



OPEN

Contextual spatial modelling in the horizontal and vertical domains

Tobias Rentschler^{1,2,3,4}✉, Martin Bartelheim^{1,5}, Thorsten Behrens^{1,6,7}, Marta Díaz-Zorita Bonilla^{1,5}, Sandra Teuber^{1,4}, Thomas Scholten^{1,2,4} & Karsten Schmidt^{1,2,3,4}

Multi-scale contextual modelling is an important toolset for environmental mapping. It accounts for spatial dependence by using covariates on multiple spatial scales and incorporates spatial context and structural dependence to environmental properties into machine learning models. For spatial soil modelling, three relevant scales or ranges of scale exist: quasi-local soil formation processes that are independent of the spatial context, short-range catenary processes, and long-range processes related to climate and large-scale terrain settings. Recent studies investigated the spatial dependence of topsoil properties only. We hypothesize that soil properties within a soil profile were formed due to specific interactions between different features and scales of the spatial context, and that there are depth gradients in spatial and structural dependencies. The results showed that for topsoil, features at small to intermediate scales do not increase model accuracy, whereas large scales increase model accuracy. In contrast, subsoil models benefit from all scales—small, intermediate, and large. Based on the differences in relevance, we conclude that the relevant ranges of scales do not only differ in the horizontal domain, but also in the vertical domain across the soil profile. This clearly demonstrates the impact of contextual spatial modelling on 3D soil mapping.

Soils are crucial for agriculture, forestry, biodiversity, biofuels production, and global carbon cycling^{1,2}. The growing world population requires changes in food production towards sustainability through policies and management practices^{1,3}. Spatial knowledge of soil properties and processes can support management practices to increase the productivity which depends on water content, nutrient availability, soil acidity, and other soil quality indicators¹. Relevant data can be obtained from sophisticated soil models and spatial predictions based on machine learning. Such predictions of soil properties are based on two fundamental paradigms: the state factor equation (Eq. 1)⁴ and the universal model of spatial variation (Eq. 2)⁵. The state factor equation formalises the concept of soil forming factors⁶:

$$S = f(cl, o, r, p, t, \dots) \quad (1)$$

where S is a soil property at a certain location that develops as a non-linear function f of the soil forming factors climate (cl), organisms (o), relief (r), parent material (p), time (t), and other potentially unknown factors (\dots), which may include other soil properties⁷, spatial location⁷, and spatial context⁸. Thus, the universal model of spatial variation includes the deterministic state factor equation as well as a stochastic part of (apparently) random variation:

$$Z(s) = Z^*(s) + \varepsilon'(s) + \varepsilon' \quad (2)$$

where $Z(s)$ is the soil property, $Z^*(s)$ the deterministic component, i.e., the soil forming factors from Eq. (1), $\varepsilon'(s)$ the stochastic component of (apparently) random variation, and ε' random noise. While $Z^*(s)$ can be modelled

¹SFB 1070 ResourceCultures, University of Tübingen, 72074 Tübingen, Germany. ²Cluster of Excellence Machine Learning: New Perspectives for Science, University of Tübingen, 72076 Tübingen, Germany. ³eScience-Center, University of Tübingen, 72074 Tübingen, Germany. ⁴Department of Geosciences, Chair of Soil Science and Geomorphology, University of Tübingen, 72070 Tübingen, Germany. ⁵Department of Pre- and Protohistory and Medieval and Post-Medieval Archaeology, University of Tübingen, 72070 Tübingen, Germany. ⁶Soilution GbR, Soil and Spatial Data Science, 06484 Quedlinburg, Germany. ⁷Swiss Competence Center for Soils (KOBO), Data Science, BFH-HAFL, CH-3052 Zollikofen, Switzerland. ✉email: t.rentschler@uni-tuebingen.de

with machine learning methods by correlating the outcome with the independent environmental covariates of the state factor equation, the stochastic component $\varepsilon'(s)$ is often modelled with geostatistical methods such as kriging^{9,10} or Bayesian approaches^{11,12}. However, with multi-scale environmental covariates Behrens et al.¹³ showed that $\varepsilon'(s)$ can actually be treated as part of $Z^*(s)$ and mention short-range catenary and long-range teleconnected aeolian systems as example because both depend on relief (r) and climate (cl) but on higher hierarchical levels¹⁴.

Recent studies incorporating multi-scale environmental covariates into machine learning models showed increasing prediction accuracy, parsimony, and computational efficiency^{8,15–18}. Further, some of the referenced studies enable and foster complex geoscientific and pedological interpretations with respect to soil forming processes^{13,18–21}. The higher accuracy of the multi-scale approaches compared to approaches without accounting for the spatial context is related to the spatial variation and the interactions of soil properties and soil forming processes that are effective on multiple scales at the same time as well as over the temporal dimension. Consequently, machine learning models in combination with feature engineering techniques can account for contextual information $\varepsilon'(s)$ that reflects interacting, hierarchical, and scale dependent soil forming processes in the horizontal domain. Recent studies focused on topsoil only (mostly 0–30 cm)^{14,20,22}, but neglected subsoil (> 30 cm). However, the whole soil continuum is relevant for environmental processes and several studies have shown that the factors of soil formation are also effective in the vertical domain, for example climate (cl)^{23,24}, vegetation (o)^{23–25}, relief (r)^{19,24,26,27}, parent material (p)^{22,28}, and other environmental properties (\dots)^{28,29}. Therefore, we investigated the influence of spatial context in modelling the horizontal and the vertical domains and hypothesised that soil forming processes in subsoil relate differently to the spatial context compared to topsoil.

Behrens et al.²⁰ presented two approaches to identify the appropriate scale space for spatial modelling: using the empirical variogram to determine the range of geostatistical models (kriging), and using machine learning model sequences where coarser scales are added or dropped sequentially from the covariate space. Our objectives with this paper were (i) to test if different spatial scales may be relevant for the vertical domain, and (ii) to interpret and discuss how and why different spatial scales may influence soil properties and soil forming processes across a soil profile. To do so, we closely followed the approaches of multi-scale contextual spatial modelling to derive multi-scale terrain derivatives with the Gaussian pyramid^{17,18,20}, and the empirical variogram. We created a soil dataset of 130 soil profiles with up to 5 depth increments depending on the local soil depth. This dataset includes measurements of soil quality indicators, soil properties derived from pedo-transfer functions, and a soil quality index.

