical and biochemical processes.

Such pedological knowledge should be incorporated into predictive models to increase prediction efficiency and to link soil maps to dynamic modelling (Walter *et al*., 2006). Incorporating pedologic knowledge into predictive models is still a challenge because complex interrelations between soil-forming processes are not easy to capture and quantify. However, it is attractive because such incorporation could provide maps that represent the soil physical, chemical and biological processes better and bring benefits for both DSM and process modelling (Angelini *et al.*, 2016). Researchers have focused on the use of narrative data in DSM, the description and prediction of variation of soil properties down the profile and acquisition of new insights, hypotheses and further questions.

Narratives of pedogenesis

In conventional soil survey, the relations between soil properties and more readily observed environmental features have been expressed as narratives. Narratives of pedogenesis and soil distribution may be translated into a form that takes advantage of digital technology and remote sensing to reach the full potential of digital soil mapping, in particular when survey extents are large, data are sparse and resources are limited.

Local and regional models of pedogenesis are always qualitative and expressed through narratives, diagrams and sets of rules. McKenzie & Gallant (2006) used the system by Butler (1982) to develop a local model of pedogenesis, first as a narrative, which emphasizes landscape history, provenance of soil parent materials and pedogenic process, and then expressed as parsimonious and easy to update rules for digital soil mapping. The rules rely on just a few terrain and geophysical variables. This approach is explicit and repeatable and encourages improvements to predictions when and where they are needed. However, its prediction power is limited as discussed in the covariate selection section, where human knowledge can sometimes be biased and unable to comprehend the utility of a range of covariates.

Depth functions

Different natural and anthropogenic soil-forming processes result in different soils with different soil property depth profiles. Several attempts have been made to use pedometric methods to map the three-dimensional variation of soil properties using depth functions (Kempen *et al.*, 2011; Meersmans *et al.*, 2009).

Minasny *et al.* (2016) identified seven typologies of soil depth functions that describe soil property change with depth relating to soil-forming processes: uniform, gradational, wetting front, peak, minima–maxima (minmax), exponential and abrupt (Figure 5). Although the depth functions may or may not reflect soil horizon boundaries, they can infer soil processes, as in the examples below.

1. The exponential or power function is the most widely used soil depth function; it is mainly used to describe the decline of soil organic matter or carbon content with depth (largest in the A horizon and less in the subsequent horizons). The model assumes the distribution of organic matter added to soil by plant litter or decaying roots. The function can also be used to describe soil development, where the rate of soil weathering decreases with increasing soil thickness (Stockmann *et al.*, 2014).
2. The movement of water through a soil profile creates wetting front-type depth functions. This function has been found in weathering profile results from diffusion processes (Kirkby, 1985), depletion of leachable elements or minerals (Brantley *et al.*, 2008), depth distribution of soil horizons (Beaudette *et al.*, 2016) and gleyed horizons (Leblanc *et al.*, 2016).
3. Peak functions can show accumulation (maxima) of some soil properties such as clay content with depth because of eluviation and illuviation processes, or *in situ* formation or discontinuity in soil parent materials. Peak functions can also indicate compaction and anthropogenic influences can create variation such as multiple peaks (Minasny, 2012; Myers *et al.*, 2011)

Meersmans *et al*. (2009) considered anthropogenic influence (tillage) on soil formation and profile morphology for modelling the vertical distribution of soil organic carbon (SOC) using an exponential decay function with a constant SOC density to tillage depth. They concluded that SOC stock near the surface is determined by land use and soil type, whereas SOC near the bottom of the soil profile depends only on soil type (i.e. texture and drainage).

Kempen *et al*. (2011) mapped the three-dimensional distribution of soil properties by combining pedologically defined soil depth functions with geostatistical modelling. Five depth functions or horizon building blocks were defined, and for each soil type, the structure of the depth function was obtained by stacking a subset of model horizons. The parameters of the depth function for each of the horizons were interpolated by universal kriging with auto- and cross-covariance models. The depth functions of four peat soils (P, mP, PY and mPY) and two mineral soils (ES and PZ) are shown in Figure 6. This approach is useful in areas where soil formation results from distinct discrete anthropogenic or geologic disturbances.

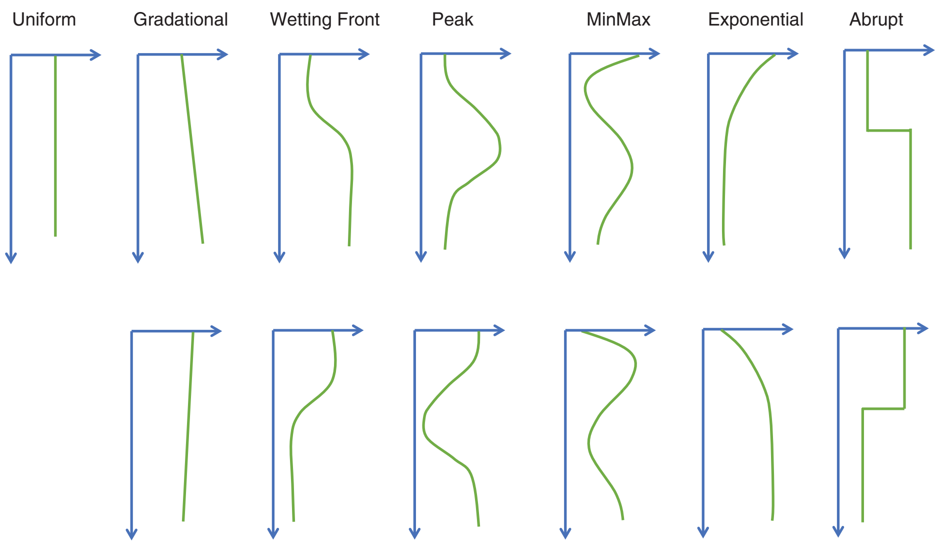

Figure 5 General typologies of depth functions (redrawn from Minasny *et al*., 2016).

Structural equation modelling

Angelini *et al*. (2016) introduced a statistical technique known as structural equation modelling (SEM) (Grace & Keeley, 2006) to integrate knowledge about interrelations between soil properties and predict these properties simultaneously. They identified the main soil forming processes in a 22 900 km^2 region in the Argentinian Pampas and assigned the main soil properties affected to each process. Based on this, they defined a conceptual model that summarizes the relations between soil-forming processes, controlling factors and affected soil properties (Figure 7), converted it to an SEM graphical model and then applied the SEM to predict seven key soil properties. Although the accuracy of the maps was poor based on cross-validation and independent validation, they illustrated that SEM can improve the consistency between multiple predicted soil properties and bridge the gap between empirical and mechanistic methods for soil–landscape modelling. Moreover, model error and measurement error can be distinguished explicitly by SEM.

Because of the many soil processes operating within the soil profile, Angelini *et al.* (2017) tested SEM for multilayer and multivariate soil mapping. They applied SEM to the model and predicted the lateral and vertical distribution of the cation exchange capacity (CEC), organic carbon (OC) and clay content of three major soil horizons, A, B and C in the same Argentinian Pampas area. They concluded that SEM is useful for predicting several soil properties simultaneously for multiple horizons.

