Soil mapping today: computer-generated predictive soil maps – their role in soil survey and land evaluation

David G Rossiter

David Rossiter is a native of Ithaca, New York, in the Finger Lakes region where the theory of continental glaciation was confirmed in the USA by Louis Agassiz. He earned a BSc in soil science under Marlin Cline and, after detours through computer science, a PhD based on the Automated Land Evaluation System working with Armand van Wambeke, an authority on tropical soils. He has worked as a soil surveyor in North Carolina, system manager of the Veterinary College Computing Facility, technical adviser to inter alia *the Venezuelan Ministry of Environment; finally settling in the Netherlands at the University of Twente, Faculty of Geoinformation and Earth Observation. After nominal retirement, he was appointed Associate Professor in Soil & Crop Sciences at Cornell, where he teaches a graduate course in geospatial analysis, and is guest researcher at ISRIC – World Soil Information on soil mapping, databases and land evaluation. His written output is online at http://www.css.cornell.edu/faculty/dgr2/*

david.rossiter@isric.org

Abstract

Predictive soil mapping, also called digital soil mapping, is the now-dominant means of producing soil maps to be used as digital layers in land resource assessment and land surface modelling. It has the advantage of consistency and has enabled the production of global and regional maps of soil properties at moderate (250 m) to fine (30 m) grid-cell resolutions. These also provide internal measures of their reliability, so can be used in sensitivity and scenario analyses. Reliability depends on the density of observations so, in areas where agricultural development projects are undertaken, their quality is doubtful but they do provide a good starting point for zoning and more detailed study.

Introduction

The increasing quantification and computerisation in society has resurrected soil surveying and the value of soil surveys. This has been made possible by three factors: almost magical computer power (recently Google Earth Engine); open databases of soil profile observations (especially ISRIC – World Soil Information's World Soil Information Service [WoSIS]); and free availability of a vast store of geographic gridded coverages related to the factors of soil formation (Jenny, 1941), especially *relief* (the terrain from digital elevation models), *organisms* (vegetative cover from satellite imagery) and *climate* (*eg* WorldClim). These allow the production of soil maps by so-called 'predictive soil mapping', more popularly known as 'digital soil mapping'.

Digital and predictive soil mapping

The first approaches to predictive soil mapping began in the 1990s and were well reviewed by Scull *et al* (2003) in *Progress in Physical Geography*. They chose the term *predictive* in the same sense that our German colleagues refer to *concept* maps, which in other contexts have been called 'pre-maps'. The idea is that from ancillary information and isolated observations, the surveyor makes a map of the study area and, then, uses it to plan fieldwork to confirm, modify and update the map into a product to be delivered to map users. In the same year, McBratney *et al* (2003) promoted the now-current term 'digital soil mapping' (DSM) to emphasise the indispensable role of computation in this method of making soil maps and argued that all maps are predictive of the true state of affairs. This is especially true for soil maps, where only a miniscule fraction of the soil is examined, even at its surface, let alone at depth.

Is DSM a case of doing it "because we can", or are there real advantages? Certainly, there are elements of the first – especially because it's accessible to IT-oriented people who have no or very limited field experience in traditional, landscape-based soil survey. A deluge of research papers compares digital methods in ever-finer detail (modelling methods, choice of covariates, dealing with under-sampled areas, *etc*). In defence of these authors, funding for systematic landscape-based soil survey is increasingly hard to come by, whereas funding is obtainable for point-based observations, especially for investigating to what degree spectroscopy can replace traditional laboratory analysis (*eg* Vågen *et al*, 2020). But mapping is then done by digital methods (*eg* Hengl *et al*, 2015).

Digital methods do have several advantages, as listed below.

- They are reproducible and objective: the same inputs and same model give the same output. This avoids the well-known issue of inconsistency among field soil surveyors which is only partially resolved by field correlation.

- Most methods provide an estimate of their accuracy and precision, which are related to the density of observations both in geographic space (for spatial interpolation methods) and their representativeness in covariate space. Thus, it is clear where more point observations would best be placed to improve the map (Brus, 2019).
- They ignore political and soil survey borders and, so, avoid the patchwork appearance of maps stitched together from diverse surveys.
- They predict areas where field survey is not possible, using the soil–environment relations developed in surveyed areas – obviously such maps must be used with caution.

