ARTICLE IN PRESS

Spatial Statistics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Spatial statistics and soil mapping: A blossoming partnership under pressure

Gerard B.M. Heuvelink [a,*], Richard Webster [b]

[a] ISRIC – World Soil Information and Wageningen University, Wageningen, The Netherlands
[b] Rothamsted Research, Harpenden, United Kingdom

ARTICLE INFO

Article history:
Received 6 December 2021
Received in revised form 4 February 2022
Accepted 4 February 2022
Available online xxxx

Keywords:
Digital soil mapping
Geostatistics
Kriging
Machine learning
Pedometrics
Soil

ABSTRACT

For the better part of the 20th century pedologists mapped soil by drawing boundaries between different classes of soil which they identified from survey on foot or by vehicle, supplemented by air-photo interpretation, and backed by an understanding of landscape and the processes by which soil is formed. Its limitations for representing gradual spatial variation and predicting conditions at unvisited sites became evident, and in the 1980s the introduction of geostatistics and specifically ordinary kriging revolutionized thinking and to a large extent practice. Ordinary kriging is based solely on sample data of the variable of interest—it takes no account of related covariates. The latter were incorporated from the 1990s onward as fixed effects and incorporated as regression predictors, giving rise to kriging with external drift and regression kriging. Simultaneous estimation of regression coefficients and variogram parameters is best done by residual maximum likelihood estimation. In recent years machine learning has become feasible for predicting soil conditions from huge sets of environmental data obtained from sensors aboard satellites and other sources to produce digital soil maps. The techniques are based on classification and regression, but they take no account of spatial correlations. Further, they are effectively 'black boxes'; they lack transparency, and their output needs to be validated if they are to be trusted. They undoubtedly have merit; they are here to stay. They too, however, have their shortcomings when applied to spatial data, which spatial statisticians can help overcome. Spatial statisticians and pedometricians still have much to do to incorporate uncertainty into digital predictions, spatial averages and totals over regions, and

* Corresponding author.
E-mail address: gerard.heuvelink@wur.nl (G.B.M. Heuvelink).

https://doi.org/10.1016/j.spasta.2022.100639

Please cite this article as: G.B.M. Heuvelink and R. Webster, Spatial statistics and soil mapping: A blossoming partnership under pressure. Spatial Statistics (2022) 100639, https://doi.org/10.1016/j.spasta.2022.100639.

to take into account errors in measurement and spatial positions of sample data. They must also communicate their understanding of these uncertainties to end users of soil maps, by whatever means they are made.

1. Introduction

Soil is crucial to our way of life. It provides us with much of our food and fibre and all of our timber. It accepts rain and melting snow, and it stores, filters and releases the water for our needs. It is in itself full of life. There are more than 100 000 different types of organisms living in it, and a single teaspoon of soil contains more living organisms than there are people on the planet (Magdoff and van Es, 2021). Soil is a heterogeneous intricate mixture of mineral particles and organic material in various forms of decay. It is not static. In the short term, such as the lifetime of a crop, physical, chemical and biological processes change its structure and composition. In the long term, from hundreds to millions of years, soil develops characteristics on rocks and in sediments in response to climate, tectonics and other land-forming processes. These processes are themselves highly complex, and the resulting complexity of the soil is hardly surprising.

Scientific study of the soil began around the middle of the 19th century, initially for understanding crop nutrition and as a basis for fertilizer practice. Towards the end of that century Russian scientists were studying variation in soil on a continental scale and related that variation to variation in climate. Agronomists were more concerned with variation within fields: why was crop growth so variable in what was thought to be homogeneous land? Mercer and Hall (1911) led the way with the first statistical analysis of spatial variation in field crops, which McBratney and Webster (1981) later showed could be attributed to an earlier ploughing regime. Between the two world wars several national governments initiated programmes of regional survey and mapping of soil. These were boosted in the USA after the 'dust bowl' of the 1930s and erosion under agriculture of fragile former forest soil in the southern states. After the second world war governments wanted maps of their whole countries (Young, 2017) as national inventories. Soil maps are now of greater importance as nations strive to feed their ever-growing populations on fixed areas of land, while conserving the soil and its biodiversity, maintaining its ability to store and purify water, all in a warming climate (Keesstra et al., 2016).

What is evident at all scales from field to continent is that soil is a continuum, broken only by water and bare rock, and that it is spatially correlated. Soil obeys Tobler's first law of Geography, namely 'Everything is related to everything else, but near things are more related than distant things' (Tobler, 1970). But one cannot know what the soil is like everywhere. One can at best sample on finite supports, and to map one must interpolate between sampling points. In traditional practice, which still has its role, one links observations of the soil in profile at sample sites to an understanding of landscape and perhaps natural vegetation, i.e. to what one can see above ground, whether from the ground or from above by remote sensing. Geostatistical interpolation depends more heavily on sampling to provide quantitative data from which to estimate correlation structure and predict at unobserved locations, though it may be augmented by other knowledge of the landscape. Sampling the soil need not be a serious limitation. Access is usually easy, and relevant soil properties such as its pH, salinity, concentrations of plant nutrients and pollutants are swift and cheap to measure, although measurements of soil properties, whether in the field or laboratory, incur error. Many soil properties are 'well behaved' statistically, smoothly varying spatially with near Gaussian distributions, or ones that can be readily normalized by transformation. It turns out that soil is an ideal medium for geostatistics (Webster, 2015, 2019).