Material and methods

Study area. The study area of 1000 km² is located in the Middle Guadalquivir basin, Andalusia, Spain, 50 km NE of Seville (Fig. 1). The geologic setting separates the landscape in three main parts: (i) the Sierra Morena mountain range in the North with Palaeozoic granite, gneiss, and slate, (ii) the Guadalquivir flood plain with Pleistocene marl, calcarenite, coarse sand, and Holocene sands, and loams, and (iii) Neogene terraces of coarse gravel and cobble with sands and loams in the South^{30–32}. The slopes are typically 50–1000 m long. The study area is a heterogenous agricultural landscape with arable land, citrus, and olive plantations, pastures, and Dehesa, an agro-sylvo-pastoral system. According to the USDA soil taxonomy the predominant soil types are Alfisols, Entisols, Inceptisols, and Vertisols³³.

Soil data and environmental covariates. Soil samples were taken at 130 stratified random locations in October 2018. The strata are combinations of four geomorphic positions³⁵ (flat, footslope, slope and shoulder) and the CORINE Land Cover level 2 classes arable land, permanent crops, pastures, forest, and shrub and/or herbaceous vegetation associations, which are the most common land cover classes in the study area. The point density of the sampled locations is proportional to the stratum area with a minimum of 3 samples for the smallest stratum. At each location up to five samples were taken with an auger depending on the soil thickness and bulked from 3 replicates. The sampled increments were 0–10, 10–20, 20–30, 40–60, and 70–100 cm. We defined the sampled increments from 0 to 30 cm as topsoil and 30–100 cm as subsoil.

For lab analysis, the samples were dried at 40 °C for 24 h, root fragments were removed, sieved (< 2 mm), and ground. The spectra of all 506 samples were measured with a Tensor II (Bruker Optics, Ettlingen, Germany) for reflectance in the NIR spectrum (833–3500 nm, i.e. 2860–12,000 cm⁻¹ with a resolution of 4 cm⁻¹) and a GladiATR (Pike Technologies, Madison, WI, USA) in the MIR spectrum (2270–25,000, i.e. 400–4400 cm⁻¹ with a resolution of 2 cm⁻¹). Due to the overlapping spectra, we cut the spectra at 2500 nm and kept the spectra from 833 to 2500 nm measured with the Tensor II and from 2500 to 25,000 nm measured with the GladiATR. A subset of 97 samples representative for the strata was analysed for soil organic carbon (SOC) with a Vario EL III (Elementar, Hanau, Germany), for soil acidity (pH in KCl solution; pH_{KCl}) with a ProfiLine pH 3310 and a SenTix 81 electrode (WTW, Weilheim, Germany), and for grain size fractions clay (< 2 μm), silt (2–50 μm) and sand (50–2000 μm) with a SediGraph III (Micromeritics, Norcross, GA, USA). The 97 spectra were used as dependent variables to train partial least squares regression models and to make predictions for the remaining 409 samples³⁶. Since pre-processing of the spectra³⁷ did not improve model performance, the raw spectra were used. The root mean squared errors of these models were 0.5% for SOC content, 0.4 for pH, 4% for clay, 5% for silt, and 5% for sand content. The predictions of SOC content ranged from 0.01 to 3.61%, the pH from 3.00 to 12.11, the clay content from 0.01 to 73.94%, the silt content from 0.01 to 53.54%, and the sand content from 0.01 to 101.55%. The sum of the texture fractions ranged from 90 to 110% with a standard deviation of $\sigma = 3.25\%$.

We derived the cation exchange capacity (CEC in cmol kg⁻¹) based on a pedo-transfer function developed specifically for Spain³⁸ and the water content at field capacity (θ_{FC} in cm³ cm⁻³) using a pedo-transfer function developed for Europe³⁹. The root mean square errors (RMSE) of these pedo-transfer functions were reported

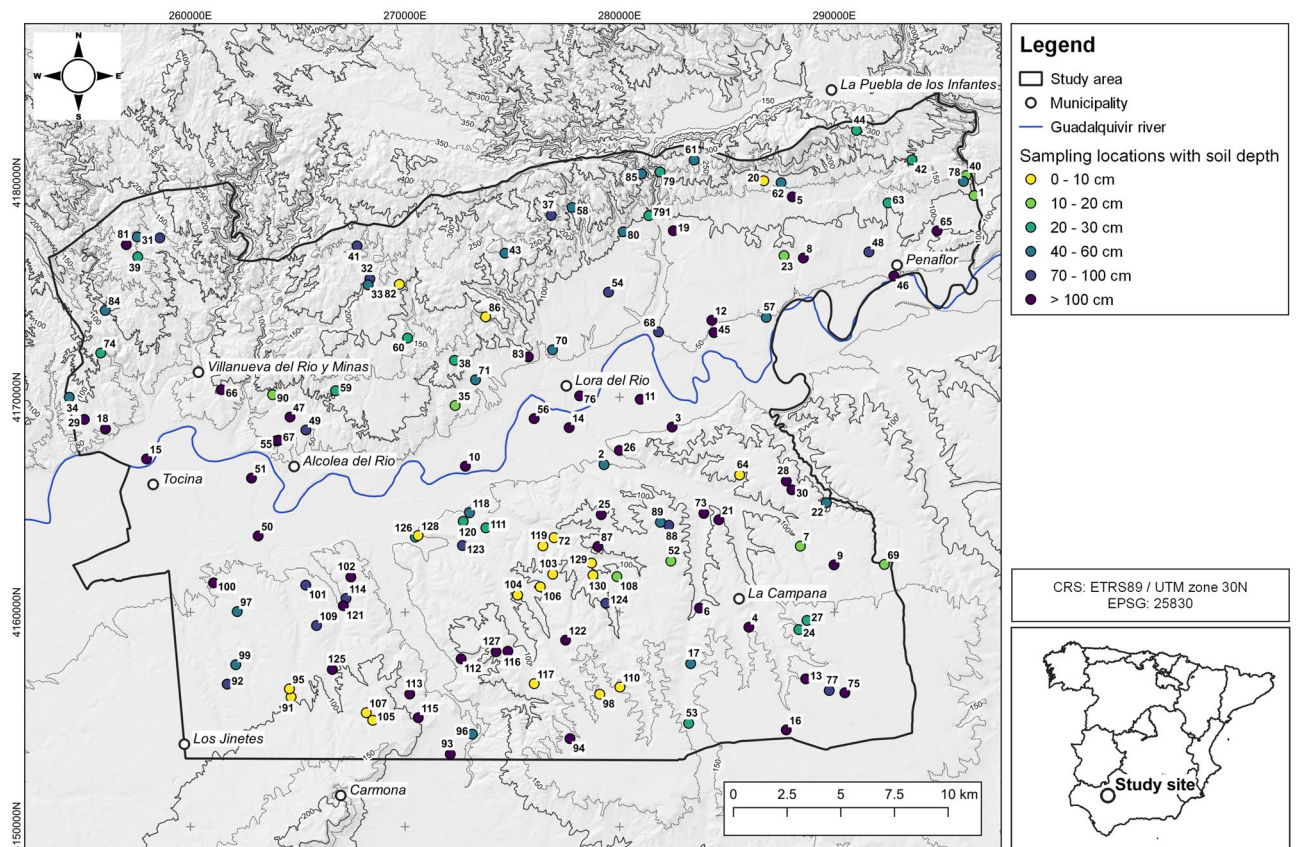


Figure 1. Map of the study area in Andalusia, Spain, with soil sampling locations and the corresponding soil depth (cm) of each profile according to the sampling design (Hill shade and contour lines derived from the digital elevation model used in this study by CNIG with QGIS³⁴. Administrative map data for the lower right panel provided by gadm.org. <https://gadm.org/license.html>).

to be $0.06 \text{ cm}^3 \text{ cm}^{-3}$ and $0.73 \text{ cmol kg}^{-1}$, respectively. Subsequently, we calculated a soil quality rating (SQR) to represent the yield potential of soils for agricultural crops based on the CEC, pH, and θ_{FC} values^{40,41}.