Several disadvantages of digital methods are also evident: first, and most notably, their theory of soil geography is a correlation between observations and environmental covariates which are supposed to represent the soil-forming factors. This is far less comprehensive than a soil-geomorphic landscape analysis made by an experienced surveyor; if these covariates do not completely represent the factors, unlike (dissimilar) soils will be grouped together. Some covariates are either non-existent or at too coarse a scale to be useful, notably surficial lithology. Moreover, the soil-forming factor *time*, or age of the landform, is hard to represent by covariates since it requires a geomorphic analysis and estimates of past climates. Second, these models can only work with the profile observations provided to them, which rarely encompass the soil-geographic space because most field sampling plans were not designed to support DSM; indeed, such plans have only recently been developed and are still being refined (Wadoux & Brus, 2020). Examining the location of available soil profiles in feature space, *ie* the multivariate space covered by the covariates, can reveal holes in this space, where no soil observations have been made. Most machine-learning methods cannot extrapolate into these, and even methods that can (*eg* multiple linear regression) are of dubious validity.

Soil maps in global modelling

Recent years have seen a growing demand for soil maps to be used in global models, most notably of earth surface fluxes such as the Community Land Model (*eg* the soil hydrologic property maps produced by Dai *et al*, 2019). Such maps of derived soil properties, related directly to soil functions, of course rely on maps of primary soil properties and, ideally, also on maps of soil classes that serve to zone pedotransfer models. Previous global maps such as the FAO-UNESCO Soil Map of the World and the Harmonized World Soil Database (IIASA *et al*, 2012) include large areas dependent on expert judgement, and are at a small scale (1 : 5 million and 1 : 1 million, respectively). Hence the demand for a higher-resolution, fully digital and more objective product.

Since the first attempt to make an internationally acceptable soil map of the world under the auspices of the Food and Agriculture Organization of the United Nations (FAO) in the 1960s, ISRIC – World Soil Information has been the key institution for collecting and harmonising soil observations, collecting maps as source materials for compiled maps, and coordinating the development of a consistent legend that has evolved over the years into today's *World Reference Base (WRB) for Soil Resources* (IUSS Working Group WRB, 2015). ISRIC is accredited as the world data centre for soils by the International Council for Science. Thus, as the demand for globally consistent digital soil maps became evident, ISRIC was from the first involved in the GlobalSoilMap.net consortium (Arrouays *et al*, 2014), but this attempt to build a global map from regional nodes faltered for a variety of pedo-political reasons, although it performed valuable work in standardisation (Science Committee, 2012).

SoilGrids

ISRIC then decided to see if a consistent global product could be produced directly. The result was SoilGrids1km, soon followed by SoilGrids250m, both under the leadership of my former PhD student Tomislav Hengl (Hengl *et al*, 2014, 2017). After Hengl left ISRIC to form his own environmental information company, EnvironmetriX (https://envirometrix.nl/), SoilGrids was developed into a second version (Poggio *et al*, 2021), the main digital soil mapper being Laura Poggio, who was recruited from the Macaulay Institute. This is the current flagship product. Grid cell size was kept at 250 m, considering the sparsity of the supporting point observations, especially in poorly surveyed areas, or those where point observations have not been (India), or cannot legally be (France), shared.

Figure 1 is a snapshot of the SoilGrids interactive map (https://soilgrids.org/). There is one map for each of the five physical, four chemical, and two derived properties and each of the six depth slices. This snapshot shows how country borders disappear into a consistent product. Figure 2 shows the prediction at a point. Notice the uncertainty bands around each prediction. A feature of many DSM methods, including the Quantile Random Forest method used in SoilGrids, is that they can provide a measure of uncertainty. Figure 3 shows the uncertainty of pH in the 0–5 cm layer in a small area south-west of Lake Naivasha, Kenya. The uncertainty is quantified as the ratio between the interquantile range (90 percent prediction interval width) and the median prediction. The pattern of uncertainty is based on the relation between observations and covariates. This layer can

be used to mask areas where predictions are too uncertain, and also for sensitivity analysis in models that use the layers.