The increasingly important role that geostatistics has played has radically changed how practitioners think of soil variation and map it. It has led to greater understanding of soil and a rich field of application. There are thousands of publications containing maps of soil properties made by kriging, and undoubtedly many more thousands of unpublished ones. Mapping this way has also thrown up new problems, solutions of which have accelerated further development of spatial statistics. In this paper we provide a brief historic account of soil mapping, with emphasis on geostatistical approaches. We accentuate contributions on the topic published in this journal during its first ten years of existence. In those ten years, however, there has been the new development in machine learning. This has tended to replace geostatistics as the main soil mapping methodology. Machine learning has its place now that sensors aboard satellites can monitor the land surface in fine detail, but it does not account for spatial correlation explicitly. One may therefore ask whether there is still a need for *spatial* statistics in soil mapping. Our assertion is that there is, and we end with a list of outstanding challenges that pedometricians and spatial statisticians need to face and solve with the backing of spatial statistics. While these challenges are formulated in the context of soil mapping, we believe that they are relevant to a much wider application domain than just soil science and help define a more general spatial statistics research agenda.

2. Spatial statistics for soil mapping

The importance of geostatistics for soil mapping has grown over the years. Although soil surveyors recognized Tobler's law long before Tobler enunciated it, they did not know how to express it in their maps. The breakthrough came in the late 1970s, in particular as theory and practice in mining found its way into soil science, leading to the first published accounts of kriging for soil mapping in 1980 (Burgess and Webster, 1980a,b). In the next 20 years kriging in one form or another dominated the literature on soil mapping. It was elaborated to take into account non-Gaussian distributions, co-regionalized variables, geographic trend and external drift. In the current century there have been further developments to combine fixed and random effects by residual maximum likelihood estimation (Lark et al., 2006) and to deal with non-stationary variance. At the same time machine learning has become popular, particularly for mapping large regions.

2.1. Conventional soil survey

Traditionally surveyors mapped the soil on foot or by vehicle. They observed the landscape, vegetation and land use, and they would supplement those observations by information from air-photographs, geological maps and other sources. They dug pits and augered at places of interest, and they drew boundaries between what they regarded as different kinds of soil. Each kind of soil on their maps, once derived, would be characterized by descriptions of one or more representative soil profiles with material analysed in the laboratory. Spatial statistics was not used formally, although the surveyors would intuitively sample more densely where spatial variation was greater than in more homogeneous areas.

Engineers were among the first scientists to recognize that the utility of such maps depended on their statistical characteristics: variances within the classes depicted would determine how well one could predict from them. Morse and Thornburn (1961) sampled to a random design soil maps made by the US Department of Agriculture to estimate means and variances of the classes mapped; Kantey and Williams (1962) analysed their own maps similarly. Somewhat later, Beckett and Webster (1971) collaborated with the engineers of the British Army to evaluate soil conditions for roads and temporary airfields. Their investigations showed not only the merits of conventional mapping techniques at the time but also their shortcomings; the techniques failed to take into account spatial correlation within the mapped classes, and they sparked the first quantitative demonstration of that correlation by Webster and Cuanalo de la C. (1975). Civil engineers are now well versed in geostatistical techniques and use these methods in site investigations. Auvinet (2019) recently reviewed the subject with a substantial appendix setting out the mathematics with which pedometricians are now familiar.

There is no doubt that the practice of conventional soil mapping worked well at some scales, at say 1:50 000 for mapping regions of several hundred of square kilometres. But it needed a scientific basis, and that was provided by Jenny (1941) in his classic text. Soil is not isolated from its environment but results from many processes acting on rocks and sediments over time. Jenny spelled out five principal sources, or factors, namely climate, organisms, relief, parent material and time. He expressed the dependence of these factors in the form of a mathematical function. It was not soluble, but it was the concept that was important.

2.2. Geostatistics and digital soil mapping

Conventional soil mapping falls short where change is gradual, where the landscape is obscured by dense forest or woodland and where the soil bears little relation to soil forming factors, such as in agricultural fields and for pollutants. In these situations the main sources of information are measurements made in the laboratory or directly in the field at geo-referenced sampling sites. It was there that geostatistics entered into soil mapping with the pioneering work of Webster and colleagues. They introduced the variogram to describe what appeared to be correlated random variation, estimated it from sample data by the method of moments and fitted authorized models by non-linear least-squares approximation, which at the time was novel. It was a logical step from there to interpolate on to fine grids by ordinary kriging for mapping. Their first paper (Burgess and Webster, 1980a) is now recognized as a landmark (Lark et al., 2019).

Other papers followed: Burgess and Webster (1980b) to estimate over blocks; Webster and Burgess (1980) to deal with local drift; Burgess et al. (1981) to optimize sampling; McBratney and Webster (1983) to incorporate random covariates in a coregionalization; Webster and Oliver (1989) to estimate conditional probabilities of exceeding, or not, specified thresholds; and McBratney and Webster (1986) to choose rationally from among two or more plausible competing variogram models. The focus was on getting the most out of the point observations by modelling anisotropy, non-stationarity and non-normality, as well as applying change-of-support algorithms and optimizing sampling designs. Kriging was not the only technique for what might be called 'contouring' from sample data. Several forms of inverse distance weighting were popular at the time. A major advantage of kriging, however, is that it minimizes the prediction error variances and reports them. In other words it provides measures of uncertainty in the predictions which few other techniques could do. Heuvelink et al. (1989) and Goovaerts (2001) went on to show how this uncertainty in soil maps could be propagated into subsequent analyses.