Variography. The variogram is a geostatistical model⁴². Empirical data is used to describe the degree of spatial autocorrelation between pairs of point measurements of environmental properties and to develop a theoretical model, which can be used for spatial modelling using kriging. The variogram consists of a set of parameters: nugget, sill, lag, and range. The nugget is the y-intercept and describes the variability at the smallest scale of the data related to noise and errors. The sill represents the maximum variability between point pairs. The lag is the radius in which point pairs are built. The spatial support that the lag corresponds to in this study is twice the spatial support of the pixel size of the downsampled Gaussian octave (see next section). The range is the maximum distance up to which the data is spatially autocorrelated. Therefore, the range is an indicator of the maximum scale of spatial context for environmental modelling^{16,20,22}.

Gaussian pyramid mixed scaling. Gaussian mixed scaling^{17,18} is a method to derive multi-scale terrain derivatives to incorporate the spatial contextual information of a landscape in a machine learning model. Gaussian mixed scaling is based on Gaussian filtering and down-sampling⁴³ and decomposes the scales of environmental covariates¹⁷. In the Gaussian pyramid, each down-sampling step removes every second column and row of the digital elevation model (DEM) of the previous step. Possible associated artifacts are minimised with a Gaussian filter before each downscaling step. The outcome of each step is called an octave. Finally, all octaves are up-sampled with an inverse scaling procedure to the original resolution of the DEM to enable subsequent data manipulation on the same resolution.

In mixed scaling, the DEM is downsampled. Then the terrain covariates are derived from each octave and are subsequently upsampled. Mixed scaling has demonstrated to be more accurate than scaling of the derived terrain covariates while providing less artefacts and a better basis for interpretation¹⁸. We used the following terrain covariates based on Zevenbergen & Thorne's equations⁴⁴ derived from the DEM of 5 m resolution: elevation, slope, sine transformed aspect, cosine transformed aspect, average curvature, profile curvature, planform curvature, flow accumulation, and topographic wetness index (Fig. 2) and calculated 11 octaves and 11 intermediate levels¹⁸ with corresponding cell sizes of the Gaussian scale space: 10, 20, 40, 80, 160, 320, 640, 1280, 2560, 5120, and 10,240 m. These covariates are the deterministic part $Z^*(s)$ and also reflect the seemingly random variation

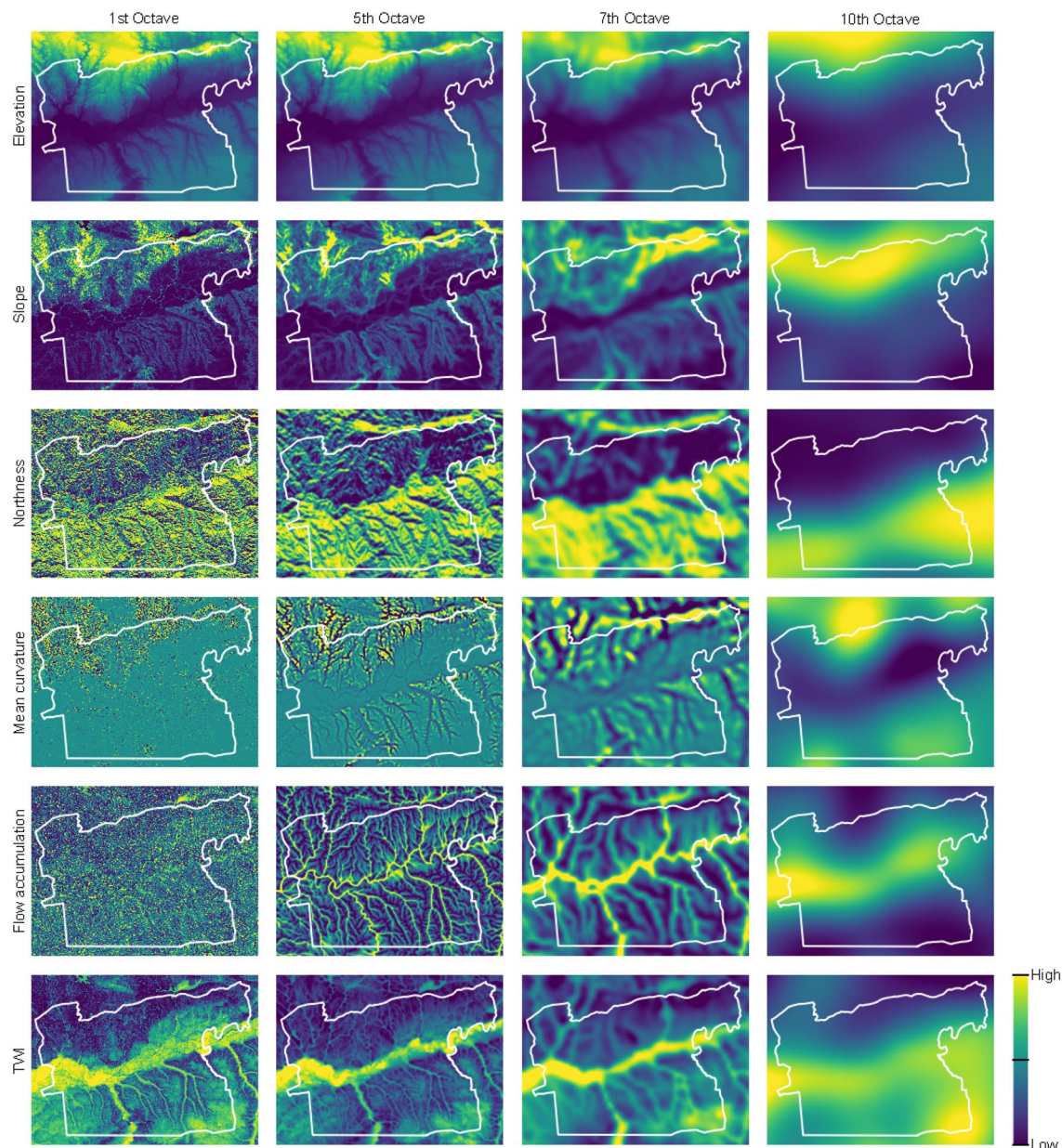


Figure 2. Selection of environmental covariates showing elevation, slope, northness, mean curvature, flow accumulation, and the topographic wetness index (TWI) at the 1st (10 m scale), 5th (160 m), 7th (640 m), and 10th octave (5120 m).

