Received: 18 June 2019 | Revised: 17 September 2019 | Accepted: 8 November 2019

DOI: 10.1111/ejss.12909

SHORT COMMUNICATION

European Journal of Soil Science WILEY

A note on knowledge discovery and machine learning in digital soil mapping

Alexandre M. J.-C. Wadoux[1] | **Alessandro Samuel-Rosa**[2] | **Laura Poggio**[3] | **Vera Leatitia Mulder**[1]

[1]Soil Geography and Landscape Group, Wageningen University and Research, Wageningen, the Netherlands

[2]Department of Agronomy, Universidade Tecnológica Federal do Paraná, Rebouças, Brazil

[3]ISRIC, World Soil Information, Wageningen, the Netherlands

Correspondence
Alexandre M. J.-C. Wadoux, Soil Geography and Landscape Group, Wageningen University, Droevendaalsesteeg 3, 6708 BP Wageningen, the Netherlands.
Email: alexandre.wadoux@wur.nl

Funding information
No funding has been received to carry out this study

Abstract

In digital soil mapping, machine learning (ML) techniques are being used to infer a relationship between a soil property and the covariates. The information derived from this process is often translated into pedological knowledge. This mechanism is referred to as knowledge discovery. This study shows that knowledge discovery based on ML must be treated with caution. We show how pseudo-covariates can be used to accurately predict soil organic carbon in a hypothetical case study. We demonstrate that ML methods can find relevant patterns even when the covariates are meaningless and not related to soil-forming factors and processes. We argue that pattern recognition for prediction should not be equated with knowledge discovery. Knowledge discovery requires more than the recognition of patterns and successful prediction. It requires the pre-selection and preprocessing of pedologically relevant environmental covariates and the posterior interpretation and evaluation of the recognized patterns. We argue that important ML covariates could serve the purpose of providing elements to postulate hypotheses about soil processes that, once validated through experiments, could result in new pedological knowledge.

Highlights

- We discuss the rationale of knowledge discovery based on the most important machine learning covariates
- We use pseudo-covariates to predict topsoil organic carbon with random forest
- Soil organic carbon was accurately predicted in a hypothetical case study
- Pattern recognition by random forest should not be equated to knowledge discovery

KEYWORDS

mapping, pedometrics, random forest, soil science, variable selection

Machine learning (ML) techniques are popular for mapping soil properties using numerous spatially explicit covariates and a set of point-measured values of the property of interest (Behrens et al., 2018; Hengl., Nussbaum, Wright, Heuvelink, & Gräler, 2018). The use of environmental information to map soil properties has also been the basis of traditional soil surveys and digital soil mapping (DSM) exercises using more traditional, geostatistical

methods (McBratney, Santos, & Minasny, 2003). More recently, the conceptional framework for soil mapping has been linked to ML (ML-DSM) with the underlying assumption that the ML model builds decision rules in a similar way as the soil surveyor does. As such, ML models have often been used for soil knowledge discovery. That is, to infer causal relationships between soil properties and forming factors and processes from the association of the former with the covariates (e.g. in Bui, Henderson, & Viergever, 2006; Guevara et al., 2018; Hengl et al., 2017; Ma, Minasny, Malone, & Mcbratney, 2019; Wiesmeier, Barthold, Blank, & Kögel-Knabner 2010; Wilford & Thomas, 2013). However, there are reasons to question the validity of our common practice of soil knowledge discovery when using ML, as previously demonstrated by Shmueli (2010) and Fourcade, Besnard, and Secondi (2018). For instance, tree-based methods, such as random forest, seek a statistical optimum following a hierarchical partitioning of data rather than accounting for soil processes and causalities in the system. Consequently, with a well-calibrated model and a sufficiently large number of covariates the resulting map can be accurate based on validation metrics, such as concordance correlation coefficient or mean squared error. This means that we can produce soil maps that can be easily reproduced, updated and validated. However, the rationale of pedological knowledge discovery based on ML is to be treated with caution, as is demonstrated in this study.

In recent years, large and abundant datasets are becoming available to scientists. Models can be applied to establish empirical relationships among the data, in what is often called "data science". This raised a number of issues yet to be fully tackled, such as the multiple interactions of many related variables, models not fully specified (i.e. not all factors are accounted for), difficulties in identifying causality from empirical data, and challenges around data exploration and understanding when including disciplinary knowledge (Blei & Smyth, 2017). This is also fundamentally different from traditional empirical scientific development where data were collected to answer a formulated hypothesis.