A shortcoming of kriging in soil science at that time was that it disregarded environmental variables, known or assumed to be correlated with target soil variables, such as the soil-forming factors embodied in Jenny's conceptual equation. These covariates might be the height or steepness of the land, annual rainfall or temperature, or vegetation density. They were available from remote sensing and known at all prediction locations, which meant that they could be included as fixed effects. Means were needed of incorporating that knowledge. Matheron (1969) had already set out the mathematics of doing that in his generalized model which he called 'universal kriging'. In practice there was a major problem; one needed the variogram of the random residuals from the drift, and one could not obtain those residuals without knowing the variogram. Nevertheless, a rough-and-ready empirical solution seemed to be to fit by ordinary least-squares regression the target variable on values of the covariates, analyse the residuals geostatistically, and then add back the predictions from the regression to the kriged prediction of the residuals. This was in essence the basic form of regression kriging. The underlying geostatistical model may be expressed as

$$Z(\mathbf{s}) \;=\; m(\mathbf{s})+\varepsilon(\mathbf{s}) \;=\; \beta_0+\sum_{i=1}^{p}\beta_i\cdot f_i(\mathbf{s})+\varepsilon(\mathbf{s}) \;. \tag{1}$$

Here Z is a soil property of interest, $\mathbf{s} \equiv \{s_1, s_2\}$ typically in two dimensions is geographic location, m is a drift (i.e., trend) term that is a linear combination of p environmental covariates f_i, and ε a correlated random residual with zero mean and variance σ^2. This is what spatial statisticians will recognize as the 'linear mixed model' (Lark, 2012). It has more parameters than the ordinary kriging model, and pedometricians have contributed greatly to proper estimation of the model parameters through residual maximum likelihood (REML) estimation (Lark et al., 2006).

Many extensions of the regression kriging model have been made, such as its application to three-dimensional and space–time mapping (Gasch et al., 2015; Cappello et al., 2021), as well as extensions to Bayesian (Steinbuch et al., 2022) and generalized linear modelling (Kempen et al., 2012a). The most important feature of all these is that environmental covariates are used to help explain the soil spatial variation. In principle, the more of these external covariates one can incorporate to improve the prediction the less important does the random component, the essence of ordinary and simple kriging, become.

2.3. Digital soil mapping using machine learning

With the advance of remote sensing and increasing availability of environmental covariates, attention was drawn to improving the modelling of the trend component in Eq. (1). This can be achieved by increasing the number of covariates incorporated in the model and by using a flexible, non-linear model structure. For example, the most recent SoilGrids model starts out with a set of more than 400 covariates (Poggio et al., 2021). Dropping the restriction that the trend is a linear combination of the covariates leads to a more flexible model and should be able to make prediction more accurate. Indeed, several case studies have shown how the introduction of machine learning algorithms has benefited digital soil mapping (e.g. Hengl et al. (2015), Chen et al. (2019)). The algorithms include random forests, gradient boosting and artificial neural networks, and the current focus in digital soil mapping research is on modelling the trend using these algorithms. While traditional regression and kriging rely on explicit statistical models, machine learning does not; it is based purely on algorithms. The approach undoubtedly has important merits, but it must be used with caution because it can be sensitive to over-fitting and lacks transparency (Spiegelhalter, 2019).

Machine learning is in essence non-spatial. It operates on a ‘regression matrix’ that stores paired observations of the dependent (target) and independent variables, without regard to their geographic coordinates. Several innovations have been proposed and implemented to make it more spatial. These include defining ‘spatial’ covariates (Hengl et al., 2018), covariates that are derived from the nearest observations and their distances (Sekulić et al., 2020), and expansion of the set of covariates with covariate values in local neighbourhoods of the prediction location (Wadoux et al., 2019b). These approaches operate only on the covariates, however, while the machine-learning algorithm itself is still non-spatial. The challenge is to account for spatial autocorrelation explicitly, basically in a way similar to extending a multiple linear regression model to the linear mixed model (Lark, 2012) or to incorporating spatial correlation in structural equation modelling (Angelini and Heuvelink, 2018).

Machine learning works well with large sets of data (‘big data’), and it is in these situations that it is finding its main role in digital soil mapping. Satellites are now providing complete cover of environmental data on land surface conditions and topographic height at fine resolutions. If there is strong correlation between them and a target soil property then any residuals from machine-learning regression are likely to be unimportant for mapping; variograms of the residuals will be almost flat, and kriging will enhance the predictions rather little (Van der Westhuizen et al., 2022).

One must therefore ask: Has kriging a future for soil mapping? Is geostatistics for digital soil mapping dead? As it happens, not all soil properties can be easily related scientifically to environmental covariates, and it is poor science to load large sets of environmental variables into machines and expect them to produce sense, even if they happen to give plausible results (Wadoux et al., 2020). Plant nutrients within fields and pollutants in larger regions from diffuse or unknown sources are examples where covariates may not be very helpful. Mapping them is likely to depend largely on sample data and geostatistical interpolation for years to come unless proximal sensors are developed specifically for them and can be used ‘on the run’. The idea is not new. Haines and Keen (1925), in a remarkable investigation, drew a plough through the soil, measured the soil’s strength by dynamometer and recorded the drawbar pull as continuous traces on paper charts. They then interpolated between the traces by hand to produce ‘contour’ maps of the soil’s strength, which they related to the proportion of clay in the soil. It was way ahead of its time; it would be another 80 years or so before proximal sensors would be developed and tested to map soil properties ‘on-the-run’ (Adamchuk and Viscarra Rossel, 2010).

3. The importance of soil mapping for spatial statistics

Soil science is only one of the domains in which spatial statistics plays a part, and it would be a gross exaggeration to claim that spatial statistics needs soil science. As mentioned above, however, the soil does provide a perfect medium for geostatistics. Pedometricians and spatial statisticians working in soil science have greatly contributed to the development and extension of geostatistical methods, such as regression kriging (Hengl et al., 2004), residual maximum likelihood estimation for the mixed linear model (Lark, 2012), non-stationarity in the mean and variance (Lark, 2009; Wadoux et al., 2018), change of support (Orton et al., 2016), and optimization of sampling designs (Van Groenigen and Stein, 1998; Brus, 2014).

From its outset, the journal *Spatial Statistics* has been clear about the importance of applications, not only to steer the development of new methods but also to demonstrate the practical use and relevance of these methods. To illustrate this we need only to quote from the journal's aims and scope: "Spatial Statistics publishes articles on the theory and application of spatial and spatio-temporal statistics. It favours manuscripts that present theory generated by new applications, or in which new theory is applied to an important practical case. A purely theoretical study will only rarely be accepted". The important role of soil applications is demonstrated by the fact that in the first ten years of its existence, the journal published 25 articles on the application of spatial statistics in soil science, most of which were related to modelling soil spatial variation and soil mapping.