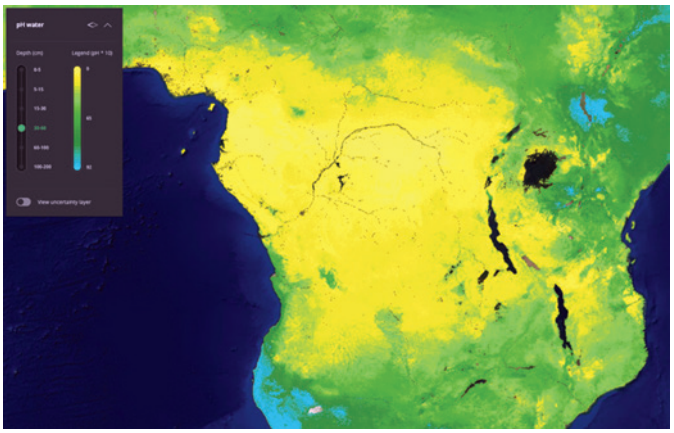

Figure 1. SoilGrids 2.0: Average pH (×10) of the 30–60 cm depth slice (Source: screen shot from SoilGrids v2.0 from ISRIC – World Soil Information; CC-BY 4.0)

Figure 2. Predicted soil properties at a grid cell west of Kampala, Uganda (Source: screen shot from SoilGrids v2.0 from ISRIC – World Soil Information; CC-BY 4.0)

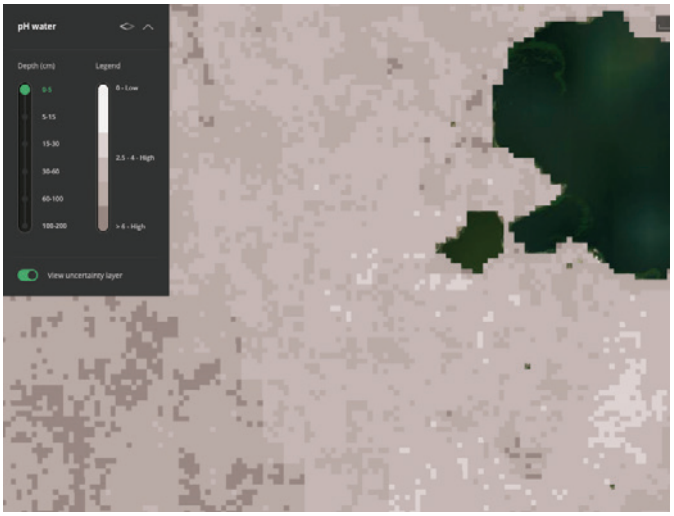

Figure 3. Uncertainty of pH prediction, south-west of Lake Naivasha, Kenya (Source: screen shot from SoilGrids v2.0 from ISRIC – World Soil Information; CC-BY 4.0)

The grids can be accessed for download in several ways, and are included as assets in Google Earth Engine (GEE). This allows the immense processing power and store of geographic information in GEE to be used along with soil properties. For example, Figure 4 shows a colour composite from a principal component analysis (clay, sand, bulk density, soil organic carbon [SOC], pH, cation exchange capacity [CEC], coarse fragments × 6 depth slices) of 42 layers, performed in GEE, of the Naivasha area. This is an objective zoning and clearly shows regions with substantially different soil properties. These can be used in cluster analysis to identify zones of relatively homogeneous soils for stratified sampling or field survey.

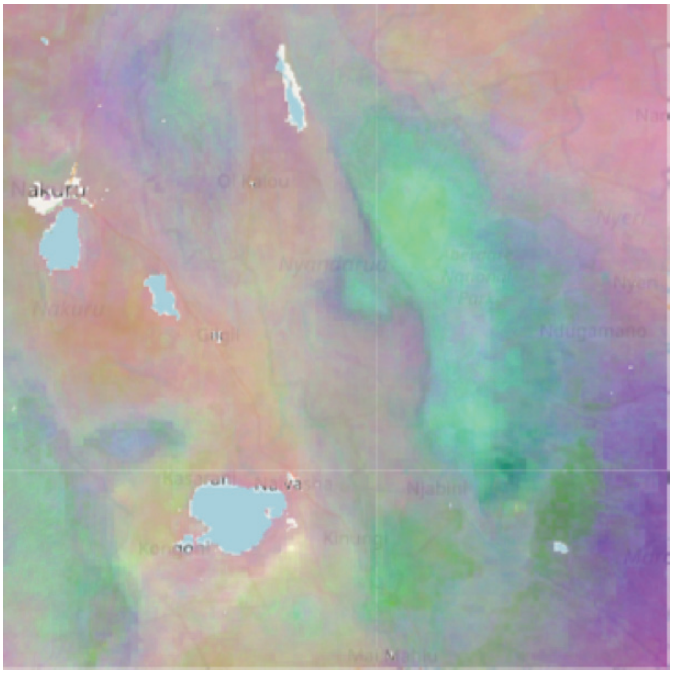

Figure 4. R–G–B colour composite from principal components 1, 2, 3 from 42 SoilGrids layers (clay, sand, bulk density, SOC, pH, CEC, coarse fragments ×6 depth slices), Naivasha area (Source: author's analysis)

