European Journal of Soil Science

WILEY

OPINION

Can Digital Soil Mapping Be Causal?

Lei Zhang[1] | Alexandre M. J.-C. Wadoux[2]

[1]Climate and Ecosystem Science Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA | [2]College of Science and Engineering, James Cook University, Cairns, Queensland, Australia

Correspondence: Lei Zhang (lei.zhang@lbl.gov)

Received: 30 September 2025 | **Revised:** 5 December 2025 | **Accepted:** 11 January 2026

Keywords: causality | correlation | knowledge discovery | machine learning | pedometrics | spatial variation

ABSTRACT

All too often, it is unclear whether digital soil mapping (DSM) models can support causal interpretation. A common practice in DSM studies is to interpret the importance of covariates for prediction. This carries an implicit causal assumption that is rarely stated and even more rarely justified. Because DSM relies entirely on observational data, it is widely assumed that causal inference is not possible. But is it? Here, we discuss the conditions under which causal inference with observational data is possible and two views of causality. We show that while under each of the views causal inference may be possible, a so-called generative view is the one most capable of satisfying the conditions for causal inference in DSM. Generative causality treats causation as the system of processes that produce observed associations, rather than relying on associations themselves, as is common in current DSM studies. Realizing this perspective requires DSM to shift towards models in which soil-forming factors influence soil properties through explicitly modelled processes, which some would call process-informed DSM. Since these processes are 'fully determined' by the modeller's specification, they offer a structured means to control confounding and open the door to applying existing causal inference frameworks. While generative DSM is formally possible, we should ultimately ask whether causal inference ought to be a primary goal, since the primary strength of DSM lies not in establishing causality but in delivering accurate predictions and highlighting patterns that warrant further investigation.

1 | Introduction

Digital soil mapping is the practice of predicting soil properties or classes at places where they have not been measured. This is commonly done with statistical modelling of the spatial correlation between locations, as is done in geostatistical modelling and prediction with kriging, and by relating those properties to spatially exhaustive covariates through statistical or machine-learning models. The covariates may include elevation, vegetation, climate and others that serve as proxies for the factors of soil formation: parent material, relief, climate, organisms and time. The idea is not new. More than 80 years ago, Jenny (1941) argued that particular combinations of these factors give rise to processes creating a particular soil. This was expressed through a relationship in the form of a deterministic regression.

Without moving away from associational structure between covariates and soil properties, pedometrics has long recognized that soil is too complex to be modelled in a fully deterministic way. We cannot describe all the factors, nor all the processes, their interactions and their changes through time. While in principle soil formation is deterministic, in practice we must regard at least part of the soil variation as if it were the outcome of a random process (Webster 2000). This variation is not necessarily unstructured and information can be extracted with an appropriate statistical model. This practice aligns with Hempel (1965)'s notion of statistical-inductive explanation: natural laws are deterministic, yet when initial conditions cannot be fully specified, explanations and predictions can be expressed probabilistically (i.e., using a statistical-inductive explanation, see Wadoux et al. 2021).

Highlights

- Implicit causal assumption is usually made in digital soil mapping.
- Digital soil mapping can support causal inference if some conditions are met.
- A generative view is the one most capable of satisfying the conditions for causal inference in DSM.
- Strength of digital soil mapping lies primarily in its predictive capacity.

To this end, a common procedure in DSM is to find the model and association with covariates that best predict the soil property of interest, as judged by some measure of fit. When the model is validated, that is, when it is well explained by probabilities under statistical laws, we then look for a physical cause of its success. In practice this means interpreting the most important covariates. For example, Gomes et al. (2025) mapped potential soil water repellency of Danish topsoil with a machine learning model fitted to 7500 measurements. The accurate predictions, as assessed by cross-validation, led the authors to examine the effect that the covariates, such as some soil properties and remote sensing indices, had on the outcomes using variable importance metrics. While in DSM we have long cautioned against interpreting correlations found in the model as causal (Wadoux et al. 2020), there is, in many studies, an implicit causal assumption that is rarely stated yet justified in the analysis. All too often, we actually do not know if the DSM model can support causal interpretation.