Knowledge discovery from DSM

Knowing the correlation between field soil observations and environmental covariates does not necessarily indicate direct causation; however, it does help to stimulate ideas, gain new insights and hypotheses, and formulate further questions for research as shown in Figure 8. In other words, the occurrence of patterns from the soil and covariate relations could indicate knowledge discovery (Bui *et al.*, 2014). Knowledge discovery from DSM is discussed further in Henderson *et al.* (2005) and Scull *et al*. (2005).

Bui *et al.* (2006, 2009) found that SOC in Australian soils responded to vegetation, biomass, soil moisture and temperature patterns. The Australian continental modelling revealed a hierarchy of interacting variables. Climate was the most important factor at a continental scale, with different climatic variables dominating in different regions. The results support Dokuchaev's zonal soils theory, which emphasized climate. Lithology is important in defining broad-scale spatial patterns of soil properties, whereas topography controls shorter-range variation of soil properties (Bui *et al*., 2006). Based on these findings, Bui *et al*. (2014) investigated whether there were similar patterns in soil chemistry and plant and microbial communities. They demonstrated that soil salinity and pH are important filters for plant species richness.

The regolith thickness prediction of Wilford and Thomas (2013) in the 'Mapping the thickness of soil horizons' section determined that geology and weathering intensity are primary controls on regolith thickness and that terrain is a secondary, more local, control. Bui (2016) pointed out that these important results showed that an emphasis on catenary relations (i.e., using only terrain variables to account for soil patterns) would be misguided for large areas. Also, the results challenged the view that Australia's current climate has little relevance to soil properties at present because of the age of Australian landscapes (Taylor, 1983).

For the soil calcium carbonate distribution estimated by Wilford *et al*. (2015) and discussed in the 'Digitally mapping pedological processes' section, parent material, climate, specifically rainfall, temperature and seasonality are regional to national-scale controls, whereas topography, parent material and carbonate-rich groundwater discharge sites are local-scale controls. In another example in this section, Kuhnert *et al*. (2010)

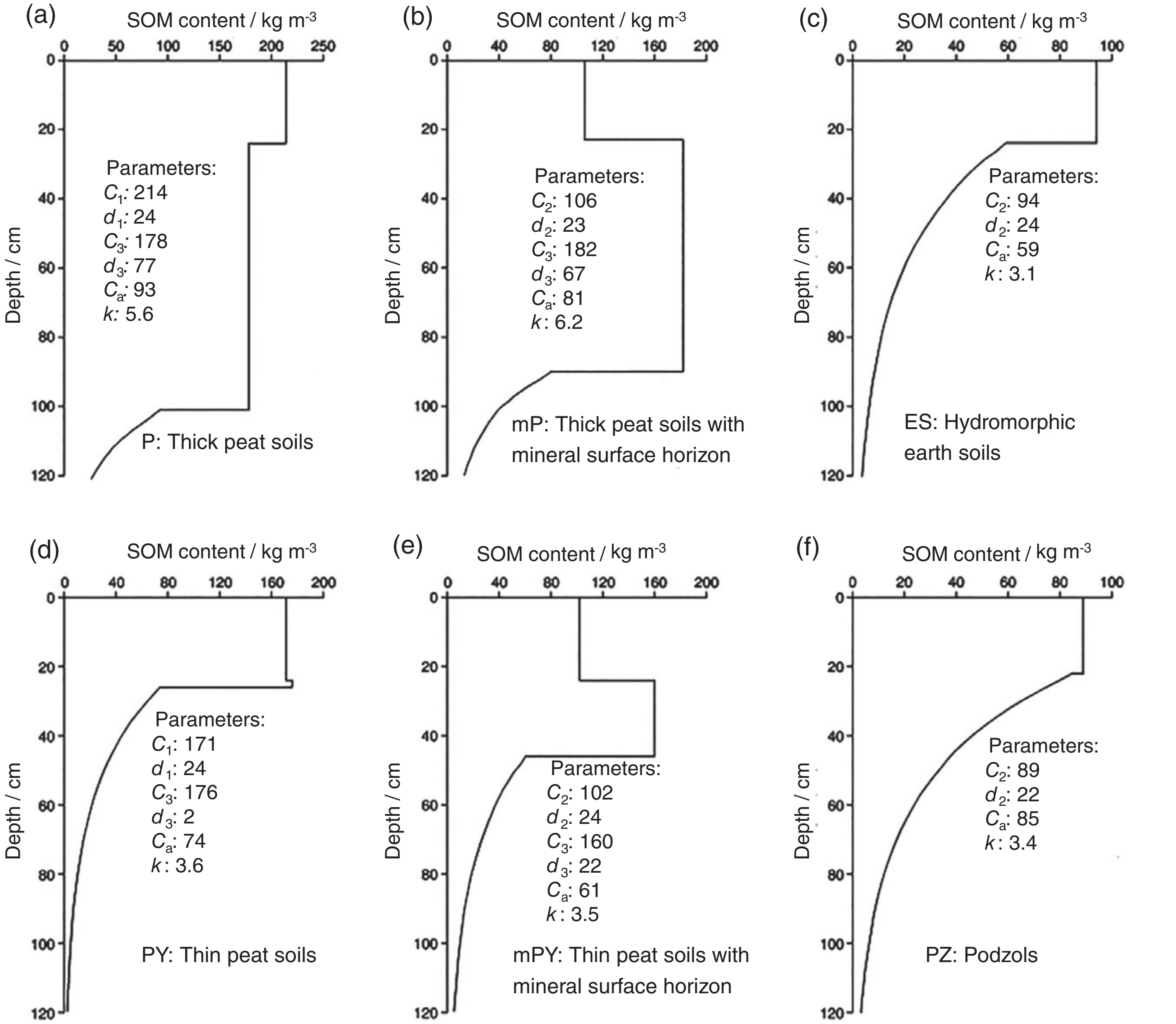

Figure 6 Predicted depth functions for six soil types at one location. The depth functions of the other four soil types that had zero probability are not shown here (adapted from Kempen *et al*., 2011). SOM, soil organic matter.

concluded that lithology, isothermality and annual mean moisture index are the most important predictors of gullying in a semiarid northern Australian river basin, whereas mean annual rainfall and slope are key factors in the Murray-Darling Basin (Hughes & Prosser, 2012).

Mechanistic pedological models in DSM

The DSM framework meets the increasing need for quantitative soil information, but is of limited use in situations of complex terrain where observations cannot be obtained. Moreover, the DSM framework expresses soil variation in spatial terms without the time dimension and excludes knowledge of the dynamic relations and mechanism between soil properties and landscape.