$e'(s)$ in the universal model of spatial variation as well¹³. In respect to the average length of slopes in the study area, the 1st to 4th octaves (10–80 m) describe small scale properties of the terrain. The 5th to 8th octaves (160–1280 m) represent features of the terrain on the intermediate scale and are related to catenary soil forming processes, such as erosion, sediment transport, and reallocation. The 9th to 11th octaves (2560–10,240 m) describe large scale supraregional features of the landscape, i.e. the geomorphic signature of the landscape¹⁷, such as the mountain range in the north and the flood plain of the Guadalquivir river in the centre of the study area.

Machine learning and validation. Machine learning methods have proven to be able to extract relationships between soil data and independent environmental covariates. Amongst others, random forests are applied often for spatial soil modelling in 3D^{18,19,28,45}. Random forests is an ensemble of classification and regression trees^{46,47}. In a decision tree binary splits are used recursively to homogenize the predictor variables in relation to the dependent variable, thus minimizing node impurity. Random forests use a bootstrapping approach, where a random set of predictor variables is tested at each split of a tree. The final regression model results from averaging all outputs of the decision tree ensemble. Random forests are relatively robust against overfitting and multi-collinearity, and provide a tool which facilitates interpretations of the models⁴⁷. In this study, we used the random forest implementation in R^{48–50}. For model evaluation, we used Pearson's correlation coefficient (R^2).

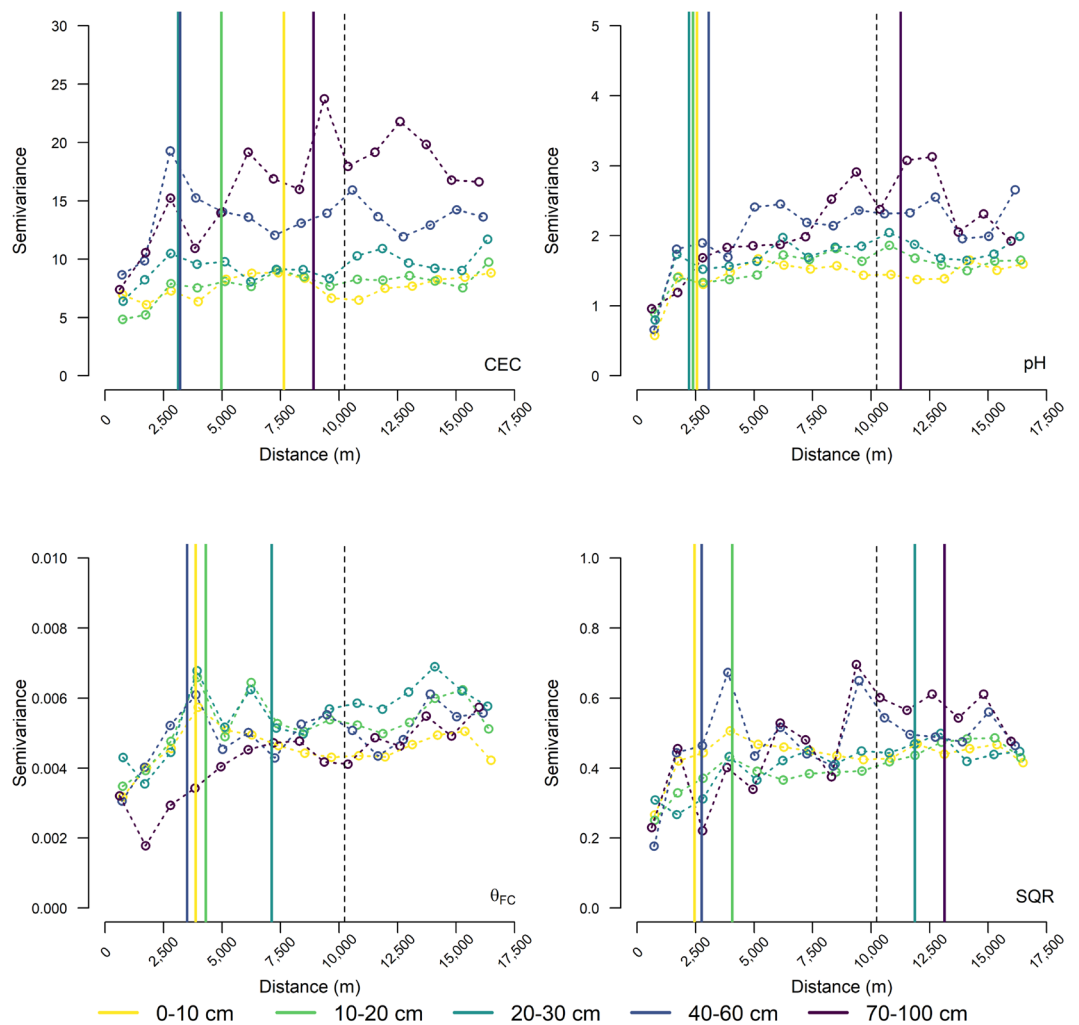


Figure 3. Empirical isotropic variograms for cation exchange capacity (CEC), soil acidity (pH), and water content at field capacity (θ_{FC}), and the soil quality rating (SQR) at the five soil depths (0–10, 10–20, 20–30, 40–60, and 70–100 cm). The solid vertical lines indicate the range of the theoretical spherical isotropic variograms for each depth increment and the dashed vertical line indicates the maximum range of the contextual machine learning models (range of θ_{FC} at 70–100 cm was 29,000 m and outside the plot boundaries).

The models were trained with grid tuning in search for the optimal model configuration⁵¹ and validated with ten times repeated ten-fold cross-validation using the caret package⁵².

We applied additive and subtractive multi-scale models²⁰ to analyse the influence of the increasing information horizon⁵³. We compared and analysed repeated model training and sequential adding of octaves representing larger scales to the training set as well as sequential subtraction (dropping) of smaller octaves from the complete set of covariates. The additive model sequence starts with the covariates at the original resolution of the DEM and subsequently adds the covariates of the next higher octave to the covariate space for model training. For the subtractive model sequence, the covariates of the smallest octaves of the covariate space are dropped sequentially. In this way a certain scale is represented by two models that contain either all smaller octaves (additive model sequence) or all larger octaves (subtractive model sequence) of the complete covariate space. With this approach increases or decreases of R^2 values of the model sequences indicate which octaves and scales are more relevant for the model training²⁰. For better visual analysis, the model results are smoothed with locally estimated scatterplot smoothing (loess⁴⁸). In total, we trained and validated 920 models (11 octaves with intermediate levels for four soil properties in five depths with additive and subtractive modelling, respectively).