Hereafter, we examine current DSM practices based on observational data and outline the conditions under which causal inference might be possible. We then situate these current practices within a successionist view of causality and contrast this with a generative view. Both perspectives are then explored to clarify their implications for making causal inference in DSM studies.

2 | Observational Data in Digital Soil Mapping

DSM relies almost exclusively on observational data, that is, data collected from field surveys and soil inventories, rather than from controlled experiments. It is widely accepted that observational data alone do not support causal inference (Yuan et al. 2017; Byrnes and Dee 2025). Confounding factors may influence both covariates and soil properties, and the timing or sequence of events is unknown. Further, soil measurements and observations capture only a snapshot emerging from a complex, dynamic system, without clear temporal sequencing. This makes it impossible to understand which variable precedes the other. If we apply this logic, no DSM study can be causal.

Take as an example Figure 1 which illustrates some relationships found in a global soil dataset (Batjes et al. 2024). We examined temperature (T), vegetation (V) and soil organic carbon (S; SOC). While V is positively correlated with T (Figure 1a), and S is positively correlated with V (Figure 1b), the direct relationship between T and S can appear negative globally (Figure 1c), while positive and negative correlations also appear depending on which variables or regions are considered (Figure 1d,e). Controlling for V reveals a

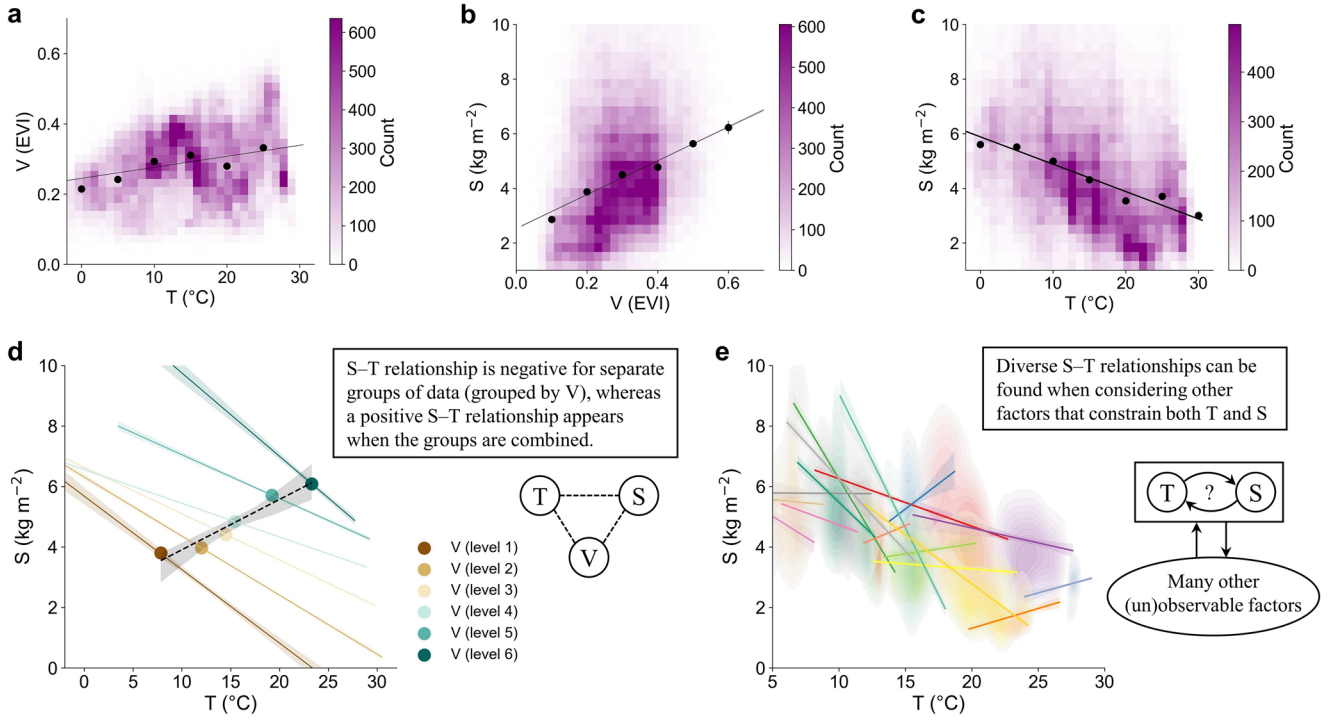

FIGURE 1 | (a–c) Relationships among T (temperature), V (enhanced vegetation index, EVI) and S (soil organic carbon stock). (d) The T–S relationship reverses: Within each V level (six groups from 0 to 0.6 with an interval of 0.1), it is negative, but averaging across groups produces a positive trend, illustrating Simpson's paradox. (e) T–S relationships vary widely across ecoregions (each line shows the regression for one ecological region).

negative T–S relationship within each subset, but aggregating the data shows a positive trend, a clear example of Simpson's paradox (Figure 1d). Stratifying by ecoregion produces yet more diverse T–S patterns (Figure 1e). These results show that observed associations are confounded by unmeasured factors and are influenced by data partitioning. The variability of statistical relationships indicates the risk of misjudgement of true causal relationships can be hard to spot. Associations from observational data can mislead, and the current model based on correlation found in the data does help for pattern detection, but not causal inference.