In this paper we investigate whether the (complete lack of) knowledge of soil-forming factors and processes influences ML-DSM performance. We do this by testing whether pseudo-covariates can be used to accurately predict a soil property. These pseudo-covariates do not represent any soil-forming factors, nor are they related to soil-forming processes.

We use the random forest (RF) model. RF is an ensemble tree-based ML technique widely used in DSM. An ensemble of trees is built based on a bootstrap sample of the training data. All tree predictions are averaged, and these averages are taken as the final predictions. The RF algorithm introduces an additional random perturbation to reduce the chance of overfitting during the splitting of a tree, by selecting a reduced subset of covariates at each split. The RF algorithm relies on three user-defined parameters, the number of trees, the number of covariates selected at each split and the size of terminal nodes. For more details on the parameters, we refer to Wadoux, Brus, and Heuvelink (2019, sec. 2.2). In this study we implemented the default RF model from the R package ranger (Wright & Ziegler, 2017) using fine-tuned parameter values. Prediction accuracy was assessed by the concordance correlation coefficient (CCC) and the root mean square error (RMSE), derived using an independent validation set. The bias was assessed by the mean error (ME). The most important predictors in the RF model were identified using the mean decrease in the variance of the response as a variable importance measure (Wright & Ziegler, 2017).

We tested the methodology on a hypothetical case study. The test area was defined as the Pangaea continent, in the shape of 250 million years ago. The soil property of interest is the (pseudo) topsoil organic carbon (SOC, g kg^{-1}). These values were taken from existing data: a 200 km × 200 km SoilGrids (Hengl et al., 2017) tile, located between longitude 17 to 22 and latitude 52 to 54 degrees. The SOC tile was considered as our variable of interest and cropped to the extent of the Pangaea. Two sets of 500 and 1,000 points were selected by simple random sampling for calibration and validation, respectively. In total, 41 pseudo-covariates were used as predictors in the RF model. They were created using publicly available pictures of pedometricians' heads, shoulders and upper chests. The three colour channels (red, blue and green) were reduced to their first principal component (PC). The first PCs of all pictures were geo-referenced, cropped to the extent of the Pangaea and resampled to match the resolution of the SoilGrids tile. An example set of the first PCs for eight pictures used as pseudo-covariates is shown in Figure 1.

The results show that RF with the pseudo-covariates predicts SOC accurately. The CCC on the calibration set is 0.65, whereas the RMSE is 19.68 g kg^{-1}. There is a negligible bias (ME = 0.48 g kg^{-1}). The predictions yielded satisfactory prediction accuracy on the validation set, as shown by a CCC of 0.67, a RMSE of 19.77 g kg^{-1} and a bias close to 0 (ME = −0.01 g kg^{-1}). This also shows that the RF model was not overfitted. Figure 2 shows the 20 most important predictors to the RF model. Pseudo-covariate Alex McBratney was the most important predictor, with a variance of the response value of 21,645 g kg^{-2}. The mean decrease in the variance of the response steadily decreases for the other covariates, until reaching a value of 4,775 g kg^{-2} for the least contributing predictor (i.e., Georges Matheron, not shown).