Results

Variography. The empirical variograms for CEC, pH, θ_{FC} , and SQR at the five soil depths (Fig. 3) show an overall increase of semivariance with increasing distance. The largest increase in semivariance was below a distance of 2500 m corresponding to the ninth octave. The ranges of the (theoretical) spherical isotropic variograms²⁰ varied between 3000 and 9000 m for CEC, 2500–11,000 m for pH, 3500–29,000 m for θ_{FC} , and 2500–13,000 m for SQR (Fig. 3). The largest range of the spherical isotropic variograms of 13,000 m was near

the 11th octave (scale 10,240 m) which was chosen as highest octave. However, in some cases the empirical variograms showed increasing semivariance beyond the range of the theoretical variogram as also reported by Behrens et al.²⁰.

Machine learning. For most of the model sequences with the soil quality indicators (CEC, pH, θ_{FC} , and SQR; Fig. 4; Supplementary Tab. S1) at five depth increments, successive addition (left panels in Fig. 4) of coarser scales of the covariates generally increased the model accuracy (R^2), while successive (right panels in Fig. 4) removal of finer scales generally decreased model accuracy at distinct scales. Both, increase and decrease were not continuous in most model sequences.

In some model sequences with the additive approach, the model accuracies started to increase at a scale of 80–160 m for CEC, pH, and θ_{FC} (Fig. 4) in topsoil. While this was also the case for CEC and pH in subsoil in 40–60 cm depth, the increase for θ_{FC} starts with the first octave similarly to the increment of 70–100 cm. At the increment 70–100 cm models for CEC showed a marginal decrease over all scales and models for pH show a slight decrease. The subtractive model sequences for topsoil and 40–60 cm mostly showed a decrease in R^2 after removal of the scales larger than 160–320 m. The model sequences for CEC and θ_{FC} at 70–100 cm showed marginal decreases over the whole sequence, whereas the sequence for pH showed a slight increase in R^2 . The increase of the additive model sequence for SQR in 0–10 cm (Fig. 4) started at a scale of 40 m, in the second and third depth increment at 320–640 m, the model sequence for SQR in 40–60 cm showed a continuous increase across all scales, and in 70–100 cm an increase to 80 m with a plateau from 160 to 640 m, and an additional increase from 1280 m onwards. In the subtractive model sequence, the decreases in R^2 start with the corresponding scales for the depth increments from 0 to 60 cm, whereas there is no decrease for the sequence for the 70–100 cm increment.

Compared to models that used the terrain derivatives of the original resolution of the DEM of 5 m only, the model sequences mostly improved predictive accuracy of the models for all soil properties, CEC, pH, θ_{FC} , and SQR. The increase ranged from 10% for CEC (0–60 cm) to 20 in SQR (70–100 cm) in the explained variance. Two models, CEC, and pH at 70–100 cm, did not benefit from increasing scales.

Discussion

The a priori approximation of the variogram showed that the strongest increase of semivariance was below 2500 m. This indicates that the strongest spatial relation for most soil properties and depth increments is found below 2500 m and the corresponding 9th octave. The increase of semivariance above this range was covered with two additional octaves to account for potential spatial dependencies beyond the range of the theoretical variogram²⁰. The ranges of the variograms were different for the four soil properties as well as the five depth increments.

For all four soil properties, the samples from 70 to 100 cm had the largest ranges (9000–29,000 m), whereas models for the increments 0–60 cm had smaller ranges. The larger ranges for the lower depth interval may reflect the large-scale variation of the regional geology, which varies over distances of several ten kilometres. This points to soil formation processes like weathering of Palaeozoic bedrock in the Sierra Morena mountain range, and large-scale translocation processes such as sediment transport by water and river meandering in the Guadalquivir flood plain⁵⁴ formed mainly during the Pleistocene, and terrace formation during the Neogene⁵⁵. Further, large-scale climate changes during the Holocene⁵⁶ might have altered the weathering conditions of the whole area and is reflected in stable soil properties like soil texture that controls water holding capacity and CEC to a large extent⁵⁷, and therefore, soil quality, but also soil acidity due to the presence of carbonates in Neogene sediments³². However, a trend for the maximum range of the variogram in respect to soil depth, i.e., an in-situ continuously increasing range with increasing soil depth, was not apparent. This may be due to the complex interactions of the vertical and horizontal domains over space and time and especially due to human impacts⁵⁸. Processes that affect soil quality are ploughing and uniform land management over longer time periods^{58,59} as they homogenize soil properties in the horizontal domain, at least over the size of agricultural fields as well as in the vertical domain, reflected in the ploughing depth. This effect is more pronounced on soil properties in the upper depth increments compared to the lower depth intervals. Consequently, the intense agricultural land use and other human activities may induce disruption of the steady-state soil processes in the upper depth increments^{58,60}.

The increase of R^2 of the additive machine learning model sequences and the decrease of R^2 of the subtracting model sequences showed that the models improved with the multi-scale approach. This is due to the relation of CEC and θ_{FC} (with their texture and pH components) as well as pH to the large-scale subregional covariates that are functioning as proxies for the parent material (p), climate or other large scale translocation drivers as reported by Behrens et al.^{17,18,20}: the higher octaves mainly reflect the geological properties and the geomorphic signature⁶¹ of the three-part study area since they allow for a spatial segmentation of the landscape units by aggregation due to scaling^{18,62,63}. Therefore, the shape of the landscape, which is determined by the geological settings, is used inversely to separate the geological zones (Fig. 4).