One of the drivers of spatial statistics in soil science has come from what is known as precision agriculture. With modern machines farmers can now vary the amounts of fertilizer and other agricultural chemicals on-the-run in response to crops' requirements. In principle, a soil map in digital form can be held in an on-board computer, and the precise amounts of chemical to be applied calculated as a machine travels across the land. Oliver (2010) assembled a set of articles setting out both the theory and potential applications of geostatistics to provide the necessary information. As it happens, however, the cost of sampling and analysing the soil to map the main plant nutrients available nitrogen (N), phosphorus (P) and potassium (K) is prohibitive for all but the most valuable crops such as lettuce (Panagopoulos et al., 2006). Nevertheless, recent advances in remote sensing and visible–infrared spectroscopy promise to overcome this stumbling block for P and K (Breure et al., 2022). Pest control is another aspect of farming on which spatial statistics has potential benefits for both farmers and the environment. Evans et al. (2003) showed that the cost of sampling soil to map infestation by the potato cyst nematodes could be justified by farmers' savings on expensive toxic nematicide where there are few or no nematodes. Metcalfe et al. (2016) analysed the spatial relations between soil and the pernicious grass weed *Alopecurus myosuroides* with a view to its control by locally targeted spraying with herbicide.

Most farmers do not have the most modern machines and detailed digitized soil maps. They have divided variable land into relatively homogeneous parcels—fields and paddocks—based on centuries of experience. Soil scientists are now helping them to refine those divisions geostatistically; the aim is to delineate homogeneous zones that differ substantially from neighbouring ones. Oliver and Webster (1989) and Bourgault et al. (1992) suggested how this might be done by modifying the multivariate *k*-means algorithm to take into account the spatial correlation among the sampling points. The idea is now being put into practice, in particular by scientists at the university of Bari. Guastaferro et al. (2010) compared several algorithms for the purpose, and Diacono et al. (2013) and Buttafuoco et al. (2021) have put them into effect on farms. Converting optimized spatial predictions into less-than optimal classifications might seem reverse engineering, but if that helps farmers then so be it.

4. Future roles of spatial statistics for soil mapping

We have described above how spatial statistics and machine learning have revolutionized soil mapping in the past four decades. Digital soil mapping has largely replaced conventional soil mapping; it is cheaper, more transparent and therefore reproducible, and more easily updated (Kempen et al., 2012b; Minasny and McBratney, 2016). The number of journal articles on digital soil mapping has grown exponentially during the past 20 years (Piikki et al., 2021), and the trajectory seems not yet to be flattening. We also noted that machine learning, which is in essence algorithmic and not spatial, has tended to oust geostatistics for soil mapping in recent years.

There are nevertheless many situations in which machine learning will not work, for example, where there are few covariates, where available environmental variables are only weakly related to the target variable, and when there are too few data for calibration. One must also realize that training a model on only a regression matrix of paired observations of dependent and independent variables and ignoring their spatial interrelations is inevitably sub-optimal. We have already pointed out that such relations were crucial to understanding and weed control in the study by Metcalfe et al. (2016) cited above. This paper is therefore also a call on spatial statisticians to show that spatial statistics still has an important role to play in modelling soil variation and soil mapping.

Scientific developments are notoriously difficult to plan and predict, but to provide some guidance and inspiration, below we set out eight challenges that pedometricians and spatial statisticians could tackle to strengthen the role of spatial statistics. We do so in the context of soil mapping, but many are relevant to a much wider set of applications.

1. *Machine learning*. Machine learning is here to stay. Its flexibility in modelling the relation between dependent and independent variables is a great asset. It has been found to improve substantially the accuracy of soil maps when used appropriately (e.g. Hengl et al. (2015)) with many data and covariates at fine resolutions. It is less effective where data are sparser, and here spatial statisticians should embrace the techniques and make them more 'spatial'. We want machine-learning algorithms that incorporate spatial dependence in the models they select, and use it for calibration, prediction and simulation. This goes much further than simply adding 'spatial context' to covariates as in Hengl et al. (2018), Sekulić et al. (2020) and Wadoux et al. (2019b). Perhaps there is merit in extending the Long-Short-Term Memory neural network (Li et al., 2021; Sungmin and Orth, 2021) from time to space, or by incorporating Moran's I or the variogram in machine learning (Nikparvar and Thill, 2021). Or perhaps it is as easy as using a geographically weighted machine learning approach (Ye et al., 2017; Li, 2019). These are recent developments that have not yet made it to this journal, which has had only seven citations on 'machine learning' in the Web of Science in its first ten years of existence, and all of these referred to use of standard machine learning to spatial data.
2. *Uncertainty quantification of spatial averages and totals*. Pedometricians take pride in the fact that their maps of predictions are routinely accompanied by maps of the uncertainties of those predictions. A soil map is considered incomplete unless its uncertainties have been quantified explicitly (Heuvelink, 2014; Veronesi and Schillaci, 2019). This is one of the strengths of kriging. Machine learning was and still lags behind, but with the development of quantile regression forests (Meinshausen, 2006), an approach became available that is now widely used in soil mapping (e.g. Vaysse and Lagacherie (2017), Szatmári and Pásztor (2019), Poggio et al. (2021)). This method, however, cannot quantify the uncertainty of spatial averages and totals, which block kriging can do, because it does not model the spatial dependence. Also here spatial statistics must come to the rescue, and it should look for more advanced ways than geostatistical modelling of the residuals from machine learning (e.g. Szatmári et al. (2021)). There is a great need to be able to quantify uncertainty of spatial averages and totals. For instance, in the context of mitigating the effects of climate change, Heuvelink et al. (2021) mapped the space–time variation of stocks of organic carbon in soil in Argentina, but they could not detect statistically significant changes over time at point support level, because of the large uncertainty. They noted that uncertainty would average out under spatial aggregation, but they could not quantify it. Similar problems are faced in other application fields, such as space–time modelling of above-ground biomass. In attempts to quantify prediction uncertainties of spatial averages, Plaza et al. (2018) assumed that prediction errors were spatially uncorrelated; Harris et al. (2021) in contrast took the opposite approach by assuming that errors were perfectly spatially correlated. The truth lies somewhere in between these extremes, and we need spatial statistical approaches to find it.