Closer inspection of DSM products reveals some problems. Figure 5 shows a small area of Figure 4, at the south-east edge of Lake Naivasha. Notice the coarse grid cells (62.5 ha). The predictions are for each grid cell based on a point prediction at its centre. The resolution can be increased to the resolution of the covariates, subject to sufficient computer power. For example, the iSDA project (https://www.isda-africa.com/) has produced DSM maps at 30 m resolution (0.9 ha) for all of Africa except deserts (Hengl *et al*, 2021). An example in the same area as Figure 5 is shown in Figure 6. One may question whether this fine detail truly reflects differences in soil properties or is an artefact of the model. A similar effect, at the coarser resolution, can be seen in Figure 5. Recall that the machine learning models depend on point observations, which evidently are not sufficiently dense in covariate space. Yet the iSDA map is publicised as

"providing information at the scale of individual small farms across Africa" – notice the subtle qualifier "the scale of", not directly claiming that an individual small-scale farmer could use this map, as the cited paper makes clear. This illustrates a wider problem with DSM products: they can be oversold or misunderstood by prospective users. Of course, conventional soil maps have also been misunderstood, especially the concept that within a mapping unit at the design resolution there may be contrasting soils.

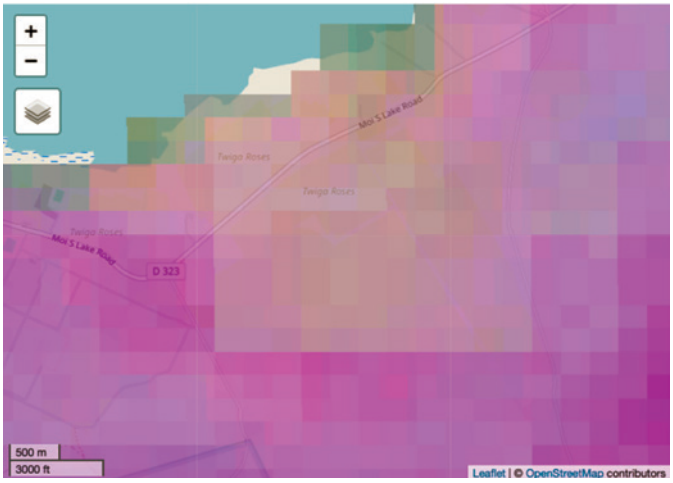

Figure 5. Detail of Figure 4, south-east edge of Lake Naivasha (Source: author's analysis)

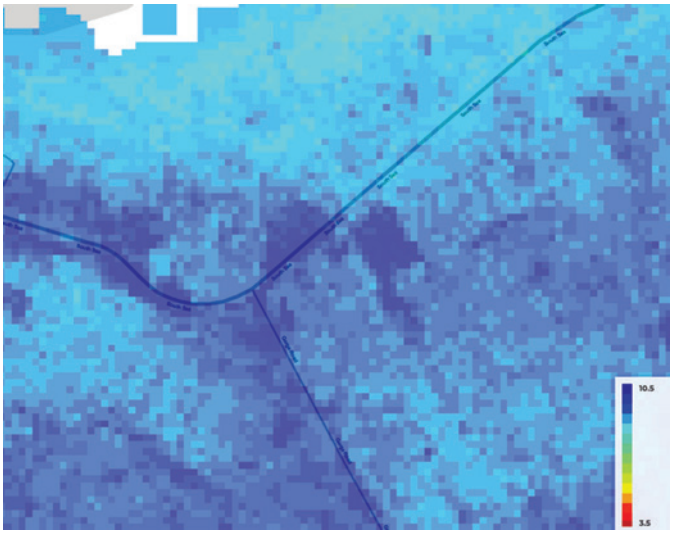

Figure 6. pH of 0–20 cm, iSDA 30 m resolution, south-east edge of Lake Naivasha (Source: iSDAsoil: Open Soil Data for Africa)

The theory underlying global-scale digital soil mapping is the 'Homosoil' idea (Mallavan *et al*, 2010), as operationalised in the SCORPAN framework (McBratney *et al*, 2003). The theory holds that under identical soil-forming conditions, identical soils form. This has been challenged on the grounds of chaos theory and contingency (Phillips, 2001), and in practical terms it is impossible to fully quantify the environment, over time, in which a soil developed. Still, the theory is used in machine-learning models to relate soils anywhere to a rich set of environmental covariates related to soil formation. These models are then applied everywhere. So, for example, soil observations and environmental covariates from Tanzania, Malaysia and Brazil are used together in one global model which is then applied everywhere in those countries.