The increase of R^2 of the additive model sequences started at different scales. This means that for the different soil properties and each sampling increment, different scales are relevant in the study area and supports our hypothesis. The modelling sequences for topsoil increments tended to increase with intermediate octaves, whereas increases for subsoil increments were invariant. These differences in the additive model sequences can be explained regarding the state factor equation and the universal model of spatial variation. On the small scales the influence of the terrain covariates (r) on soil formation processes in topsoil is small, whereas the influence on soil processes in subsoil is larger. This may be related to the influence of catenary soil forming processes as erosion, sediment transport, and allocation^{19,64}, which is active on intermediate scales (160–1280 m) and related to slope length. In contrast, subsoil is not influenced by recent erosion processes but by the geological and

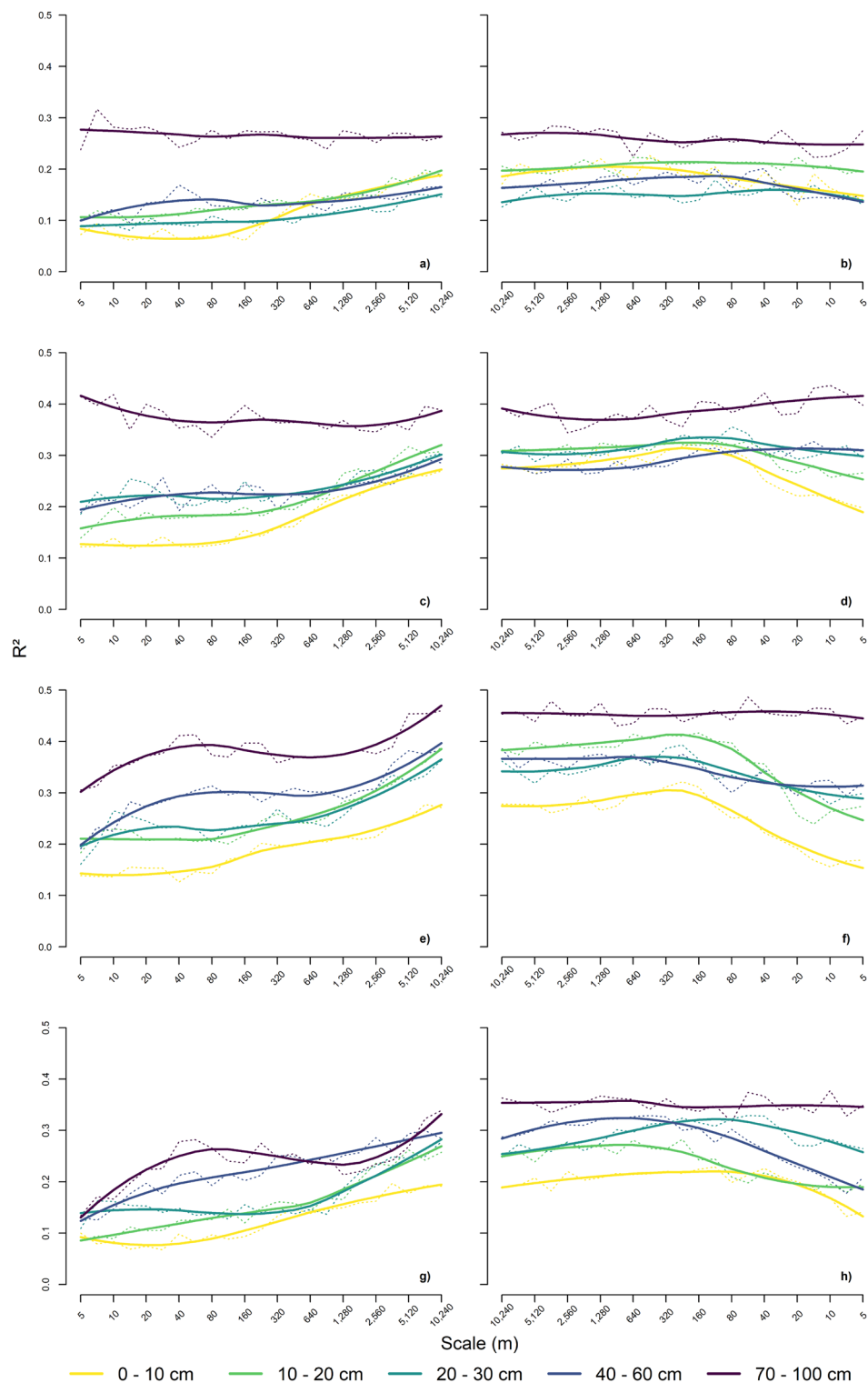


Figure 4. Results of the contextual machine learning models for cation exchange capacity (CEC; first row), soil acidity (pH; second row), water content at field capacity (θ_{FC} ; third row) and the soil quality rating (SQR; fourth row) at the five soil depths (0–10, 10–20, 20–30, 40–60, and 70–100 cm). The left panels (a, c, e, and g) show the additive models, and the right panels (b, d, f, and h) show the subtractive models. The correlation of the model predictions with the measured values are represented by the dotted lines, and smoothed trends are represented by solid lines (see Supplementary Tab. S1 for details).

geomorphological history of the landscapes, incorporated into the model with the increasing Gaussian scale space as proxy data as explained above. The large-scale nature of the highest octaves also reflects the spatial context, i.e. the universal model of spatial variation, and the segmented study area that is, however, not limited to the geomorphic signature as described above, but also comprises land use types that can be differentiated on the supraregional level and that are related to the geomorphic setting (predominantly Dehesa and olive plantations in the mountain range, citrus plantations and irrigation-intensive crops at the floodplain and its proximity, and olive plantations and non-irrigated cereals at the Neogene terraces). The importance of the large-scale covariates in the modelling, thus, also reflects the human activities and their impact on soil formation through intensive agriculture as the accessibility for agricultural machines and the construction of irrigation facilities is affected by the geomorphology. This impact can either occur through erosion, sediment transport, and reallocation on the intermediate scale with catenary processes, but also through the different cultivated crops, irrigation, fertilization, and land use change over time that is closely related to socio-economic and political factors⁶⁵.

The scaled terrain covariates did not contain additional relevant information for the model sequences for CEC and pH at 70–100 cm. Maybe further relevant information lies beyond the range recommended by the variogram. In general, Fig. 4 does not show a plateau in the R^2 at larger scales, which was found in previous studies^{8,16,17,20}. This indicates, that either additional larger octaves or predictors such as Euclidean distance transforms should be tested to increase modelling accuracies. It is recommended to increase the scales as far as possible with respect to interpretability and domain knowledge. Modelling beyond this maximum scale falls beyond the so-called information horizon³³ and is only recommended when interpretability is not the objective.

To identify the effect of smallest relevant scales we used subtractive model sequences. The R^2 of the subtractive model sequences decreased with the 6th octave and below for the soil quality indicators and for the 3rd–4th octave for topsoil and 6–7th octave for subsoil. Smaller scales did not improve the model performance, since they comprised irrelevant information or noise. Thus, they were not relevant for the individual models. This is related to the effective scales of the physical and chemical processes that are relevant for soil formation. One reason might be that the minimum scales of our input data required to achieve the most accurate model results are coarser than the original (finest) scale of the selected environmental covariates¹⁸. Regarding the agricultural use of the area and the size of the agricultural fields of a few hundred meters, the minimum range of multi-scale covariates, i.e., 80 m (4th octave), also may be linked to homogenization of soil with ploughing, fertilisation, and irrigation as mentioned above. Thus, the management practices may have disrupted steady-state soil processes on the small to intermediate scales^{58,60}. Subsequently, the relevant scales were not only linked to small and large-scale geomorphologic patterns, but also to the physical, chemical, and biotic factors of farming⁵⁸, which reshaped the landscape and may have blurred the scales reflecting the natural pattern of the landscape.