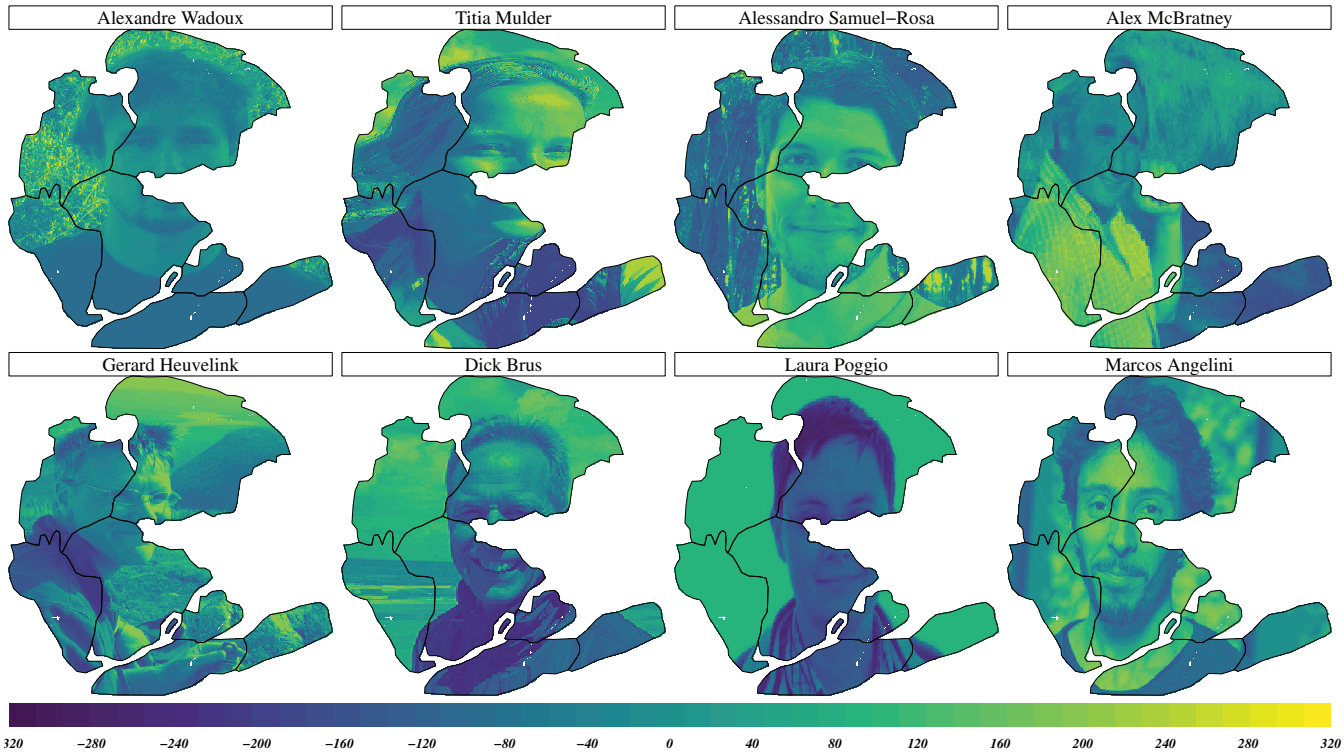

FIGURE 1 Example set of eight (out of 41) first principal components (PCs) of pedometricians' head, shoulders and upper chest images used as pseudo-covariates in the RF model

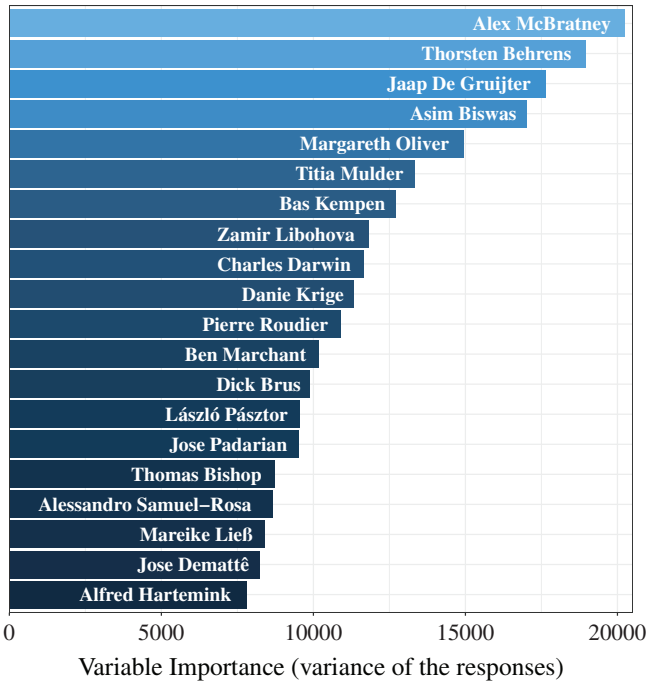

FIGURE 2 Set of 20 most important predictors of the random forest (RF) model

The results show that using pseudo-covariates without pedological meaning as predictors of SOC in a ML model produces maps having a similar accuracy compared to existing ML-DSM. This map was produced using an ML model that recognizes relevant patterns in the data. Fourcade et al. (2018) noted that any image containing a spatial pattern can be fitted with a ML model to the values of a dependent variable (SOC in this study). Once the numerical rules of the ML model have been established, they can be applied to new locations to predict soil properties. This shows that the rationale of knowledge discovery based on ML is to be treated with caution, as an accurate soil map can be produced without any pedological knowledge.