Several people have claimed strategies that enable causal inference from observational data, with the seminal framework of Pearl (2009) being among the most influential. Although the terminology varies across frameworks, they generally agree on three core conditions. First, one must articulate a causal model that specifies how variables are thought to influence one another. This is typically expressed using a directed acyclic graph (DAG), which makes explicit the causal pathways. Second, the model must satisfy the assumption of causal sufficiency (i.e., we ensure no unobserved confounders): all common causes of the exposure and the outcome must be observed and included, so that spurious associations can be blocked. Third, there must be compatibility between the structural model and observed data, often referred to as the faithfulness or stability assumption. This means that any statistical independencies observed in the data must correspond to the independencies implied by the causal graph. These conditions are conceptually distinct: the first concerns the formulation of assumptions, the second concerns the measured variables and the third concerns the consistency between the causal structure and the data. They jointly determine whether causal effects are identifiable from observational data using tools such as adjustment criteria or do-calculus.

Having the possibility for causal inference with observational data is particularly attractive for DSM because it allows us to ask questions of the form: *what happens on Y (a soil property) if we change X (environmental covariates)?* without the requirement for controlled experiments. For example, a plausible causal pathways for SOC stock could be defined as:

$$\begin{array}{ccccccc} \textit{air temperature} & \longrightarrow & \textit{vegetation} & \longrightarrow & \textit{organic inputs} & \longrightarrow & \textit{microbial processing} \longrightarrow \textit{SOC} \\ & \searrow & & & & \nearrow & \\ & & & \textit{soil temperature, moisture} & & & \end{array} \quad (1)$$

These pathways, along with others not shown here, are grounded in theory and pedological knowledge rather than mere correlation. They provide a structured way to reason about interventions, such as changing vegetation cover, managing crop residues differently or altering temperature and its effect on microbial respiration, and assessing their consequences for SOC. Such pathways have been abundantly discussed in the pedology and biogeochemistry literature, yet they are seldom operationalized in DSM. The reasons are clear: the conditions required for causal inference are difficult to satisfy in typical DSM settings.

In the context of DSM, these conditions can typically be satisfied only partially. Domain knowledge does support the formulation of plausible causal pathways among soil-forming factors and soil properties, and confounding can sometimes be mitigated through data stratification or by imposing process-based constraints. However, causal sufficiency is seldom achievable: many relevant drivers (e.g., historical land-use, biological processes subsurface conditions) remain unmeasured or are only crudely proxied by available covariates. Faithfulness is difficult to satisfy in DSM because any causal representation is scale-specific, while many covariates are obtained at scales incompatible with this representation, and further are measured imprecisely. This makes the independencies observed in DSM datasets difficult to align with those implied by any assumed causal model. Pearl's framework gives the formal 'how', that is, how to encode variables and estimate causal effects from observational data, but application to observational soil data is challenging.

