Geoderma 383 (2021) 114725

Contents lists available at ScienceDirect

Geoderma

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

Hypotheses, machine learning and soil mapping

Alexandre M.J.-C. Wadoux[a,*], Alex B. McBratney[a]

[a] Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, New South Wales, Australia

ARTICLE INFO

Handling Editor: Kristin Piikki

Keywords:
Soil science
Data science
Theory
Epistemology
Knowledge discovery
Pedology

ABSTRACT

Hypotheses are of major importance in scientific research. In current applications of machine learning algorithms for soil mapping the hypotheses being tested or developed are often ambiguous or undefined. Mapping soil properties or classes, however, does not tell much about the dynamics and processes that underly soil genesis and evolution. When the interest in the soil map is for applications in a context different than soil science, such as for policy making or baseline production of quantitative soil information, the interpretation should be made in light of this application. If otherwise, we recommend soil scientists to provide hypotheses to accompany their research. The hypothesis is formulated at the beginning of the research and, in some cases, motivates data collection. Here we argue that when applying data-driven techniques such as machine learning, developing hypotheses can be a useful end point of the research. The spatial pattern predicted by the machine learning model and the correlation found among the covariates are an opportunity to develop hypotheses which are likely to require additional analyses and datasets to be tested. Systematically providing scientific hypotheses in digital soil mapping studies will enable the soil science community to build on previous work, and to increase the credibility of data-driven algorithms as a means to accelerate discovery on soil processes.

In recent years, there has been an increasing number of publications using data-driven, empirical algorithms for digital soil mapping (DSM, Lagacherie et al., 2006; Ma et al., 2019). In particular, machine learning (ML) algorithms are popular for mapping soil properties or classes using soil point information and a large number or environmental covariates (Walter et al., 2006; Hengl et al., 2018). Typically these studies assess and quantify the spatial variability of the soil using one or several ML algorithms. These studies are performed with the explicit purpose of evaluating spatial variation from a set of observations. This hinges upon an empirical discovery of the relationships among covariates and a set of observations, whereby the discovery is driven by the ML algorithm and fundamentally relies on pattern recognition (Pennock, 2004).

Research in soil science traditionally relies on hypothesis testing or development, in a deductive or inductive approach. Deduction is a syllogistic logic, from the general (the theory) to the specific (the observations). The scientist holds a theory from an educated guess. From this theory a hypothesis is formulated, and confrontation against observations is made to corroborate or refute it. For example:

> All soil properties vary in the geographic space.
> Soil organic carbon (SOC) is a soil property.
> We deduct that SOC concentration varies in space.

Inductive reasoning, conversely, goes from the specific to the general by inferring a theory or an explanation from the observations, theory or explanation from which predictions are made to new observations. Thus:

> In alkaline soils, SOC is stabilized through interaction with calcium.
> Solonetz are alkaline soils.
> We predict by induction that SOC in Solonetz is stabilized through interaction with calcium.

In practice, there is a constant interplay between deduction and induction, because a hypothesis derived by induction can be tested by deduction. In soil science, both approaches are operating. One may start with observations and develop possible explanations (induction) while others may have a possible explanation for a phenomenon and test it using a controlled experiment or by collecting new data (deduction).

In current use of ML for DSM, however, it is not clear which approach is taken since hypotheses are not developed nor tested. A map of the topsoil organic carbon might reveal patterns of location-specific concentration in an area, but the fundamental problems of, for example, what are physical and chemical stabilization mechanisms, why pedoclimatic conditions impact SOC protection and how the organic material interacts with other soil elements, remain beyond questions.

* Corresponding author.
E-mail address: alexandre.wadoux@sydney.edu.au (A.M.J.-C. Wadoux).

https://doi.org/10.1016/j.geoderma.2020.114725
Received 1 September 2020; Accepted 5 September 2020
Available online 19 September 2020

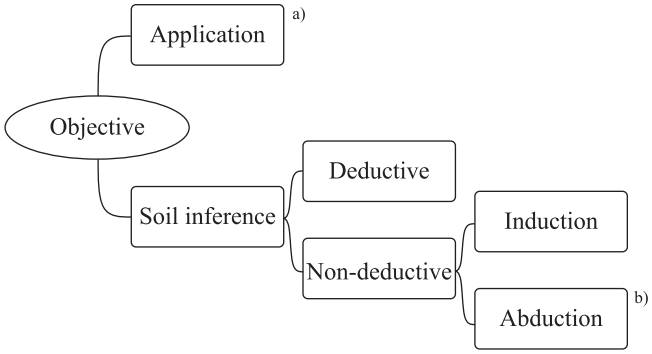

Fig. 1. Illustration of the path followed by the scientist using machine learning algorithm for digital soil mapping. The scientist must decide whether the objective of the study is a) an application, or b) for scientific purpose by abductive reasoning. In a) the map is for application in a context different than soil science while in b) the objective is the development of plausible hypotheses from the predicted pattern and the correlations found in the data by the machine learning model.

Predicting a pattern is different from providing an explanation (Kitchin, 2014). Explanation requires contextual and pedological knowledge on the interaction of the soil property with the environmental covariates. The danger in ML for DSM is to have research driven by the technique (an algorithm is available) or by the data availability rather than by a scientific question. In fact, most authors on studies on ML for DSM refrain from formulating a hypothesis. More worryingly, they consider the predicted map an end in itself.