A direct comparison of the multi-scale model sequences with the empirical variograms is limited, since most of the relevant covariates have a spatial support that is smaller than the first lag of the variogram with a spatial support of 1500 m which corresponds to the 8th octave. Consequently, the multi-scale approach can account for soil forming processes on smaller scales than the variogram with automatically calculated lag distances. More research on this topic is necessary. However, even if we cannot quantify this effect, the better prediction results underpin the great potential of multi-scale modelling to predict soil processes and properties in landscapes that are both diverse in terms of their natural conditions and in terms of land use and land management.

In summary, since the different soil properties as well as the different soil depths showed varying relevant scales, we suggest that 3D soil models may require different sets of multi-scale environmental covariates to account for different soil forming processes of the state factor equation. Further, there may be landscapes where models for a certain soil depth do not benefit from multi-scale contextual information, because human activity may interfere with steady-state soil processes. An example are models for topsoil using satellite data that typically only reflect the upper centimetres of the Earth surface²⁴.

Thus, disentangling natural and human signals over scales with multi-scale contextual modelling of soil processes and properties can pave the way from soil science to many other fields of research. Disciplines that are related to landscape development, e.g. (pedo-)archaeology, archaeobiology, geology, and climatology, benefit from the better understanding of spatial scales to reconstruct land use practices^{66,67} and paleo-climate conditions¹³. Since time can also be interpreted via spatial scales and is part of the state factor equation, the space-for-time substitution⁶⁸ concept in geosciences and ecology⁶⁹ that focuses on landscape development should incorporate scale in spatial analysis.

Conclusion

There are two fundamental paradigms in digital soil mapping: the state factor equation with structural dependencies of soil formation and the universal model of spatial variation with spatial dependencies of soil formation. The multi-scale contextual spatial modelling approach is suitable to incorporate both structural and spatial dependencies into machine learning models. We hypothesised that different spatial scales may be relevant for different soil depths and interpreted how the differences may be related to the spatial context and soil formation.

We found:

- Different spatial scales are relevant in the horizontal (soil properties) and the vertical (soil depth) domain, and
- The relevant scales can be linked to the state factor equation and its factors relief (r), parent material (p) and organisms (o), i.e., land use and human impact.

These examples show, that pedologic and (physical and human) geographic domain knowledge is important to link the factors of the state factor equation and to interpret the relations to potential soil forming processes in

different soil depths. Thus, we suggest using multi-scale environmental covariates with different spatial support to account for the specific soil forming processes in specific soil depths when modelling soil properties in 3D.

Data availability

The original terrain data is published under the CC-BY 4.0 license by Centro Nacional de Información Geográfica (CNIG) of the Spanish government; last accessed March 31st 2020 (<https://doi.org/10.7419/162.09.2020>). The soil data and the processed terrain data is published at PANGAEA (<https://doi.pangaea.de/10.1594/PANGAEA.938522> and <https://doi.pangaea.de/10.1594/PANGAEA.938774>).

Received: 10 January 2022; Accepted: 25 May 2022

Published online: 09 June 2022

References

- Godfray, H. C. J. *et al.* Food security: The challenge of feeding 9 billion people. *Science* **327**, 812–818. <https://doi.org/10.1126/science.1185383> (2010).
- Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Quéré, C., Myneni, R. B., Piao, S., Thornt, P. Carbon and Other Biogeochemical Cycles. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (ed Stocker, T. F. *et al.*) 465–470 (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013).
- Bouma, J. Soil science contributions towards Sustainable Development Goals and their implementation: Linking soil functions with ecosystem services. *J. Plant Nutr. Soil Sci.* **177**, 111–120. <https://doi.org/10.1002/jpln.201300646> (2014).
- Jenny, H. *Factors of Soil Formation. A System of Quantitative Pedology* (McGraw-Hill Book Company, New York, 1941).
- Matheron, G. Principles of geostatistics. *Econ. Geol.* **58**, 1246–1266. <https://doi.org/10.2113/gsecongeo.58.8.1246> (1963).
- Dokuchaev, V. V. *The Russian Chernozem* (St. Petersburg, 1883).
- McBratney, A., Mendonça Santos, M. & Minasny, B. On digital soil mapping. *Geoderma* **117**, 3–52. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4) (2003).
- Behrens, T. *et al.* Hyper-scale digital soil mapping and soil formation analysis. *Geoderma* **213**, 578–588. <https://doi.org/10.1016/j.geoderma.2013.07.031> (2014).
- Krige, D. G. A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. Chem. Metall. Min. Soc. S. Afr.* **52**, 119–139 (1951).
- Burgess, T. M. & Webster, R. Optimal interpolation and isarithmic mapping of soil properties. *J. Soil Sci.* **31**, 333–341. <https://doi.org/10.1111/j.1365-2389.1980.tb02085.x> (1980).
- Aitkenhead, M. J. & Aalders, I. H. Predicting land cover using GIS, Bayesian and evolutionary algorithm methods. *J. Environ. Manag.* **90**, 236–250. <https://doi.org/10.1016/j.jenvman.2007.09.010> (2009).
- Banerjee, S., Carlin, B. P. & Gelfand, A. E. *Hierarchical Modeling and Analysis for Spatial Data* 2nd edn. (CRC Press, Boca Raton, 2015).
- Behrens, T., MacMillan, R. A., Viscarra Rossel, R. A., Schmidt, K. & Lee, J. Teleconnections in spatial modelling. *Geoderma* **354**, 113854. <https://doi.org/10.1016/j.geoderma.2019.07.012> (2019).
- MacMillan, R., Jones, R. & McNabb, D. H. Defining a hierarchy of spatial entities for environmental analysis and modeling using digital elevation models (DEMs). *Comput. Environ. Urban Syst.* **28**, 175–200. [https://doi.org/10.1016/S0198-9715\(03\)00019-X](https://doi.org/10.1016/S0198-9715(03)00019-X) (2004).
- Behrens, T., Zhu, A.-X., Schmidt, K. & Scholten, T. Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma* **155**, 175–185. <https://doi.org/10.1016/j.geoderma.2009.07.010> (2010).
- Behrens, T., Schmidt, K., Zhu, A. X. & Scholten, T. The ConMap approach for terrain-based digital soil mapping. *Eur. J. Soil Sci.* **61**, 133–143. <https://doi.org/10.1111/j.1365-2389.2009.01205.x> (2010).
- Behrens, T., Schmidt, K., MacMillan, R. A. & Viscarra Rossel, R. A. Multiscale contextual spatial modelling with the Gaussian scale space. *Geoderma* **310**, 128–137. <https://doi.org/10.1016/j.geoderma.2017.09.015> (2018).
- Behrens, T., Schmidt, K., MacMillan, R. A. & Viscarra Rossel, R. A. Multi-scale digital soil mapping with deep learning. *Sci. Rep.* **8**, 15244. <https://doi.org/10.1038/s41598-018-33516-6> (2018).
- Rentschler, T. *et al.* Comparison of catchment scale 3D and 2.5D modelling of soil organic carbon stocks in Jiangxi Province, PR China. *PLoS ONE* **14**, e0220881. <https://doi.org/10.1371/journal.pone.0220881> (2019).
- Behrens, T. *et al.* The relevant range of scales for multi-scale contextual spatial modelling. *Sci. Rep.* **9**, 14800. <https://doi.org/10.1038/s41598-019-51395-3> (2019).
- Reichstein, M. *et al.* Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204. <https://doi.org/10.1038/s41586-019-0912-1> (2019).
- Kerry, R. & Oliver, M. A. Soil geomorphology: Identifying relations between the scale of spatial variation and soil processes using the variogram. *Geomorphology* **130**, 40–54. <https://doi.org/10.1016/j.geomorph.2010.10.002> (2011).
- Jobbágy, E. G. & Jackson, R. B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecol. Appl.* **10**, 423–436 (2000).
- Taghizadeh-Mehrjardi, R. *et al.* Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. *Remote Sens.* **12**, 1095. <https://doi.org/10.3390/rs12071095> (2020).
- Murphy, B. W., Wilson, B. R. & Koen, T. Mathematical functions to model the depth distribution of soil organic carbon in a range of soils from New South Wales, Australia under different land uses. *Soil Syst.* **3**, 46. <https://doi.org/10.3390/soilsystems3030046> (2019).
- Aldana Jague, E. *et al.* High resolution characterization of the soil organic carbon depth profile in a soil landscape affected by erosion. *Soil Tillage Res.* **156**, 185–193. <https://doi.org/10.1016/j.still.2015.05.014> (2016).
- Milne, G. Normal erosion as a factor in soil profile development. *Nature* **138**, 548–549. <https://doi.org/10.1038/138548c0> (1936).
- Rentschler, T. *et al.* 3D mapping of soil organic carbon content and soil moisture with multiple geophysical sensors and machine learning. *Vadose Zone J.* <https://doi.org/10.1002/vzj2.20062> (2020).
- Moghadas, D., Taghizadeh-Mehrjardi, R. & Triantafyllis, J. Probabilistic inversion of EM38 data for 3D soil mapping in central Iran. *Geoderma Reg.* **7**, 230–238. <https://doi.org/10.1016/j.geodrs.2016.04.006> (2016).
- Civis, J. *et al.* Cuenca del Guadalquivir. In *Geológica de España* (ed. Vera, J. A.) 543–550 (Igme, Madrid, 2004).
- Wolf, D. & Faust, D. Western Mediterranean environmental changes: Evidences from fluvial archives. *Quat. Sci. Rev.* **122**, 30–50. <https://doi.org/10.1016/j.quascirev.2015.04.016> (2015).
- Aguirre, J. *et al.* An enigmatic kilometer-scale concentration of small mytilids (Late Miocene, Guadalquivir Basin, S Spain). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **436**, 199–213. <https://doi.org/10.1016/j.palaeo.2015.07.015> (2015).
- Gómez-Miguel, V. *Mapa de Suelos de España* (Centro Nacional de Información Geográfica (CNIG), Madrid, 2005).