Researchers have found that there are strong links between the soils and geomorphology of the landform where they occur (Welivitiya *et al.*, 2016). Landscape evolution models try to model soil development as governed by weathering, erosion or deposition. The dynamics of changes in soil attributes can also have a positive effect on the long-term processes (landform evolution) or time-varying conditions (e.g. climate change) (Cohen *et al.*, 2010). For example, soil texture, soil organic carbon and surface stone cover can affect landscape evolution dynamics and the spatial variation and magnitude of erosion (Minasny *et al.*, 2015). In an eroding landscape, the bedrock–saprolite contact is closer to the surface, which in turn accelerates soil formation (Heimsath *et al*., 1997) from bioturbation, uprooting of bedrock material (Phillips & Marion, 2006), more intense chemical weathering and more active physical weathering

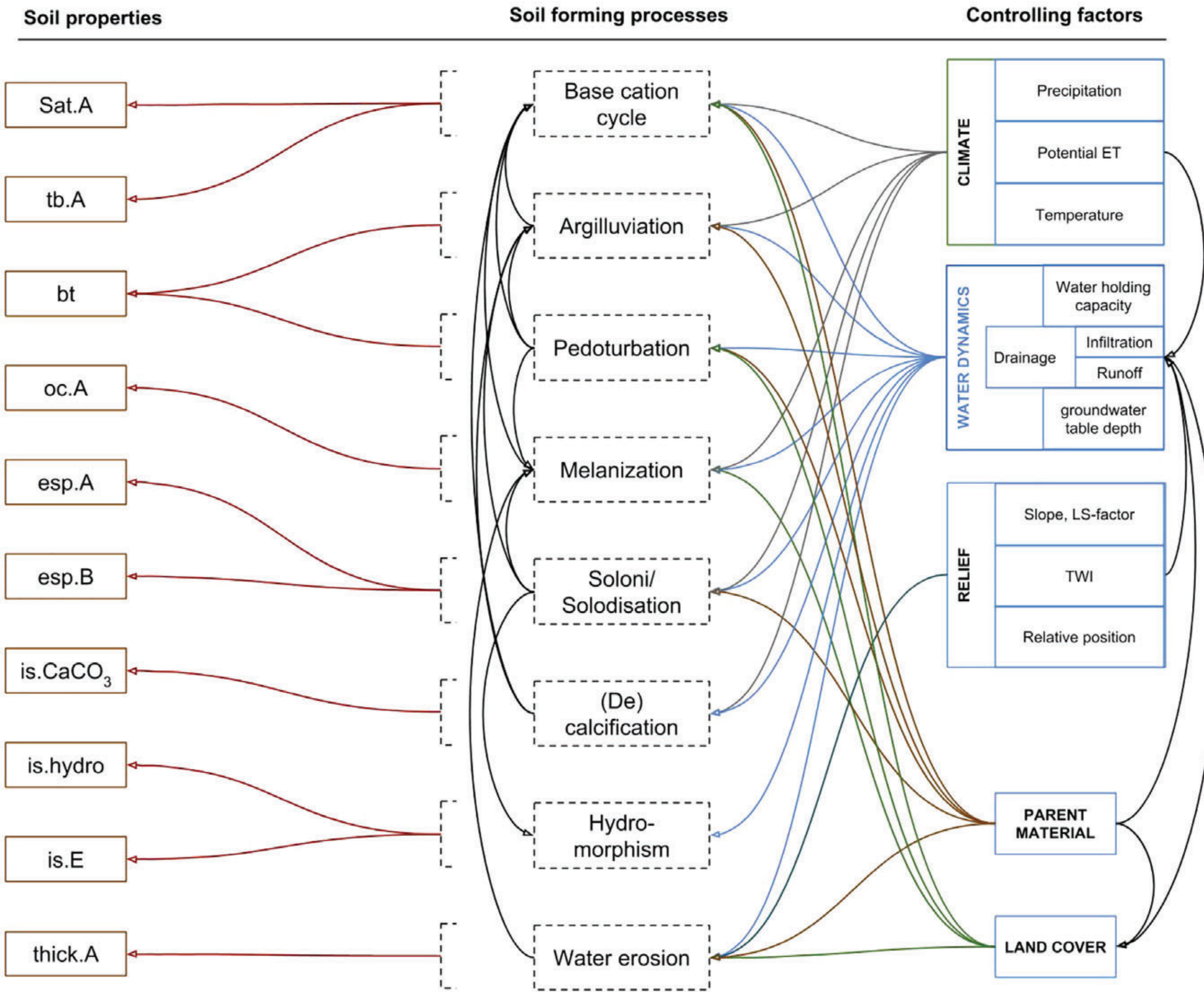

Figure 7 Conceptual model depicting relations between and across soil forming and controlling processes. Sat.A, base saturation of A horizon; tb.A, total bases of A horizon; bt, clay ratio B/A horizons; oc.A, soil organic carbon of A horizon; esp.A and esp.B, exchangeable sodium percentage of A and B horizons, respectively; is.$CaCO_3$, presence of calcium carbonate; is.hydro, presence of hydromorphic conditions; is.E, presence of E horizon; thick. A, thickness of A horizon (based on Angelini *et al*., 2016).

(Anderson *et al.*, 2013). Process-based soil–landscape evolution models have been developed to understand the spatiotemporal variation of soil properties on a dynamic landform (e.g., Lebedeva *et al.* (2010)).

Some studies of soil–landscape modelling work with a hypothetical landscape based on an assumption of a steady-state condition, and validation of soil–landscape models with limited field data (Braun *et al.*, 2016; Minasny *et al*., 2015). Willgoose & Sharmeen (2006) developed a physically based model named ARMOUR to simulate spatial and temporal changes of armouring (the process of surface coarsening) and weathering processes on a one-dimensional hillslope. By simplifying ARMOUR, Cohen *et al.* (2009) reformulated it as a state-space matrix model named mARM that can simulate complex hillslope soil surface particle-size evolution. Extending the mARM model, Cohen *et al*. (2010) developed mARM3D, which can model soil profile particle-size distribution at a large spatial extent. The SSSPAM (Welivitiya *et al*., 2016) generalized the formulation of mARM3D and extended the previous research to test more general conditions. Other models such as MILESD (Vanwalleghem *et al.*, 2013) and LORICA (Temme & Vanwalleghem, 2016) incorporated various chemical and biological processes in the simulation. The MILESD is formulated on the rudimentary framework of landscape-scale models for soil redistribution (Minasny & McBratney, 1999; Minasny & McBratney, 2001) and the pedon-scale soil formation model (Salvador-Blanes *et al.*, 2007). The LORICA modifies the three-layers module to represent the soil profile in MILEDS to incorporate additional layers and combines with the landform evolution model LAPSUS.

Ideally, we can use such soil–landscape models to simulate soil development in an area. However, little or no research has

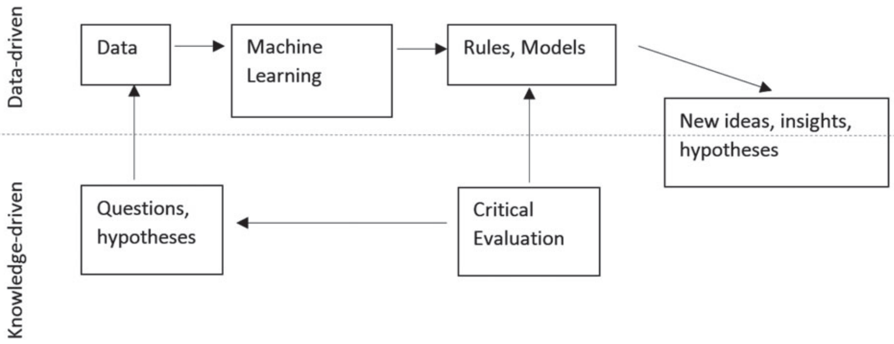

Figure 8 Generic workflow illustrating how the expert uses knowledge of the domains to define the study, control data selection and evaluate results (redrawn from Bui, 2016).

been carried out regarding use of such models. Bonfatti *et al.* (2018) estimated soil thickness and its variation over time using soil production function (SPF) and a landscape evolution model (LEM). The SPF calculated the rate of soil production and the LEM calculated erosion and deposition patterns. They observed that variation in soil thickness largely depended on the landscape position, and soil thickness was predicted more accurately in the upland areas, with a trend towards an equilibrium condition, than in the valley bottom areas without a trend towards equilibrium. These results indicate that the assumption of a steady-state condition might not hold over long timescales.