3 | The Two Views of Causality

Understanding causality in soil science requires us to step back from the three conditions previously described and ask a more fundamental question: what do we mean when we say that one factor causes another? This question can be answered with two viewpoints, starting with the one that underpins most of DSM today.

The first viewpoint, often called the successionist view of causation, interprets causal relationships in terms of stable and repeatable regularities: when a particular configuration of conditions is reliably associated with an outcome, the former is treated as a causal factor for the latter. Although successionism is sometimes expressed in temporal terms (i.e., a cause precedes an effect), its core idea is regular association rather than chronology, which means it applies to static spatial data such as those used in DSM. Much of DSM implicitly relies on this logic: by identifying consistent relationships between soil properties and covariates across a landscape, we infer what tends to co-occur and exploit these regularities for prediction. For example, in montane temperate regions, SOC often increases with elevation because of low temperatures and slow decomposition, whereas in arid mountain systems, SOC may decline with elevation due to sparse vegetation and limited inputs. However, a successionist view remains essentially about patterns without guaranteeing that they reflect underlying processes. As Figure 1 illustrates, such associations can be fragile and changing when data are partitioned, when additional variables are included, or when confounding structure is altered. This does not make causal inference in DSM impossible under a successionist logic, but it is challenging to satisfy the three conditions previously described.

A second viewpoint, often called generative, holds that causation arises from the operation of mechanisms within systems, not from regular association alone. To explain causally in science is to specify the internal structures and

processes that bring about an effect. In this view, the emphasis falls on the mechanisms that generate outcomes. An observed association between A and B is therefore not sufficient for a causal claim. Rather, A must initiate a process through which B is produced. For example, nitrogen fertilizer does not merely co-occur with soil acidification. It triggers microbial nitrification, releasing hydrogen ions into the soil, which in turn lowers pH and drives acidification. Here the causal link rests not on association alone but on a well-specified chain of processes connecting the initial intervention to the final state, with empirical regularities providing support for the underlying mechanism.

These two views on causality have been extensively discussed in the scientific literature. As Harré (1972) has argued, even if two variables are consistently associated, even if the direction of influence is clear, and even if the effect appears inevitable under certain conditions, we may still lack a causal explanation. Causal knowledge is not exhausted by establishing that a set of elements is associated with another; what must be understood is the system of processes through which such associations arise.

SOC mapping provides a clear illustration. Under a successionist approach, one might fit a machine-learning model to predict SOC across a landscape using covariates such as remote sensing indices, slope, clay content, rainfall and temperature. The output may reveal that areas with low slope and moderate rainfall tend to hold more carbon. This is an empirical regularity that aids prediction, but it does not by itself explain why these conditions co-occur with higher SOC. In Harré's sense, such a model may be externally accurate yet remains descriptive.

A generative approach would reframe the exercise. Instead of stopping at observed associations, DSM would be guided by the known processes through which SOC is produced, transformed and stabilized: plant productivity generates litter inputs; microbial activity decomposes organic matter; soil texture and mineralogy regulate stabilization mechanisms. With this structure in view, the statistical model is no longer a black box relating predictors to outcomes, but one organized around the mechanisms that connect vegetation, climate and soil properties to soil carbon. What emerges is not just a pattern but an explanation: environments with high plant productivity and moderate moisture accumulate more soil carbon because inputs exceed decomposition losses, and fine-textured soils stabilize carbon more effectively. We argue that, among the available viewpoints, a generative view of DSM is the one most capable of satisfying the three conditions for causal inference described above.

4 | Generative Causality to Digital Soil Mapping

A generative DSM view extends the conventional two-term soil scheme:

$$\textit{soil forming factors} \rightarrow \textit{soil}, \tag{2}$$

which was formally described for DSM in McBratney et al. (2003) with the *scorpan* model, where soil in space and time is expressed as a function of soil-forming factors, namely, soil, climate, organisms/vegetation, parent material, time and spatial location. McBratney et al. (2003) describe that the only unknown is the form of the function linking soil to the forming factors, and that 'we shall not consider the direction of causality'. This reflects a successionist view: soil-forming factors and soil properties are associated in consistent ways, and these regularities are used for prediction.