3. *Change of support along the soil profile.* We noted above the need for spatial statistics to handle spatial aggregation properly. The problem was cast in the context of a lateral change of support, but the problem occurs in the vertical dimension too. There it is even more challenging for soil science, and for several reasons. Soil data are often collected by horizons which vary in thickness and depth rather than at fixed vertical intervals. In the latter case pedologists may wish to predict over smaller depth intervals than those on which the measurements were made, i.e. on smaller vertical supports. Orton et al. (2016) have made important steps to solve the problem, but methods need to be developed to make them operational. These analyses are further complicated by the large anisotropies between lateral and vertical spatial soil variation and the fact that soil vertical variation is far from stationary (see Li et al. (2016)). Also, although properties tend to be fairly uniform within soil horizons they can change abruptly from one horizon to another.
4. *Sampling designs for soil mapping.* Much research has been done to optimize sampling designs for spatial and space–time soil mapping (e.g. Marchant et al. (2013), Brus (2014), Ließ (2015), Brus (2019)). Nevertheless, the methods need to keep pace with developments in modelling and to adjust to machine learning (Wadoux et al., 2019a). In a recent paper that published ten challenges for the future of pedometrics (Wadoux et al., 2021b), the question was raised whether we can "develop new sampling strategies, mapping models and decision making tools that can synergistically use and consider different sources of soil data (qualitative soil profiles descriptions, soil maps, soil sensing data, citizen data) that have different precision, geographical supports, resolutions and extents".
5. *Incorporating attribute and positional measurement uncertainty.* There are many sources of uncertainty in soil mapping, but one that is often ignored is that arising from errors of measurement and observations. This is serious, because these uncertainties can be substantial (Van Leeuwen et al., 2022). Measurement error can be readily incorporated in kriging (stemming from the work of Delhomme (1978)), but incorporating it in machine learning is much less well-developed (Van der Westhuizen et al., 2022). Note that many soil data sets have both attribute and positional measurement uncertainty. The latter can have major repercussions for soil modelling when high-resolution covariates are matched with erroneous soil sampling locations. It is useful to note that positional error can be translated into measurement error and incorporated in kriging systems. When a perfect measurement at a location, say $\mathbf{s} + \mathbf{a}$, is taken to represent a soil property at location $\mathbf{s}$, with $\mathbf{a}$ a possible positional error, then short-distance spatial variation means that the soil property at $\mathbf{s}$ may differ from that at $\mathbf{s} + \mathbf{a}$. Therefore, even if the measurement at location $\mathbf{s} + \mathbf{a}$ is error-free, it still represents the soil property value at $\mathbf{s}$ with error. The magnitude of this error can be derived from the short-distance spatial variation.
6. *Cross-validation for spatially clustered data.* In addition to spatially explicit quantification of uncertainty (see point 2 above), many digital soil mappers also routinely assess the overall quality of their maps using independent validation data. This is best done by probability sampling (Brus et al., 2011), but that incurs an additional cost to survey, and in many cases one has to make do with cross-validation. Here, concerns have been raised about cross-validation locations being geographically close to calibration locations, and 'spatial cross-validation' has been proposed (e.g. Brenning (2005), Ploton et al. (2020)). But this approach starts from the wrong premise that spatial autocorrelation should be avoided, and it tends to produce validation metrics of accuracy that are too pessimistic (Wadoux et al., 2021a). In contrast, if data are spatially clustered then standard cross-validation produces results that are too optimistic. Spatial statisticians need to develop workable and less-biased assessments of map accuracy where probability sampling is impracticable or unaffordable. Note that these validation exercises must not only evaluate the accuracy of the predictions but also the uncertainties of those predictions. This can be done with the prediction interval coverage probability (PICP, Shrestha and Solomatine (2006)) and accuracy plots (Goovaerts, 2001). It is intriguing that different mapping methods can yield quite different uncertainty maps, while validation by PICP and accuracy plots cannot reveal that one approach is better than another (Vaysse and Lagacherie, 2017; Szatmári and Pásztor, 2019).

7. *Generalized linear geostatistical model.* The generalized linear geostatistical model and its Bayesian extension (Diggle and Ribeiro Jr., 2007) have been used for mapping soil classes and binary soil properties, but there have been only a few cases (e.g. Kempen et al. (2012a), Steinbuch et al. (2022)). Two reasons for there being so few examples are (a) the statistical and computational complexity; and (b) the difficulty of modelling strong spatial dependence, because the model assumes that the response is spatially independent conditional on the signal (Diggle and Ribeiro Jr. (2007), Section 1.4). It might be useful to explore other models of spatial distribution of categorical soil variables, such as discrete Markov random fields and autologistic models (Caragea and Berg, 2014; Hristopulos, 2020).
8. *Uncertainty communication and spatial decision making under uncertainty.* While much attention is paid to quantifying uncertainty in digital soil maps, and its importance is recognized by research scientists, it turns out to be more difficult to visualize and communicate uncertainty to end users and to advise them how uncertainty maps can be used for analysing risk and making decisions. Milne et al. (2015), Yoo et al. (2018) and Zheng et al. (2021) discuss the matter, but the topic needs much more attention, by pedometricians (Wadoux et al., 2021b) and spatial statisticians alike.