The Homosoil theory, which assumes homology of soil-forming factors between a reference area and the region of interest (including climate, physiography and parent materials), is most useful for extrapolation to areas without field surveys or locally calibrated DSM products. These latter can be produced by the same methods as used in SoilGrids, but need sufficient observations in the area to be mapped.

A key step in the production of a global map is the harmonisation of soil observations from different soil surveys. This is the task of ISRIC's World Soil Information Service (WoSIS) database (Batjes *et al*, 2020). This contains observations for over 50 years; soils are dynamic so, for some properties such as soil organic matter, models may struggle to find strong correlations. Another problem is the poor georeference of older observations. The machine-learning model depends on correctly determining the environmental covariates at the observation's location by (GIS) map overlay. If the recorded location is not correct, the correlation will be poor. I have seen an example at the Kawanda Agricultural Research Station, near Kampala (Uganda), where several reference soil profiles had been described by a USAID mission in the 1980s (before high-precision GPS) and these data had been incorporated into WoSIS. The landscape at the station is the classic soil catena in the same landscape where Geoffrey Milne developed this concept in East Africa in the 1930s, with a period of about 500 m from hill crest to toe slope. Along with colleagues, I navigated to the recorded coordinates for the points and it was immediately clear that these could not be at the correct positions on the catena, nor were the soils at these locations similar to the recorded soils. We contacted the retired technician who had assisted the Americans, and he led us to the actual locations of their observations, where it was clear these were correct landscape positions and soils. In this case, I could record the correct georeference and alter the WoSIS record, but there must be many georeferences that distort the models.

Soil properties are interesting, but soil surveyors also use soil classes as carriers of holistic information, especially for technology transfer. To date, the results of DSM for soil classes have been disappointing. This is partly because the observations used as the basis for DSM were classified inconsistently (as we all know from arguments around the soil pit) and over a long period, and partly because the soil classes themselves