Runge (1973) suggested that soil formation is analogous to energy fluxes and developed 'energy models,' which are somewhat of a hybrid between the state-factor model of Jenny (1941) and process-systems model of Simonson (1959, 1978). The model emphasized two intensity factors, water available for leaching, organic matter production and time. Although Runge (1973) defined a qualitative model, Rasmussen *et al.* (2005) developed a quantitative pedogenic energy model that converts influxes of precipitation and organic matter (approximated by NPP) to soils on the basis of their contributions to an energy balance.

Linking models of pedoscale mechanics and soilscape would be advantageous to soil classification. According to Finke (2012), developing and applying three-dimensional simulation models of soil behaviour could help soil classification because the present classifications are based on morphological descriptions. However, such ideas are currently embryonic because no simulation model can presently simulate properties directly, such as soil colour, type or structure (Brevik *et al*., 2016).

Ma *et al*. (2019) used the soil–landscape model SSSPAM for predicting the spatial pattern and evolution of sand content for an area in the Hunter Valley, New South Wales, Australia. In this way, they attempted to improve the pedological knowledge in DSM techniques. They also tested the possibility of a process-based model used as a covariate in DSM, thus combining the empirical spatial data with pedological knowledge to predict soils in space and time.

The ultimate aim is to simulate soil distribution across the landscape with mechanistic models, rather than using empirical models, but there are still many issues in the use of such models, such as the following.

1. Soil is treated as a residual material, and soil formation is limited by weathering processes and soil redistribution by geomorphic processes.
2. A steady-state assumption is applied in many landscape evolution models. This should be questioned because of lack of equilibrium in the real landscape (Phillips, 2007).
3. On the equilibrium basis, the implicit initial condition and default process properties are used based on limited experiments and field observations, such as size-selective sediment transport and deposition.
4. Process coverage is limited; that is, only rare soil–landscape evolution models incorporate carbon dynamics, and most soil carbon models do not take the landscape component into account.
5. The simulation process is incomplete without the feedback effects of vegetation, biota or humans, or without coupling with other models such as climate or vegetation.
6. Computation of certain process representations is time consuming and the adaptation of the model code for use in high-performance computing is often complex.
7. Most calibration data represent the present-day situation and only a limited amount of data is available to represent geographical and temporal domains.

These challenges of model incompleteness, computational efficiency, and model calibration and validation issues should be a focus of future research. For example, the research required includes the following.

- Optimize the coupled four-dimensional soil–landscape evolution model to be available in a global context by converting soil formation into external forcings (boundary conditions).
- Balance increased process coverage and reduced computation complexity.
- Explicit representation of certain processes.
- Link with mineral equilibria models, energy and mass balance (thermodymics) models.
- Increase sensitivity analysis, and more substantive model calibration and validation.
- Combine with DSM. For example, combining the empirical spatial data with the mechanical process-based model to predict soils in space and time.

Conclusions

The creation of synergies between pedology and DSM has been developed by more accessible mathematical and statistical methods. A primary link between pedology and DSM is that pedogenesis and soil–landscape processes strongly influence spatial variation in soil properties. Here, we summarize some major findings for developments in DSM and pedology in this review.

1. Soil classes can be mapped accurately using DSM, especially when considering selection of covariates, complex models, taxonomic distance and depth functions of soil properties.
2. Information can be extracted from soil maps by updating or refining, extrapolating and disaggregating when there is only soil map information available. The Homosoil method can be applied where legacy soil information is difficult to obtain or even non-existent.
3. Occurrence and thickness of soil horizons, whole soil profile and soil parent material can be predicted successfully using DSM.
4. DSM can provide useful information about pedogenic processes, addition (aeolian dust deposition), removal (soil erosion), transformation (weathering of primary minerals, formation of clay minerals and calcification) and translocation (clay illuviation). Currently these processes have been mapped individually, but future work should focus on mapping all four processes simultaneously.
5. Pedological knowledge can be incorporated into DSM, but DSM can also lead to knowledge discovery.
6. There is a potential to use process-based soil–landscape evolution models in DSM.

Based on these important findings, it is clear that DSM is not solely about making maps. It can help in forming new hypotheses and gaining new insights into soil and soil processes. The combination of data-driven and knowledge-based methods promotes even greater interactions between pedology and DSM.

References

Adhikari, K., Minasny, B., Greve, M. B. & Greve, M. H. 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. *Geoderma*, **214–215**, 101–113.

Amundson, R. & Jenny, H. 1997. On a state factor model of ecosystems. *Bioscience*, **47**, 536–543.

Anderson, R. S., Anderson, S. P. & Tucker, G. E. 2013. Rock damage and regolith transport by frost: An example of climate modulation of the geomorphology of the critical zone. *Earth Surface Process. Landforms*, **38**, 299–316.

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

Angelini, M. E., Heuvelink, G. B. M., Kempen, B. & Morrás, H. J. M. 2016. Mapping the soils of an Argentine Pampas region using structural equation modelling. *Geoderma*, **281**, 102–118.

Bazaglia Filho, O., Rizzo, R., Lepsch, I. F., Prado, H.d., Gomes, F. H., Mazza, J. A., *et al*. 2013. Comparison between detailed digital and conventional soil maps of an area with complex geology. *Revista Brasileira de Ciência do Solo*, **37**, 1136–1148.

Beaudette, D. E., Roudier, P. & Skovlin, J. 2016. Probabilistic representation of genetic soil horizons. In: *Digital Soil Morphometrics* (eds. A. E. Hartemink & B. Minasny), pp. 281–293. Springer, Dordrecht, The Netherlands.

Behrens, T., Schmidt, K., Zhu, A.-X. & Scholtena, T. 2010. The ConMap approach for terrain-based digital soil mapping. *European Journal of Soil Science*, **61**, 133–143.

Bell, J. C., Cunningham, R. L. & Havens, M. W. 1992. Calibration and validation of a soil–landscape model for predicting soil drainage class. *Soil Science Society of America Journal*, **56**, 1860–1866.

Bockheim, J. G., Gennadiyev, A. N., Hammer, R. D. & Tandarich, J. P. 2005. Historical development of key concepts in pedology. *Geoderma*, **124**, 23–36.

Bonfatti, B. R., Hartemink, A. E., Vanwalleghem, T., Minasny, B. & Giasson, E. 2018. A mechanistic model to predict soil thickness in a valley area of Rio Grande do Sul, Brazil. *Geoderma*, **309**, 17–31.