5. Conclusion

During the past four decades soil mapping has benefited tremendously from spatial statistics. Countless soil maps have been produced by kriging in its various forms in that time. Spatial statistics has also benefited from soil mapping, because practical applications have been powerful catalysts for development of new methods. It is gratifying that many spatial statisticians have been inspired and motivated to tackle real-world problems for which pedologists have sought solutions.

The fruitful partnership between pedologists and spatial statisticians is under pressure because software for machine learning is readily available to analyse huge sets of dense environmental data from satellites and other sources, and map makers see little need to incorporate spatial statistical techniques into their procedures. This last is unfortunate; those who abjure spatial statistics fail to understand its functions and potential benefits both for prediction and understanding. There is an educational role for spatial statisticians; they need to explain to pedologists what those functions and benefits are; they need to show that there is added value in incorporating spatial dependence in the choice of models, in calibration, prediction and simulation. We have listed several challenges that require spatial statistics. Addressing these will not only advance soil mapping but will also have positive rewards in understanding the Earth and environmental sciences much more broadly.

References

Adamchuk, V.I., Viscarra Rossel, R.A., 2010. Development of on-the-go proximal sensor systems. In: Viscarra Rossel, R.A., McBratney, A.B., Minasny, B. (Eds.), Proximal Soil Sensing. Springer, Dordrecht, pp. 15–28.
Angelini, M.E., Heuvelink, G.B.M., 2018. Including spatial correlation in structural equation modelling of soil properties. Spatial Stat. 25, 35–51.
Auvinet, G.Y., 2019. Geotechnical engineering in spatially variable soft soils. The case of Mexico city. The 9th Arthur Casagrande Lecture. In: López Acosta, N.P., Martínez Hernández, E., Espinosa Santiago, A.L. (Eds.), The Proceedings of the XVI Pan-American Conference on Soil Mechanics and Geotechnical Engineering (XVI PCSMGE). International Society for Soil Mechanics and Geotechnical Engineering, London, pp. 13–107.
Beckett, P.H.T., Webster, R., 1971. The development of a system of terrain evaluation over large areas. Royal Eng. J. 85, 243–258.
Bourgault, G., Marcotte, D., Legendre, P., 1992. The multivariate (co)variogram as a spatial weighting function in classification methods. Math. Geol. 24 (5), 463–478.
Brenning, A., 2005. Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat. Hazards Earth Syst. Sci. 5 (6), 853–862.
Breure, T.S., Haefele, S.M., Hannam, J.A., Corstanje, R., Moreno-Rojas, S., Webster, R., Milne, A.E., 2022. A loss function to evaluate agricultural decision making under uncertainty: a case study of soil spectroscopy. Precis. Agric. in press.
Brus, D.J., 2014. Statistical sampling approaches for soil monitoring. Eur. J. Soil. Sci. 65 (6, SI), 779–791.
Brus, D.J., 2019. Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma 338, 464–480.
Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil. Sci. 62 (3), 394–407.
Burgess, T.M., Webster, R., 1980a. Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. J. Soil. Sci. 31 (2), 315–331.