In some specific cases, the soil map is of special interest when considered in a context other than soil science. For land planning, the map suggests which management practices are the most appropriate for field-specific soil conditions. In environmental protection, the soil map advises stakeholders and drives policy making. If the production of quantitative soil information serves for another purpose than increasing the understanding of soil, this is yet rarely, if ever, made clear in scientific publications. We argue that, in this case, the study should focus on describing the context with examples on the application of the map in that context. For example, maps of the temporal evolution of the SOC may well be confronted to a baseline SOC map (e.g. the coarse default IPCC Tier 1 global map, as is done in Heuvelink et al. (2020)), with examples of local actions to be taken for scenarios of SOC concentration change over time.

When the objective is to produce scientific knowledge and understand, however, providing a map is not sufficient. The question is then the type of scientific reasoning that is adopted (Fig. 1). ML algorithms are empirical, which supposes an inductive reasoning (i.e. from the data to the theory). Because of their complexity, however, calibrated ML models preclude analysis of their structure and the process that underlies the prediction. This hinders inductive reasoning because no theory can be readily extracted from a calibrated ML model. We argue that the reasoning behind data-driven algorithms relies more on abduction (Peirce, 1960), which is a similar, but weaker form of inference compared to induction. While induction relies on the data to infer a theory, abduction relies on data to infer a possible explanation for a phenomenon (Miller and Goodchild, 2015). Thus abduction begins with data accumulation independently of any surmise. A data-driven algorithm then interrogates the data to tease subtle correlations that are often inherently invisible to the human because of the multivariate and non-linear nature of the data. The discovered pattern might ultimately serve for the development of new hypotheses (Fayyad et al., 1996; Hazen, 2014).

In practical terms this means that the digital soil map should not be the end, but rather the starting point of the analysis. The revealed pattern and calibrated ML model should trigger questions which will certainly require further analysis and data to be answered. In fact, there is already a hypothesis behind the model construction. The hypothesis is that soil spatial variability is driven by a set of environmental variables representing soil formation. The veracity of this hypothesis is offset by the problem of the spatial scale. The drivers of some processes change with scale, and, for example with SOC, we may not be able to identify from the predicted map the mechanisms of stabilization (e.g. we may not have input information on the controlling factors at profile scale, like micro-climate or metal oxides). On a number of occasions, formulating a hypothesis at the beginning of the study (deductive reasoning) is not even possible because the scientist uses legacy data which were originally collected for purposes other than the present mapping exercise, and in fact often with an unknown (to the mapper) purpose. This obliges the researcher to adopt abductive reasoning for scientific knowledge production, where first the ML algorithm is used to find patterns and predict from the store of data and secondly, hypotheses are developed from the fitted relationships between the point soil information and the environmental covariates, by connecting patterns to possible processes. For example, more SOC in clayey areas or in natural vegetation vs. less SOC in cultivated areas triggers some assumptions and hypotheses on the mechanisms behind SOC stabilization. In this sense, ML is used as a "hypothesis discovery" tool (Wadoux et al., 2020).

We conclude on the importance that studies using ML for DSM clarify their objective. When the objective is to produce quantitative soil information, the context and applicability of the map in this context should be then elucidated. When, conversely, the objective is to increase our scientific understanding of soil formation and genesis, hypotheses should be proposed. When using data-driven algorithms such as ML, it is sensible to let the algorithm find the correlation in the data, and to analyse the pattern with the ambition to develop potential questions worthy of further investigation. In this sense, one would take full advantage of the data-driven algorithm which searches for pattern in datasets and can reveal previously undetected correlations. In short, let the algorithm find the pattern and the soil scientist the hypotheses that follow.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Magazine 17, 37.
Hazen, R.M., 2014. Data-driven abductive discovery in mineralogy. Am. Mineral. 88, 2165–2170.
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., Angelini, M.E., Poggio, L., Bai, Z., Batjes, N.H., van den Bosch, H., Bossio, D., Estella, S., Lehmann, J., Olmedo, G.F.. et al. 2020. Machine learning in space and time for modelling soil organic carbon change. European J. Soil Sci., In Press.
Kitchin, R., 2014. The Data Revolution: Big Data, Open Data, Data Infrastructures And Their Consequences. SAGE Publications UK, London, England.
Lagacherie, P., McBratney, A.B., 2006. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An Introductory Perspective. Elsevier, Amsterdam, the Netherlands, pp. 3–22.
Ma, Y., Minasny, B., Malone, B.P., McBratney, A.B., 2019. Pedology and digital soil mapping (DSM). European J. Soil Sci. 70, 216–235.
Miller, H.J., Goodchild, M.F., 2015. Data-driven geography. GeoJournal 80, 449–461.
Peirce, C.S., 1960. Collected papers of Charles Sanders Peirce, vol. 2 Harvard University Press, Cambridge, USA.
Pennock, D., 2004. Designing field studies in soil science. Canadian J. Soil Sci. 84, 1–10.
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. European J. Soil Sci. 71, 133–136.
Walter, C., Lagacherie, P., Follain, S., 2006. Integrating pedological knowledge into digital soil mapping. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An Introductory Perspective. Elsevier, Amsterdam, the Netherlands, pp. 281–615.