Brantley, S. L., Bandstra, J., Moore, J. & White, A. F. 2008. Modelling chemical depletion profiles in regolith. *Geoderma*, **145**, 494–504.

Brantley, S. L., Martin, B. G. & Ragnarsdottir, K. V. 2007. Crossing disciplines and scales to understand the critical zone. *Elements*, **3**, 307–314.

Braun, J., Mercier, J., Guillocheau, F. & Cécile, R. 2016. A simple model for regolith formation by chemical weathering. *Journal of Geophysical Research: Earth Surface*, **121**, 2140–2171.

Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A. & Jordán A. 2016. Soil mapping, classification, and pedologic modeling: History and future directions. *Geoderma*, **264,** 256–274.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A. & Edwards, T. C., Jr. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, **239–240**, 68–83.

Bui, E. N. 2004. Soil survey as a knowledge system. *Geoderma*, **120**, 17–26.

Bui, E. N. 2016. Data-driven Critical Zone science: A new paradigm. *Science of the Total Environment*, **568**, 587–593.

Bui, E. N., Gonzalez-Orozco, C. E. & Miller, J. T. 2014. Acacia, climate, and geochemistry in Australia. *Plant and Soil*, **381**, 161–175.

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

Bui, E. N., Henderson, B. L. & Viergever, K. 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. *Global Biogeochemical Cycles*, **23**, GB4033.

Bui, E. N. & Moran, C. J. 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. *Geoderma*, **103**, 79–94.

Bui, E. N. & Moran, C. J. 2003. A strategy to fill gaps in soil survey over large spatial extents: An example from the Murray-Darling basin of Australia. *Geoderma*, **111**, 21–44.

Butler, B. E. 1982. A new system for soil studies. *Journal of Soil Science*, **33**, 581–595.

Campling, P., Gobin, A. & Feyen, J. 2002. Logistic modeling to spatially predict the probability of soil drainage classes. *Soil Science Society of America Journal*, **66**, 1390–1401.

Carré, F. & Jacobson, M. 2009. Numerical classification of soil profile data using distance metrics. *Geoderma*, **148**, 336–345.

Cattle, S. R., Meakin, S. N., Ruszkowski, P. & Cameron, R. G. 2003. Using radiometric data to identify aeolian dust additions to topsoil of the

Hillston district, western NSW. *Australian Journal of Soil Research*, **41**, 1439–1456.

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W. & Odgers N.P. 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. *Geoderma*, **274,** 54–67.

Chaplot, V., Walter, C., Curmi, P., Lagacherie, P. & King, D. 2004. Using the topography of the saprolite upper boundary to improve the spatial prediction of the soil hydromorphic index. *Geoderma*, **123**, 343–354.

Cohen, S., Willgoose, G. R. & Hancock, G. R. 2009. The mARM spatially distributed soil evolution model: A computationally efficient modeling framework and analysis of hillslope soil surface organization. *Journal of Geophysical Research: Earth Surface*, **114**, F03001.

Cohen, S., Willgoose, G. R. & Hancock, G. R. 2010. The mARM3D spatially distributed soil evolution model: Three-dimensional model framework and analysis of hillslope and landform responses. *Journal of Geophysical Research: Earth Surface*, **115**, F04013.

Collard, F., Kempen, B., Heuvelink, G. B. M., Saby, N. P. A., Forges, A. C. R., Lehmann, S., *et al*. 2014. Refining a reconnaissance soilmap by calibrating regression models with data from the same map (Normandy, France). *Geoderma Regional*, **1**, 21–30.

Curmi, P., Durand, P., Gascuel-Odoux, C., M´erot, P., Walter, C. & Taha, A. 1998. Hydromorphic soils, hydrology and water quality: Spatial distribution and functional modelling at different scales. *Nutrient Cycling in Agroecosystems*, **50**, 127–142.

Dokuchaev, V. V. 1883. Russian chernozem. In: *Selected Works of V.V. Dokuchaev, Volume 1. Moscow, 1948* (ed. S. Monson), pp. 14–419. (Transl. into English by N.Kaner). Israel Program for Scientific Translations Ltd. (for USDA-NSF), Jerusalem, Israel.

Finke, P. A. 2012. On digital soil assessment with models and pedometrics agenda. *Geoderma*, **171–172**, 3–15.

Florinsky, I. V. 2012. The Dokuchaev hypothesis as a basis for predictive digital soil mapping (on the 125th anniversary of its publication). *Eurasian Soil Science*, **45**, 445–451.

Gastaldi, G., Minasny, B. & McBratney, A. B. 2012. Mapping the occurrence and thickness of soil horizons. In: *Digital Soil Assessments and Beyond* (eds. B. Minasny, B. P. Malone & A. B. McBratney), pp. 145–148. CRC Press, Leiden, The Netherlands.

Gatehouse, R. D., Williams, I. S. & Pillans, B. J. 2001. Fingerprinting windblown dust in south-eastern Australian soils by uranium-lead dating of detrital zircon. *Australian Journal of Soil Research*, **39**, 7–12.

Grace, J. B. & Keeley, J. E. 2006. A structural equation model analysis of postfire plant diversity in California shrublands. *Ecological Applications*, **16**, 503–514.

Gray, J. M. & Bishop, T. F. A. 2016. Change in soil organic carbon stocks under 12 climate change projections over New South Wales, Australia. *Soil Science Society of America Journal*, **80**, 1296–1307.

Grinand, C., Arrouays, D., Laroche, B. & Martin, M. P. 2008. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma*, **143**, 180–190.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. & Nelson, C. 2000. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, **12**, 19–30.

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

Grunwald, S. J. A., Thompson, J. L. & Boettinger, J. L. 2011. Digital soil mapping and modeling at continental scales: Finding solutions for global issues. *Soil Science Society of America Journal*, **75**, 1201–1213.

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T. & Schröder, B. 2012. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. *Geoderma*, **185–186**, 37–47.

Hartemink, A. E., Hempel, J., Lagacherie, P., McBratney, A. B., McKenzie, N. J., MacMillan, R. A., *et al*. 2010. GlobalSoilMap.net–A new digital soil map of the world. In: *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation* (eds. J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink & S. Kienast-Brown), pp. 423–428. Springer Science, Dordrecht, The Netherlands.

Heimsath, A. M., Dietrich, W. E., Nishiizumi, K., Finkel, R. C., Mass, A. & National, L. L. 1997. The soil production function and landscape equilibrium. *Nature*, **388**, 358–361.

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

Hengl, T., Toomanian, N., Reuter, H. I. & Malakouti, M. J. 2007. Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. *Geoderma*, **140**, 417–427.

Heung, B., Bulmer, C. E. & Schmidt, M. G. 2014. Predictive soil parent material mapping at a regional-scale: A random Forest approach. *Geoderma*, **214–215**, 141–154.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E. & Schmidt, M. G. 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geroderma*, **265**, 62–77.

Hilgard, E. W. 1860. *Report on the Geology and Agriculture of the State of Mississippi*. E. Barksdale State Printer, Jackson, MS.

Hole, F. D. & Hironaka, M. 1960. An experiment in ordination of some soil profiles. *Soil Science Society of America Journal*, **24**, 309–312.