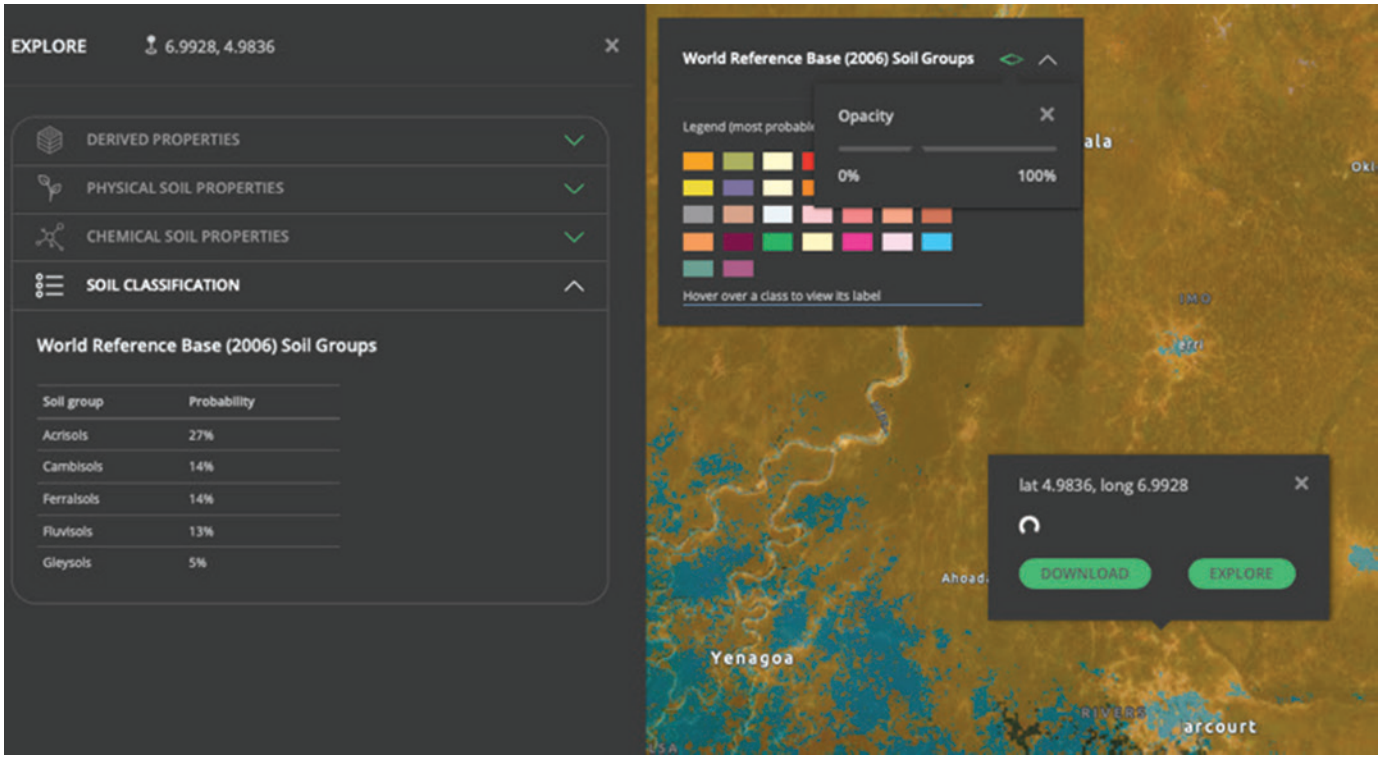

Figure 7. Predicted WRB Reference Soil Groups and their probability, near Port Harcourt (Nigeria). Map legend shows the most probable (Source: screen shot from SoilGrids v2.0 from ISRIC – World Soil Information; CC-BY 4.0)

are defined with rigid limits. DSM is, however, able to predict the probability of occurrence of each class and, thus, measures of uncertainty in the prediction (*eg* Shannon entropy). Again, this can be used in sensitivity analysis for models that use soil classes as inputs. Figure 7 shows the predicted WRB Reference Soil Groups and their probability, near Port Harcourt (Nigeria) – note the high uncertainty.

DSM and the modern soil surveyor's toolkit

So, how does DSM fit into the modern soil surveyor's or land evaluator's toolkit, especially in un- or under-surveyed areas? Maps produced by DSM can certainly be used in the absence of site-specific information to get a first idea of the soil properties within a grid cell. The uncertainty layers give an idea of the confidence the surveyor can have in the provided information. These maps are quite useful as pre-maps or, in Bayesian terms, prior information, either for field survey or local DSM using higher-resolution or area-specific covariates. Hengl *et al* (2021) explain these trade-offs in the context of the intended use of the DSM products:

> *"While there have been criticisms of the absolute accuracy of the iSDA soil maps, it is important to consider this in the context of real-world applications of the resource, for example in the generation of site-specific fertiliser recommendations. In this case, additional data collection would be required such as land use history, previous fertiliser applications and historic yields. However, we see this resource as a low-cost alternative to lab-based soil test that has value in reducing uncertainty around soil properties compared to having no information, which is especially relevant in a smallholder agriculture context. Our initial predictions are not likely to be correct enough to support informed management at the farm scale immediately. We can, however, propose our initial predictions as being relevant as a starting point, or base, that drives and informs additional new sampling, for each specific parcel of interest. … Promotion of first steps for basic improved crop management does not perhaps demand an exceptionally high accuracy of soil data. For example, a good estimate of soil pH can already help to inform which crops may be most suitable to grow / to not grow or if liming may be needed before any other agrochemicals are used" (emphasis added).*

Thus, DSM products can certainly be used in development projects, with appropriate cautions and understanding of the products – how they were produced, their data sources and their limitations. In

most areas where development work is proceeding, the major limitation to accurate DSM products is the density of appropriate soil observations (Loiseau *et al*, 2021); sampling campaigns to provide these can and should go hand in hand with field soil survey to inform and evaluate the predictive models.