Burgess, T.M., Webster, R., 1980b. Optimal interpolation and isarithmic mapping of soil properties. II. Block kriging. J. Soil. Sci. 31 (2), 333–341.
Burgess, T.M., Webster, R., McBratney, A.B., 1981. Optimal interpolation and isarithmic mapping of soil properties. IV. sampling strategy. J. Soil. Sci. 32 (4), 643–659.
Buttafuoco, G., Quarto, R., Quarto, F., Conforti, M., Venezia, A., Vitti, C., Castrignanò, A., 2021. Taking into account change of support when merging heterogeneous spatial data for field partition. Precis. Agric. 22 (2, SI), 586–607.
Cappello, C., De Iaco, S., Palma, M., Pellegrino, D., 2021. Spatio-temporal modeling of an environmental trivariate vector combining air and soil measurements from Ireland. Spatial Stat. 42, 100455.
Caragea, P.C., Berg, E., 2014. A centered bivariate spatial regression model for binary data with an application to presettlement vegetation data in the Midwestern United States. J. Agric. Biol. Environ. Stat. 19 (4), 453–471.
Chen, L., Ren, C., Li, Y., Wang, Y., Zhang, B., Wang, Z., Lu, L., 2019. A comparative assessment of geostatistical, machine learning, and hybrid approaches for mapping topsoil organic carbon content. Int. J. Geogr. Inf. Sci. 8, 174.
Delhomme, J.P., 1978. Kriging in the hydrosciences. Adv. Water Resour. 1 (5), 251–266.
Diacono, M., De Benedetto, D., Castrignanó, A., Rubino, P., Vitti, C., Abdelrahman, H.M., Sollitto, D., Cocozza, C., Ventrella, D., 2013. A combined approach of geostatistics and geographical clustering for delineating homogeneous zones in a durum wheat field in organic farming. NJAS 64-65, 47–57.
Diggle, P.J., Ribeiro Jr., P.J., 2007. Model-Based Geostatistics. Springer, New York.
Evans, K., Webster, R., Barker, A., Halford, P., Russell, M.D., Stafford, J., Griffin, S., 2003. Mapping infestations of potato cyst nematodes and the potential for spatially varying applications of nematicides. Precis. Agric. 4, 149–162.
Gasch, C.K., Hengl, T., Gräler, B., Meyer, H., Magney, T.S., Brown, D.J., 2015. Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D+T: The Cook Agronomy Farm data set. Spatial Stat. 14 (A), 70–90.
Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103 (1-2), 3–26.
Guastaferro, F., Castrignanò, A., De Benedetti, D., Sollito, D., Troccoli, A., Catarelli, B., 2010. A comparison of different algorithms for the delineation of management zones. Precis. Agric. 11, 600–620.
Haines, W.B., Keen, B.A., 1925. Studies in soil cultivation. III. Measurements on the Rothamsted classical plots by means of dynamometer and plough. J. Agric. Sci. 15, 395–406.
Harris, N.L., Gibbs, D.A., Baccini, A., Birdsey, R.A., de Bruin, S., Farina, M., Fatoyinbo, L., Hansen, M.C., Herold, M., Houghton, R.A., Potapov, P.V., Suarez, D.R., Roman-Cuesta, R.M., Saatchi, S.S., Slay, C.M., Turubanova, S.A., Tyukavina, A., 2021. Global maps of twenty-first century forest carbon fluxes. Nat. Clim. Chang. 11 (3), 234–240.
Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., de Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLOS ONE 10 (6), e0125814.
Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120 (1-2), 75–93.
Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.
Heuvelink, G.B.M., 2014. Uncertainty quantification of GlobalSoilMap products. In: Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A., McBratney, A.B. (Eds.), GlobalSoilMap. Basis of the Global Spatial Soil Information System. CRC Press, Boca Raton, pp. 335–340.
Heuvelink, G.B.M., Angelini, M.E., Poggio, L., Bai, Z., Batjes, N.H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G.F., Sanderman, J., 2021. Machine learning in space and time for modelling soil organic carbon change. Eur. J. Soil. Sci. 72 (4), 1607–1623.
Heuvelink, G.B.M., Burrough, P.A., Stein, A., 1989. Propagation of errors in spatial modelling with GIS. Int. J. GIS 3, 303–322.
Hristopulos, D.T., 2020. Random Fields for Spatial Data Modeling. Springer, Dordrecht.
Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. Dover Publications, New York.
Kantey, B.A., Williams, A.A.B., 1962. The use of soil engineering maps for road projects. Trans. South Afr. Inst. Civil Eng. 4, 149–159.
Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerda, A., Montanarella, L., Quinton, J.N., Pachepsky, Y., van der Putten, W.H., Bardgett, R.D., Moolenaar, S., Mol, G., Jansen, B., Fresco, L.O., 2016. The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. Soil 2 (2), 111–128.
Kempen, B., Brus, D.J., Heuvelink, G.B.M., 2012a. Soil type mapping using the generalised linear geostatistical model: A case study in a Dutch cultivated peatland. Geoderma 189, 540–553.
Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., de Vries, F., 2012b. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Sci. Am. J. 76 (6), 2097–2115.
Lark, R.M., 2009. Kriging a soil variable with a simple nonstationary variance model. J. Agric. Biol. Environ. Stat. 14, 301–321.
Lark, R.M., 2012. Towards soil geostatistics. Spatial Stat. 1, 92–99.
Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil. Sci. 57 (6), 787–799.
Lark, R.M., Heuvelink, G.B.M., Bishop, T.F.A., 2019. Burgess, T.M. & Webster, R. 1980. Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. Journal of Soil Science, 31, 315-331. Eur. J. Soil. Sci. 70, 7–10.
Li, Lianfa, 2019. Geographically weighted machine learning and downscaling for high-resolution spatiotemporal estimations of wind speed. Remote Sens. 11 (11), 1378.
Li, H.Y., Marchant, B.P., Webster, R., 2016. Modelling the electrical conductivity of soil in the Yangtze delta in three dimensions. Geoderma 269, 119–125.