Hudson, B. D. 1992. The soil survey as paradigm-based science. *Soil Science Society of America Journal*, **56**, 836–841.

Hughes, A. O. & Prosser, I. P. 2012. Gully erosion prediction across a large region: Murray-Darling Basin. *Australian Journal of Soil Research*, **50**, 267–277.

Hughes, A. O., Prosser, I. P., Stevenson, J., Scott, A., Lu, H., Gallant, J., *et al*. 2001. Gully erosion mapping for the National Land and Water Resources Audit. In: *CSIRO Land and Water Technical Report 26/01*. ACT, Canberra, Australia.

Jafari, A., Ayoubi, S., Khademi, H., Finke, P. A. & Toomanian, N. 2013. Selection of a taxonomic level for soil mapping using diversity and map purity indices: A case study from an Iranian arid region. *Geomorphology*, **201**, 86–97.

Jafari, A., Finke, P. A., Van deWauw, J., Ayoubi, S. & Khademi, H. 2012. Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: Comparing logistic regression approaches to predict diagnostic horizons and soil types. *European Journal of Soil Science*, **63**, 284–298.

Jang, H. J., Minasny, B., Stockmann, U. & Malone, B. 2016. Spatial pedological mapping using a portable X-ray fluorescence spectrometer at the Tallavera grove vineyard, Hunter Valley. *Korean Journal of Soil Science and Fertilizer*, **49**, 635–643.

Jenny, H. 1941. *Factors of Soil Formation: A System of Quantitative Pedology*. McGraw-Hill, New York, NY.

Jenny, H. 1961. Derivation of state factor equations of soils and ecosystems. *Soil Science Society of America Proceedings*, **25**, 385–388.

Jenny, H., Salem, A. E. & Wallis, J. R. 1968. Interplay of soil organic matter and soil fertility with state factors and soil properties. “Study week on organic matter and soil fertility”. *Pontificiae Academiae Scientiarvm Scripta Varia*, **32**, 5–36.

Kempen, B., Brus, D. J. & Stoorvogel, J. J. 2011. Three-dimensional mapping of soil organic matter content using soil type-specific depth functions. *Geoderma*, **162**, 107–123.

Kempen, B., Brus, D. J., Stoorvogel, J. J., Heuvelink, G. B. M. & Vries, F.d. 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. *Soil Science Society of America Journal*, **76**, 2097–2115.

Kirkby, M. J. 1985. A basis for soil profile modelling in a geomorphic context. *European Journal of Soil Science*, **36**, 97–121.

Kravchenko, A. N., Bollero, G. A., Omonode, R. A. & Bullock, D. G. 2002. Quantitative mapping of soil drainage classes using topographical data and soil electrical conductivity. *Soil Science Society of America Journal*, **66**, 235–243.

Kuhnert, P. M., Henderson, A.-K., Bartley, R. & Herr, A. 2010. Incorporating uncertainty in gully erosion calculations using the random forests modelling approach. *Environmetrics*, **21**, 493–509.

Lacoste, M., Lemercier, B. & Walter, C. 2011. Regional mapping of soil parent material by machine learning based on point data. *Geomorphology*, **133**, 90–99.

Lagacherie, P. 2005. An algorithm for fuzzy pattern matching to allocate soil individuals to pre-existing soil classes. *Geoderma*, **128**, 274–288.

Lagacherie, P. & McBratney, A. B. 2007. Spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. In: *Digital Soil Mapping: An Introductory Perspective* (eds. P. Lagacherie, A. B. McBratney & M. Voltz), pp. 3–24. Elsevier, Amsterdam, The Netherlands.

Lagacherie, P., Robbez-Masson, J. M., Nguyen-The, N. & Barthès, J. P. 2001. Mapping of reference area representativity using a mathematical soilscape distance. *Geoderma*, **101**, 105–118.

Lebedeva, M. I., Fletcher, R. C. & Brantley, S. L. 2010. A mathematical model for steady-state regolith production at constant erosion rate. *Earth Surface Processes and Landforms*, **35**, 508–524.

Leblanc, M. A., Gagné, G. & Parent, L. E. 2016. Numerical clustering of soil series using profile morphological attributes for potato. In: *Digital Soil Morphometrics* (eds. A. E. Hartemink & B. Minasny), pp. 253–266. Springer, Dordrecht, The Netherlands.

Lin, H., Bouma, J., Pachepsky, Y., Western, A., Thompson, J. L., van Genuchten, R., *et al*. 2006. Hydropedology: Synergistic integration of pedology and hydrology. *Water Resources Research*, **42**, W05301.

Liu, F., Geng, X.-Y., Zhu, A.-X., Walter, F., Song, X.-D. & Zhang, G.-L. 2016. Soil polygon disaggregation through similarity-based prediction with legacy pedons. *Journal of Arid Land*, **8**, 760–772.

Liu, X.-Y., Wu, J.-X. & Zhou, Z.-H. 2006. Exploratory under sampling for class imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **39**, 539–550.

Lorenzetti, R., Barbetti, R., Fantappiè, M., L'Abate, G. & Costantini, E. A. 2015. Comparing data mining and deterministic pedology to assess the frequency of WRB reference soil groups in the legend of small scale maps. *Geoderma*, **237**, 237–245.

Ma, Y. X., Minasny, B., Welivitiya, W. D. D. P., Malone, B. P., Willgoose, G. R. & McBratney, A. B. 2019. The feasibility of predicting the spatial pattern of soil particle-size distribution using a pedogenesis model. *Geoderma*, **341**, 195–205.

MacMillan, R. A. 2008. Experiences with applied DSM: Protocol, availability, quality and capacity building. In: *Digital Soil Mapping With Lmited Data* (eds. A. E. Hartemink, A. B. McBratney & M. L. Mendonca-Santos), pp. 113–136. Springer, Dordrecht, The Netherlands.

Mallavan, B. P., Minasny, B. & McBratney, A. B. 2010. Homosoil, a methodology for quantitative extrapolation of soil information across the globe. In: *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation* (eds. J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink & S. Kienast-Brown), pp. 137–149. Springer, Dordrecht, The Netherlands.

Malone, B. P., Minasny, B., Odgers, N. P. & McBratney, A. B. 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. *Geoderma*, **232**, 34–44.

Malone, B. P., Sanjeev, K. J., Minasny, B. & McBratney, A. B. 2016. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. *Geoderma*, **262**, 243–253.

Manil, G. 1959. General considerations on the problem of soil classification. *Journal of Soil Science*, **10**, 5–13.

Marchetti, A., Piccini, C., Santucci, S., Chiuchiarelli, I. & Francaviglia, R. 2011. Simulation of soil types in Teramo province (Central Italy) with terrain parameters and remote sensing data. *Catena*, **85**, 267–273.

McBratney, A., Field, D. J. & Koch, A. 2014. The dimensions of soil security. *Geoderma*, **213**, 203–213.

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

McKenzie, N. J. & Gallant, J. C. 2006. Digital soil mapping with improved environmental predictors and models of pedogenesis. In: *Digital Soil Mapping: An Introductory Perspective* (eds. P. Lagacherie, A. B. McBratney & M. Voltz), pp. 327–349. Elsevier, Amsterdam, The Netherlands.