The generative alternative introduces an explicit mechanistic layer. The three-term soil scheme proposed by Gerasimov (1984) replaces the direct mapping with:

$$\textit{soil forming factors} \rightarrow \textit{processes} \rightarrow \textit{soil}, \tag{3}$$

in which soil-forming factors act through identifiable processes that produce soil properties. Although the distinction appears subtle, it aligns with a generative view of causality: factors influence outcomes through mechanisms, allowing the model to represent how interventions or system changes propagate through these processes.

In a generative DSM framework, soil processes would be treated as determined once their governing mechanisms are specified. Biogeochemical processes such as organic matter decomposition, nutrient cycling or moisture dynamics follow functional relationships with environmental and soil variables. For example, decomposition rates depend on temperature, moisture, substrate quality and microbial activity. By making these dependencies explicit, the model can propagate changes in inputs through the underlying processes to predict resulting soil properties across a landscape. In this way, generative DSM moves beyond empirical associations: system behaviour arises from specified mechanisms, providing a structured foundation for reasoning about interventions and for identifying and controlling confounding factors.

The methods that support a generative causal view of DSM already exist in various forms. They fall broadly into two families: process-informed DSM and quantitative pedogenesis modelling. The first integrates mechanistic understanding into statistical models of soil variation. In process-informed machine learning (e.g., Zhang et al. 2024), predictors are selected, transformed or constrained using knowledge of the governing soil processes so that the statistical model learns relationships that are consistent with known mechanisms rather than associations. The second family, quantitative pedogenesis models (e.g., Minasny et al. 2008; Finke and Hutson 2008), simulates soil formation explicitly through processes such as organic matter accumulation, mineral weathering, horizon development or clay translocation. In these models, the pathways from soil-forming factors to soil properties are specified directly by the modeller.

We therefore argue that the three conditions of causal inference with observational data are far more naturally satisfied within a generative DSM view than within a successionist one. A generative view requires the modeller to specify the mechanisms and processes that link soil-forming factors to soil properties, which directly supports the first condition: it yields an explicit causal model rather than an empirical association structure. Because mechanisms define how factors act through processes, they also clarify which variables must be measured to block confounding,

making the assumption of causal sufficiency more plausible than in a successionist approach where confounders remain implicit. Finally, process-based representation imposes functional relationships grounded in biogeochemistry and pedogenesis, which makes the faithfulness assumption more credible. For these reasons, generative DSM provides a stronger foundation for meeting the identification conditions required for causal inference from observational data.

Although generative DSM aligns more naturally with the conditions required for causal inference, it is not straightforward to implement. Fully mechanistic models of soil processes are difficult to build: many pedogenic and biogeochemical mechanisms are only partly understood, and key boundary conditions such as initial soil states or long-term histories are rarely known with confidence. Most process-based models in soil science are also semi-mechanistic. They rely on simplifying assumptions, parameterizations and scale-averaging that approximate, rather than fully reproduce, real processes. For this reason, adopting a generative view in DSM does not mean relying on complete mechanistic simulators. Instead, it calls for a framework in which mechanistic knowledge can be incorporated where available to clarify causal structure, guide variable selection, reduce confounding and stabilize dependence patterns. Such a framework provides stronger support for causal inference from observational data than current practices rooted in a successionist view of DSM.

5 | Concluding Remarks

Can DSM be causal? Soil dynamics result from complex chains of interacting mechanisms, yet the observational data used to study these systems rarely support straightforward causal inference. The claim advanced here is that, in principle, causal reasoning within DSM is possible. Three conditions outline clearly, what must be satisfied for causal inference with observational data. We explain that two viewpoints exist for DSM causal inference, a successionist and a generative one. We argue that, among the available viewpoints, a generative view of DSM is the one most capable of satisfying the three conditions for causal inference described above. This does not mean that a successionist view precludes causal inference, but it makes it more challenging. A generative view considers that soil-forming factors influence soil properties through explicitly modelled processes. Because these processes are 'fully determined' by the modeller's specification, they provide a structured means to control confounding and better support the application of formal causal frameworks. Process-informed DSM and quantitative pedogenesis modelling are examples of techniques that support a generative view.