Li, Q., Wang, Z., S., Wei, Li, L., Yao, Y., Yu, F., 2021. Improved daily SMAP satellite soil moisture prediction over China using deep learning model with transfer learning. J. Hydrol. 600, 126698.
Ließ, M., 2015. Sampling for regression-based digital soil mapping: Closing the gap between statistical desires and operational applicability. Spatial Stat. 13, 106–122.
Magdoff, F., van Es, H., 2021. Building Soils for Better Crops, 4 Sustainable Agriculture Research and Education, College Park.
Marchant, B.P., McBratney, A.B., Lark, R.M., Minasny, B., 2013. Optimized multi-phase sampling for soil remediation surveys. Spatial Stat. 4, 1–13.
Matheron, G., 1969. Le Krigeage Universel. Ecole des Mines de Paris, Fontainbleau.
McBratney, A.B., Webster, R., 1981. Detection of ridge and furrow pattern by spectral analysis of crop yield. Int. Stat. Rev. 49 (1), 45–52.
McBratney, A.B., Webster, R., 1983. Optimal interpolation and isarithmic mapping of soil properties. V. Co-regionalization and multiple sampling strategy. J. Soil. Sci. 34 (1), 137–162.
McBratney, A.B., Webster, R., 1986. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. J. Soil. Sci. 37 (4), 617–639.
Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.
Mercer, W.B., Hall, A.D., 1911. Experimental error of field trials. J. Agric. Sci. 4, 107–132.
Metcalfe, H., Milne, A.E., Webster, R., Lark, R.M., Murdoch, A.J., Storkey, J., 2016. Designing a sampling scheme to reveal correlations between weeds and soil properties at multiple spatial scales. Weed Res. 56 (1), 1–13.
Milne, A.E., Glendining, M.J., Lark, R.M., Perryman, S.A. M., Gordon, T., Whitmore, A.P., 2015. Communicating the uncertainty in estimated greenhouse gas emissions from agriculture. J. Environ. Manag. 160, 139–153.
Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264 (B, SI), 301–311.
Morse, R.K., Thornburn, T.H., 1961. Reliability of soil maps. In: Proceedings of the 5th International Conference of Soil Mechanics and Foundation Engineering. Dunod, Paris, pp. 259–262.
Nikparvar, B., Thill, J.-C., 2021. Machine learning of spatial data. ISPRS Int. J. Geogr. Inf. Sci. 10 (9), 600.
Oliver, M.A, 2010. Geostatistical Applications for Precision Agriculture. Springer, Dordrecht.
Oliver, M.A., Webster, R., 1989. A geostatistical basis for spatial weighting in multivariate classification. Math. Geol. 21 (1), 15–35.
Orton, T.G., Pringle, M.J., Bishop, T.F.A., 2016. A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals. Geoderma 262, 174–186.
Panagopoulos, T., Jesus, J., Antunes, M.D.C., Beltrao, J., 2006. Analysis of spatial interpolation for optimising management of a salinized field cultivated with lettuce. Eur. J. Agron. 24 (1), 1–10.
Piikki, K., Wetterlind, J., Söderström, M., Stenberg, B., 2021. Perspectives on validation in digital soil mapping of continuous attributes – a review. Soil Use Manag. 37 (1), 7–21.
Plaza, C., Zaccone, C., Sawicka, K., Mendez, A.M., Tarquis, A., Gasco, G., Heuvelink, G.B.M., Schuur, E.A.G., Maestre, F.T., 2018. Soil resources and element stocks in drylands to face global issues. Sci. Rep. 8, 13788.
Ploton, P., Mortier, F., Rejou-Mechain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., V., G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pelissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 11 (1), 4540.
Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B. M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240.
Sekulić, A., Kilibarda, M., Heuvelink, G.B.M., Nikolić, M., Bajat, B., 2020. Random forest spatial interpolation. Remote Sens. 12 (10), 1687.
Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Netw. 19 (2), 225–235.
Spiegelhalter, D., 2019. The Art of Statistics: Learning from Data. Penguin Random House, UK.
Steinbuch, L., Brus, D.J., Heuvelink, G.B.M., 2022. Mapping depth to pleistocene sand with Bayesian generalized linear geostatistical models. Eur. J. Soil. Sci. 73, e13140.
Sungmin, O., Orth, R., 2021. Global soil moisture data derived through machine learning trained with in-situ measurements. Sci. Data 8 (1), 170.
Szatmári, G., Pásztor, L., 2019. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma 337, 1329–1340.
Szatmári, G., Pásztor, L., Heuvelink, G.B.M., 2021. Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. Geoderma 403, 115356.
Tobler, W.R., 1970. Computer movie simulating urban growth in Detroit region. Econ. Geograp. 46 (2), 234–240.
Van Groenigen, J.W., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. J. Environ. Qual. 27 (5), 1078–1086.
Van Leeuwen, C.E., Mulder, V.L., Batjes, N.H., Heuvelink, G.B.M., 2022. Statistical modelling of measurement error in wet chemistry soil data. Eur. J. Soil. Sci. 73, e13137.
Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.
Veronesi, F., Schillaci, C., 2019. Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. Ecol. Indic. 101, 1032–1044.
Wadoux, A.M.J.-C., Brus, D.J., Heuvelink, G.B.M., 2018. Accounting for non-stationary variance in geostatistical mapping of soil properties. Geoderma 324, 138–147.

Wadoux, A.M.J.-C., Brus, D.J., Heuvelink, G.B.M., 2019a. Sampling design optimization for soil mapping with random forest. Geoderma 355, 113913.

Wadoux, A.M.J-C., Heuvelink, G.B.M., de Bruin, S., Brus, D.J., 2021a. Spatial cross-validation is not the right way to evaluate map accuracy. Ecol. Model. 457, 109692.

Wadoux, A.M.J-C., Heuvelink, G.B.M., Lark, R.M., Lagacherie, P., Bouma, J., Mulder, V.L., Libohova, Z., Yang, L., McBratney, A.B., 2021b. Ten challenges for the future of pedometrics. Geoderma 401, 115155.

Wadoux, A.M.J.-C., Padarian, J., Minasny, B., 2019b. Multi-source data integration for soil mapping using deep learning. Soil 5, 107–119.

Wadoux, A.M.J-C., Samuel-Rosa, A., Poggio, L., Mulder, V.L., 2020. A note on knowledge discovery and machine learning in digital soil mapping. Eur J. Soil. Sci. 71, 133–136.

Webster, R., 2015. Technological developments for spatial prediction of soil properties, and Danie Krige's influence on it. J. South. Afr. Inst. Min. Metallurgy 115, 165–172.

Webster, R., 2019. Burgess, T.M. & Webster, R. 1980. Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. Journal of Soil Science, 31, 315–331. Eur. J. Soil Sci. 70, 2–6.

Webster, R., Burgess, T.M., 1980. Optimal interpolation and isarithmic mapping of soil properties. III. Changing drift and universal kriging. J. Soil. Sci. 31 (3), 505–524.

Webster, R., Cuanalo de la C., H.E., 1975. Soil transect correlograms of North Oxfordshire and their interpretation. J. Soil Sci. 26 (2), 176–194.

Webster, R., Oliver, M.A., 1989. Optimal interpolation and isarithmic mapping of soil properties. VI. Disjunctive kriging and mapping the conditional probability. J. Soil. Sci. 40 (3), 497–512.

Van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D.P., Poggio, L., 2022. Measurement error-filtered machine learning in digital soil mapping. Spatial Stat. 47, 100572.

Ye, H., Huang, W., Huang, S., Huang, Y., Zhang, S., Dong, Y., Chen, P., 2017. Effects of different sampling densities on geographically weighted regression kriging for predicting soil organic carbon. Spatial Stat. 20, 76–91.

Yoo, E.H., Kerry, R., Ingram, B., Ortiz, B., Scully, B., 2018. Defining and characterizing aflatoxin contamination risk areas for corn in Georgia, USA: Adjusting for collinearity and spatial correlation. Spatial Stat. 28, 84–104.

Young, A., 2017. Thin on the Ground: Soil Science in the Tropics, 2 Land Resource Books, Norwich.

Zheng, H., Xu, W., Zhang, S., Jin, S., Zhang, F., Jia, J., 2021. Uncertainty analysis of human health risk assessment of soils contaminated with high concentrations of arsenic in a coal chemical industry area in northwest China. Water Air Soil Pollut 232 (11), 458.