Meersmans, J., van Wesemael, B., De Ridder, F. & Van Molle, M. 2009. Modelling the three-dimensional spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). *Geoderma*, **152**, 43–52.

Minasny, B. 2012. Contrasting soil penetration resistance values acquired from dynamic and motor-operated penetrometers. *Geoderma*, **177–178**, 57–62.

Minasny, B., Finke, P., Stockmann, U., Vanwalleghem, T. & McBratney, A. B. 2015. Resolving the integral connection between pedogenesis and landscape evolution. *Earth-Science Reviews*, **150**, 102–120.

Minasny, B. & McBratney, A. B. 1999. A rudimentary mechanistic model for soil production and landscape development. *Geoderma*, **90**, 3–21.

Minasny, B. & McBratney, A. B. 2001. A rudimentary mechanistic model for soil production and landscape development. II. A two-dimensional model incorporating chemical weathering. *Geoderma*, **103**, 161–179.

Minasny, B. & McBratney, A. B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, **32**, 1378–1388.

Minasny, B. & McBratney, A. B. 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. *Geoderma*, **142**, 285–293.

Minasny, B. & McBratney, A. B. 2010. Methodologies for global soil mapping. In: *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation* (eds. J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink & S. Kienast-Brown), pp. 429–436. Springer, Dordrecht, The Netherlands.

Minasny, B., Stockmann, U., Hartemink, A. E. & McBratney, A. B. 2016. Measuring and modelling soil depth functions. In: *Digital Soil Morphometrics* (eds. A. E. Hartemink & B. Minasny), pp. 225–240. Springer, Dordrecht, The Netherlands.

Myers, D. B., Kitchen, N. R., Sudduth, K. A., Miles, R. J., Sadler, E. J. & Grunwald, S. 2011. Peak functions for modeling high resolution soil profile data. *Geoderma*, **166**, 74–83.

Nauman, T. W. & Thompson, J. A. 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. *Geoderma*, **213**, 385–399.

Nauman, T. W., Thompson, J. A., Odgers, N. P. & Libohova, Z. 2012. Fuzzy disaggregation of conventional soil maps using database knowledge extraction to produce soil property maps. In: *Digital Soil Assessments and Beyond* (eds. B. Minasny, B. P. Malone & A. B. McBratney), pp. 203–207. CRC Press, Leiden, The Netherlands.

Noller, J. S. 2010. Applying Geochronology in predictive digital mapping of soil. In: *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation* (eds. J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink & S. Kienast-Brown), pp. 43–53. Springer, Dordrecht, The Netherlands.

Odeh, I. O. A., Chittleborough, D. J. & McBratney, A. B. 1991. Elucidation of soil-landform interrelationships by canonical ordination analysis. *Geoderma*, **49**, 1–32.

Odgers, N. P. 2010. *Bottom-up Digital Soil Mapping*. (Doctoral dissertation). The University of Sydney, Sydney, Australia.

Odgers, N. P., McBratney, A. B. & Minasny, B. 2011a. Bottom-up digital soil mapping. I. Soil layer classes. *Geoderma*, **163**, 38–44.

Odgers, N. P., McBratney, A. B. & Minasny, B. 2011b. Bottom-up digital soil mapping. II. Soil series classes. *Geoderma*, **163**, 30–37.

Odgers, N. P., Sun, W., McBratney, A. B., Minasny, B. & Clifford, D. 2014. Disaggregation and harmonisation of soil map units through resampled classification trees. *Geoderma*, **214**, 91–100.

Padarian, J., Minasny, B. & McBratney, A. B. 2018. Using deep learning for Digital Soil Mapping. *SOIL Discuss*, **5**, 79–89.

Pahlavan-Rad, M. R., Toomanian, N., Khormali, F., Brungard, C. W., Komaki, C. B. & Bogaert, P. 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. *Geoderma*, **232–234**, 97–106.

Phillips, J. D. 2007. The perfect landscape. *Geomorphology*, **84**, 159–169.

Phillips, J. D. & Marion, D. A. 2006. Biomechanical effects of trees on soil and regolith: Beyond treethrow. *Annals of the Association of American Geographers*, **96**, 233–247.

Priori, S., Fantappiè, M., Magini, S. & Costantini, E. A. C. 2013. Using the ARP-03 for high-resolution mapping of calcic horizons. *International Agrophysics*, **27**, 313–321.

Rahmati, O., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H. R. & Feizizadeh, B. 2017. Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. *Geoderma*, **298**, 118–137.

Rasmussen, C., Southard, R. J. & Horwath, W. R. 2005. Modeling energy inputs to predict pedogenic environments using regional environmental databases. *Soil Science Society of America Journal*, **69**, 1266–1274.

Roecker, S., Howell, D., Haydu-Houdeshell, C. & Blinn, C. 2010. A qualitative comparison of conventional soil survey and Digital Soil Mapping approaches. In: *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation* (eds. J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink & S. Kienast-Brown), pp. 369–384. Springer, Dordrecht, The Netherlands.

Rossiter, D. G., Zeng, R. & Zhang, G.-L. 2017. Accounting for taxonomic distance in accuracy assessment of soil class predictions. *Geoderma*, **292**, 118–127.

Rudiyanto, Minasny, B., Setiawana, B. I., Saptomoa, S. K. & McBratney, A. B. 2018. Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. *Geoderma*, **313**, 25–40.

Runge, E. C. A. 1973. Soil development sequences and energy models. *Soil Science*, **115**, 183–193.

Saey, T., Simpson, D., Vitharana, U. W. A., Vermeersch, H., Vermang, J. & Van Meirvenne, M. 2008. Reconstructing the paleotopography between the loess cover with the aid of an electromagnetic induction sensor. *Catena*, **74**, 58–64.

Salvador-Blanes, S., Minasny, B. & McBratney, A. B. 2007. Modeling long-term in-situ soil profile evolution-Application to the genesis of soil profiles containing stone layers. *European Journal of Soil Science*, **58**, 1535–1548.

Schaetzl, R. & Anderson, S. 2005. *Soils: Genesis and Geomorphology*. Cambridge University Press, New York, NY.

Scull, P., Franklin, J. & Chadwick, O. A. 2005. The application of classification tree analysis to soil type prediction in a desert landscape. *Ecological Modelling*, **181**, 1–15.

Shafique, M., Meijde, M. & Rossiter, D. G. 2011. Geophysical and remote sensing-based approach to model regolith thickness in a data-sparse environment. *Catena*, **87**, 11–19.

Shi, X., Long, R., Dekett, R. & Philippe, J. 2009. Integrating different types of knowledge for digital soil mapping. *Soil Science Society of America Journal*, **73**, 1682.

Sidorova, V. & Krasilnikov, P. 2008. The use of geostatistical methods for mapping soil horizons. In: *Soil Geography and Geostatistics: Concepts and Applications* (eds. P. Krasilnikov, F. Carré & L. Montanarella), pp. 85–106. JRC Scientific and Technical Reports. European Commission, Joint Research Centre, Ispra, Italy.