Yet, even if such a generative approach is possible in practice and would better support the conditions for causal inference from observational data in DSM, a more fundamental question remains: is it necessary to aim for causality in DSM? DSM, like any empirical modelling framework, has always had two primary objectives: prediction and understanding. Its predictive power is well established, and it can be used to generate and test hypotheses about soil-environment relationships. Insisting on formal causal inference with DSM may therefore be neither essential nor fully compatible with current practices. A more prudent stance may be to acknowledge the potential of generative, process-informed DSM for causal reasoning, while recognizing that, in most cases, the strength of DSM lies not in establishing causality but in predicting soil properties and identifying patterns worthy of further investigation.

Author Contributions

Lei Zhang: conceptualization, methodology, investigation, formal analysis, visualization, writing – original draft, writing – review and editing. **Alexandre M. J.-C. Wadoux:** conceptualization, methodology, investigation, writing – original draft, writing – review and editing.

Funding

Lei Zhang acknowledges the support from the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under Award Number DE-AC02-05CH11231.

Data Availability Statement

The data that support the findings of this study are available in WoSIS Soil Profile Database at https://doi.org/10.17027/isric-wdcsoils-20231130. These data were derived from the following resources available in the public domain: ISRIC—World Soil Information, https://www.isric.org/explore/wosis.

References

Batjes, N. H., L. Calisto, and L. M. de Sousa. 2024. "Providing Quality-Assessed and Standardised Soil Data to Support Global Mapping and Modelling (WoSIS Snapshot 2023)." *Earth System Science Data* 16: 4735–4765.

Byrnes, J. E., and L. E. Dee. 2025. "Causal Inference With Observational Data and Unobserved Confounding Variables." *Ecology Letters* 28: e70023.

Finke, P. A., and J. L. Hutson. 2008. "Modelling Soil Genesis in Calcareous Loess." *Geoderma* 145: 462–479.

Gerasimov, I. P. 1984. "The System of Basic Genetic Concepts That Should Be Included in Modern Dokuchayevian Soil Science." *Soviet Geography* 25: 1–14.

Gomes, L. C., P. L. Weber, C. Hermansen, et al. 2025. "Mapping Potential Water Repellency of Danish Topsoil." *Geoderma* 457: 117280.

Harré, R. 1972. *The Philosophies of Science: An Introductory Survey*. Oxford University Press.

Hempel, C. G. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press.

Jenny, H. 1941. *Factors of Soil Formation: A System of Quantitative Pedology*. McGraw-Hill Book Company.

McBratney, A. B., M. M. Santos, and B. Minasny. 2003. "On Digital Soil Mapping." *Geoderma* 117: 3–52.

Minasny, B., A. B. McBratney, and S. Salvador-Blanes. 2008. "Quantitative Models for Pedogenesis—A Review." *Geoderma* 144: 140–157.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press.

Wadoux, A. M. J.-C., M. Román-Dobarco, and A. B. McBratney. 2021. "Perspectives on Data-Driven Soil Research." *European Journal of Soil Science* 72: 1675–1689.

Wadoux, A. M. J.-C., A. Samuel-Rosa, L. Poggio, and V. L. Mulder. 2020. "A Note on Knowledge Discovery and Machine Learning in Digital Soil Mapping." *European Journal of Soil Science* 71: 133–136.

Webster, R. 2000. "Is Soil Variation Random?" *Geoderma* 97: 149–163.

Yuan, Z., F. Jiao, X. Shi, et al. 2017. "Experimental and Observational Studies Find Contrasting Responses of Soil Nutrients to Climate Change." *eLife* 6: e23255.

Zhang, L., G. B. M. Heuvelink, V. L. Mulder, S. Chen, X. Deng, and L. Yang. 2024. "Using Process-Oriented Model Output to Enhance Machine Learning-Based Soil Organic Carbon Prediction in Space and Time." *Science of the Total Environment* 922: 170778.