Variable importance measures are often used to draw causal conclusions about soil-forming factors and processes, i.e. for knowledge discovery. However, we can rarely attest to the validity of these conclusions because we do not know the true soil-forming factors and processes. This is nothing new, as Jenny already discussed the problem of causality in his book "Factors of soil formation". Jenny (1941, p. 118) gave the example of nitrogen and organic carbon which vary according to soil moisture. He stressed that while certain properties vary together, they do not necessarily relate to explaining processes of soil formation. This logic applies to ML-DSM studies where we only have access to the covariates, that are simple surrogates for the forming factors in the empirical ML equations. We argue that care should be taken when drawing causal conclusions from the variable importance measure, because strong assumptions based

on proxy information are implicitly made. Therefore we argue that it more sound to use ML and its variable importance measures to postulate hypotheses about forming factors and processes (see also Ma et al. (2019)). New causal pedological knowledge could then be discovered by testing these hypotheses using properly designed experiments and principles of soil genesis. In this context, we consider that ML in DSM would be more of a "hypothesis", rather than a "knowledge" discovery tool.

Bui et al. (2006) stated that "the fact that the decision tree models can be used to make extensive maps of soil properties demonstrates that knowledge discovery from the soil-landscape databases has occurred". We argue that the ability to produce maps is due to the fact that the ML algorithm used was able to recognize relevant patterns in the database even when the covariates are unrelated to soil-forming factors and processes. Efficient models (prediction wise) can be created using meaningless predictors, but only models using a hypothesis (e.g. *scorpan* (McBratney et al., 2003)) behind the model construction should be used. Therefore the covariates selected for DSM models should always represent interpretable factors that are related to the soil-forming factors and all interpretation should still be carried out with extreme caution. Pattern recognition should not be equated with knowledge discovery because knowledge discovery requires the interpretation and evaluation of the recognized patterns (Gullo, 2015). Thus, as long as we are concerned only with the production of spatial soil information i.e. soil maps via pattern recognition, using ML in DSM does not necessarily serve the purpose of discovering pedological knowledge.

DATA AVAILABILITY STATEMENT

Data and code are accessible upon request to the corresponding author.

ORCID

Alexandre M. J.-C. Wadoux https://orcid.org/0000-0001-7325-9716
Alessandro Samuel-Rosa https://orcid.org/0000-0003-0877-1320

REFERENCES

Behrens, T., Schmidt, K., Rossel, R. A. V., Gries, P., Scholten, T., & MacMillan, R. A. (2018). Spatial modelling with Euclidean distance fields and machine learning. *European Journal of Soil Science*, *69*, 757–770.

Blei, D. M., & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 8689–8692.

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

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*, 245–256.

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Duarte, Y. A., Hernández, C. A., ... Vargas, R. (2018). No silver bullet for digital soil mapping: Country-specific soil organic carbon estimates across Latin America. *Soil*, *4*, 173–193.

Gullo, F. (2015). From patterns in data to knowledge discovery: What data mining can do. *Physics Procedia*, *62*, 18–22.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., ... Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*, e0169748.

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.

Jenny, H. (1941). *Factors of soil formation: A system of quantitative pedology*. New York, NY: McGrawHill, USA.

Ma, Y., Minasny, B., Malone, B. P., & Mcbratney, A. B. (2019). Pedology and digital soil mapping (DSM). *European Journal of Soil Science*, *70*, 216–235.

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

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.

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

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2010). Digital mapping of soil organic matter stocks using random forest modeling in a semi-arid steppe ecosystem. *Plant and Soil*, *340*, 7–24.

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

Wright, M. N., Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*, 1–17.

How to cite this article: Wadoux A-C, Samuel-Rosa A, Poggio L, Mulder VL. A note on knowledge discovery and machine learning in digital soil mapping. *Eur J Soil Sci*. 2020;71: 133–136. https://doi.org/10.1111/ejss.12909