Simonson, R. W. 1959. Modern concepts of soil genesis. Outline of a generalized theory of soil genesis. *Soil Science Society Proceedings*, **23**, 152–156.

Simonson, R. W. 1978. Multiple process model of soil genesis. In: *Quaternary Soils* (ed. W. C. Mahaney), pp. 1–25. GeoAbstracts, Norwich, England.

Singer, M. J. 2005. Pedology/basic principles. In: *Encyclopedia of Soils in the Environment* (ed. D. Hillel), pp. 151–156. Elsevier, Amsterdam, Netherlands.

Smith, M. P., Zhu, A.-X., Burt, J. E. & Stiles, C. 2006. The effects of DEM resolution and neighborhood size on digital soil survey. *Geoderma*, **137**, 58–69.

Stahr, K., Clemens, G., Schuler, U., Erbe, P., Haering, V., Cong, N. D., Bock, M., Tuan, V. D., Hagel, H., Vinh, B. L., Rangubpit, W., Surinkum, A., Willer, J., Ingwersen, J., Zarei, M. & Herrmann, L. 2013. Beyond the horizons: Challenges and prospects for soil science and soil care in Southeast Asia. In: *Sustainable Land Use and Rural Development in Southeast Asia: Innovations and Policies for Mountainous Areas*. (eds. H. L. Fröhlich, P. Schreinemachers, K. Stahr & G. Clemens) Springer, Berlin, Heidelberg.

Stockmann, U., Minasny, B. & McBratney, A. B. 2014. How fast does soil grow? *Geoderma*, **216**, 48–61.

Stum, A. K., Boettinger, J. L., White, M. A. & Ramsey, R. D. 2010. Random forests applied as a soil spatial predictive model in Arid Utah. In: *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation* (eds. J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink & S. Kienast-Brown), pp. 179–190. Springer, Dordrecht, The Netherlands.

Sudduth, K. A., Kitchen, N. R., Myers, D. B. & Drummond, S. T. 2010. Mapping depth to argillic soil horizons using apparent electrical conductivity. *Journal of Environmental and Engineering Geophysics*, **15**, 135–146.

Taghizadeh-Mehrjard, R., Minasny, B., McBratney, A. B., Triantafilis, J., Sarmadian, F. & Toomanian, N. 2012. Digital soil mapping of soil classes using decision trees in Central Iran. In: *Digital Soil Assessments and Beyond* (eds. B. Minasny, B. P. Malone & A. B. McBratney), pp. 197–202. CRC Press, Leiden, The Netherlands.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B. & Triantafilis, J. 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. *Geoderma*, **253–254**, 67–77.

Taylor, J. K. 1983. The Australian environment. In *Soils: An Australian Viewpoint* (ed. CSIRO. Division of Soils), pp. 1–2. CSIRO/Academic Presss, Melbourne, Australia/London, England.

Temme, A. J. A. M. & Vanwalleghem, T. 2016. LORICA - A new model for linking landscape and soil profile evolution: Development and sensitivity analysis. *Computers & Geosciences*, **90**, 131–143.

Thompson, J. A., Prescott, T., Moore, A. C., Bell, J., Kautz, D. R., Hempel, J. W., *et al*. 2010. Regional approach to soil property mapping using legacy data and spatial disaggregation techniques. In *19th World Congress of Soil Science* pp. 17–20. Brisbane, Queensland, Australia.

Thompson, J. A., Roecker, S., Grunwald, S. & Owens, P. R. 2012. Digital soil mapping: Interactions with and applications for hydropedology. *Hydropedology*, **1**, 664–709.

Tsai, C. C., Chen, Z. S., Duh, C. T. & Horng, F. W. 2001. Prediction of soil depth using a soil-landscape regression model: A case study on forest soils in southern Taiwan. *Proceedings of the National Science Council, Republic of China. Part B, Life Sciences*, **25**, 34–39.

Valentin, C., Poesen, J. & Li, Y. 2005. Gully erosion: A global issue. *Catena*, **63**, 129–131.

Vanwalleghem, T., Stockmann, U., Minasny, B. & McBratney, A. B. 2013. A quantitative model for integrating landscape evolution and soil formation. *Journal of Geophysical Research-Earth Surface*, **118**, 331–347.

Vasques, G. M., Grunwald, S. & Myers, D. B. 2012. Associations between soil carbon and ecological landscape variables at escalating spatial scales in Florida, USA. *Landscape Ecology*, **27**, 355–367.

Walter, C., Lagacherie, P. & Follain, S. 2006. Integrating pedological knowledge into digital soil mapping. In: *Digital Soil Mapping: An Introductory Perspective* (eds. P. Lagacherie, A. B. McBratney & M. Voltz), pp. 281–300. Elsevier, Amsterdam, The Netherlands.

Welivitiya, W. D. D. P., Willgoose, G. R., Hancock, G. R. & Cohen, S. 2016. Exploring the sensitivity on a soil area-slope-grading relationship to changes in process parameters using a pedogenesis model. *Earth Surface Dynamics*, **4**, 607–625.

Wilford, J. 2012. A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis. *Geoderma*, **183**, 124–142.

Wilford, J., Caritatab, P.d. & Bui, E. 2015. Modelling the abundance of soil calcium carbonate across Australia using geochemical survey data and environmental predictors. *Geoderma*, **259-260**, 81–92.

Wilford, J. & Minty, B. 2007. The use of airborne gamma-ray imagery for mapping soils and understanding landscape processes. In *Digital Soil Mapping: An Introductory Perspective* (eds. P. Lagacherie, A. B. McBratney & M. Voltz.), pp. 207–218. Elsevier, Amsterdam, The Netherlands.

Wilford, J. R., Searle, R., Thomas, M., Pagendam, D. & Grundy, M. J. 2016. A regolith depth map of the Australian continent. *Geoderma*, **266**, 1–13.

Wilford, J. R. & Thomas, M. 2013. Predicting regolith thickness in the complex weathering setting of the central Mt Lofty Ranges, South Australia. *Geoderma*, **206**, 1–13.

Willgoose, G. R. & Sharmeen, S. 2006. A one-dimensional model for simulating armouring and erosion on hillslopes: 1. Model development and event-scale dynamics. *Earth Surface Process and Landforms*, **31**, 970–991.

Yaalon, D. H. 1975. Conceptual models in pedogenesis: Can soil-forming functions be solved? *Geoderma*, **14**, 189–205.

Yigini, Y. & Panagos, P. 2016. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. *Science of the Total Environment*, **557–558**, 838–850.

Zeraatpisheh, M., Ayoubi, S., Jafari, A. & Finke, P. 2017. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. *Geomorphology*, **285**, 186–204.

Zhang, G.-L., Liu, F. & Song, X.-D. 2017. Recent progress and future prospect of digital soil mapping: A review. *Journal of Integrative Agriculture*, **16**, 2871–2885.

Zhang, G.-L., Liu, F., Song, X.-D. & Zhao, Y.-G. 2016. Digital soil mapping across Paradigms, scales, and boundaries: A review. In *Digital Soil Mapping Across Paradigms, Scales and Boundaries* (eds. G.-L. Zhang, D. Brus, F. Liu, X.-D. Song & P. Lagacherie), pp. 3–10. Springer, Singapore.