References

Arrouays D, Grundy MG, Hartemink A *et al*, 2014. GlobalSoilMap: towards a fine-resolution global grid of soil properties. *Advances in Agronomy*, **125**, 93–134. https://doi.org/10.1016/B978-0-12-800137-0.00003-0

Batjes NH, Ribeiro E, van Oostrum A, 2020. Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data*, **12**(1), 299–320. https://doi.org/10.5194/essd-12-299-2020

Brus DJ, 2019. Sampling for digital soil mapping: a tutorial supported by R scripts. *Geoderma*, **338**, 464–480. https://doi.org/10.1016/j.geoderma.2018.07.036

Dai Y, Xin Q, Wei N *et al*, 2019. A global high-resolution data set of soil hydraulic and thermal properties for land surface modeling. *Journal of Advances in Modeling Earth Systems*, **11**(9), 2996–3023. https://doi.org/10.1029/2019MS001784

Hengl, T, de Jesus JM, MacMillan RA *et al*, 2014. SoilGrids1km – global soil information based on automated mapping. *PLoS ONE*, **9**(8), art. e105992. https://doi.org/10.1371/journal.pone.0105992

Hengl T, Heuvelink, GBM, Kempen B *et al*, 2015. Mapping soil properties of Africa at 250 m resolution: Random Forests significantly improve current predictions. *PLoS ONE*, **10**(6), art. e0125814. https://doi.org/10.1371/journal.pone.0125814

Hengl T, de Jesus JM, Heuvelink GBM *et al*, 2017. SoilGrids250m: global gridded soil information based on machine learning. *PLoS ONE*, **12**(2), art. e0169748. https://doi.org/10.1371/journal.pone.0169748

Hengl T, Miller MAE, Križan J *et al*, 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports*, **11**(1), art. 6130. https://doi.org/10.1038/s41598-021-85639-y

IIASA, FAO, ISRIC *et al*, 2012. *Harmonized World Soil Database (version 1.2)*. Rome and Laxenburg, Austria: Food and Agriculture Organization of the United Nations and International Institute of Applied Systems Analysis. http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HWSD_Documentation.pdf. Accessed 10 November Worldsoil-database/2021.

IUSS Working Group WRB, 2015. *World reference base for soil resources 2014: international soil classification system for naming soils and creating legends for soil maps, update 2015*. World Soil Resources Reports 106. Rome: Food and Agriculture Organization of the United Nations. http://www.fao.org/3/i3794en/I3794en.pdf. Accessed 10 November 2021.

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

Loiseau T, Arrouays D, Richer-de-Forges AC *et al*, 2021. Density of soil observations in digital soil mapping: a study in the Mayenne region, France. *Geoderma Regional*, **24**, art. e00358. https://doi.org/10.1016/j.geodrs.2021.e00358

Mallavan BP, Minasny B, McBratney AB, 2010. Homosoil, a methodology for quantitative extrapolation of soil information across the globe. In: Boettinger JL, Howell DW, Moore AC *et al*, eds, *Digital soil mapping*. Springer Netherlands, 137–150. https://doi.org/10.1007/978-90-481-8863-5_12

McBratney AB, Mendonça Santos ML, Minasny B, 2003. On digital soil mapping. *Geoderma*, **117**(1–2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Phillips JD, 2001. Contingency and generalization in pedology, as exemplified by texture-contrast soils. *Geoderma*, **102**(3–4), 347–370. https://doi.org/10.1016/S0016-7061(01)00041-6

Poggio L, Sousa LM de, Batjes NH *et al*, 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, **7**(1), 217–240. https://doi.org/10.5194/soil-7-217-2021

Science Committee, 2012. Specifications: tiered GlobalSoilMap.net products, release 2.3 [21/9/2012]. GlobalSoilMap.net. http://www.ozdsm.com.au/resources/GlobalSoilMap%20specs%20version%202point3.pdf. Accessed 10 November 2021.

Scull P, Franklin J, Chadwick OA, McArthur D, 2003. Predictive soil mapping: a review. *Progress in Physical Geography: Earth and Environment*, **27**(2), 171–197. https://doi.org/10.1191/0309133303pp366ra

Vågen T-G, Winowiecki LA, Desta L *et al*, 2020. Mid-infrared spectra (MIRS) from ICRAF Soil and Plant Spectroscopy Laboratory: Africa Soil Information Service (AfSIS) Phase I 2009-2013 [Data set]. World Agroforestry – Research Data Repository. https://doi.org/10.34725/DVN/QXCWP1

Wadoux AMJ-C, Brus DJ, 2020. How to compare sampling designs for mapping? *European Journal of Soil Science*, **72**(1), 35–46. https://doi.org/10.1111/ejss.12962

Photo: courtesy of Musfiqur Rahman