34. QGIS Development Team. QGIS Geographic Information System (QGIS Association, 2020).
35. Jasiewicz, J. & Stepinski, T. F. Geomorphons—a pattern recognition approach to classification and mapping of landforms. *Geomorphology* **182**, 147–156. <https://doi.org/10.1016/j.geomorph.2012.11.005> (2013).
36. Viscarra Rossel, R. A., Walvoort, D., McBratney, A. B., Janik, L. J. & Skjemstad, J. O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **131**, 59–75. <https://doi.org/10.1016/j.geoderma.2005.03.007> (2006).
37. Stevens, A. & Ramirez-Lopez, L. An introduction to the prospectr package (2014).
38. Khaledian, Y. *et al.* Modeling soil cation exchange capacity in multiple countries. *CATENA* **158**, 194–200. <https://doi.org/10.1016/j.catena.2017.07.002> (2017).
39. Tóth, B. *et al.* New generation of hydraulic pedotransfer functions for Europe. *Eur. J. Soil Sci.* **66**, 226–238. <https://doi.org/10.1111/ejss.12192> (2015).
40. Hazelton, P. & Murphy, B. *Interpreting Soil Test Results. What do all the Numbers Mean?* (CSIRO Publishing, Clayton South, 2007).
41. Pulido, M., Schnabel, S., Contador, J. F. L., Lozano-Parra, J. & Gómez-Gutiérrez, Á. Selecting indicators for assessing soil quality and degradation in rangelands of Extremadura (SW Spain). *Ecol. Indic.* **74**, 49–61. <https://doi.org/10.1016/j.ecolind.2016.11.016> (2017).
42. Pebesma, E. J. Multivariable geostatistics in S: The gstat package. *Comput. Geosci.* **30**, 683–691. <https://doi.org/10.1016/j.cageo.2004.03.012> (2004).
43. Burt, P. J. & Adelson, E. H. The Laplacian pyramid as a compact image Code. *IEEE Trans. Commun.* **31**, 532–540. <https://doi.org/10.1109/TCOM.1983.1095851> (1983).
44. Zevenbergen, L. W. & Thorne, C. R. Quantitative analysis of land surface topography. *Earth Surf. Process. Landf.* **12**, 47–56. <https://doi.org/10.1002/esp.3290120107> (1987).
45. Grimm, R., Behrens, T., Märker, M. & Elsenbeer, H. Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. *Geoderma* **146**, 102–113. <https://doi.org/10.1016/j.geoderma.2008.05.008> (2008).
46. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees* (Chapman and Hall, 1984).
47. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. <https://doi.org/10.1023/A:1010933404324> (2001).
48. R Core Team. R: A Language and Environment for Statistical Computing (Vienna, Austria, 2021).
49. RStudio Team. RStudio: Integrated Development Environment for R. Available at <http://www.rstudio.com/> (Boston, MA, 2021).
50. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
51. Schmidt, K., Behrens, T. & Scholten, T. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. *Geoderma* **146**, 138–146. <https://doi.org/10.1016/j.geoderma.2008.05.010> (2008).
52. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26. <https://doi.org/10.18637/jss.v028.i05> (2008).
53. Behrens, T. & Viscarra Rossel, R. A. On the interpretability of predictors in spatial data science: The information horizon. *Sci. Rep.* **10**, 16737. <https://doi.org/10.1038/s41598-020-73773-y> (2020).
54. Caro Gómez, J. A., Del Díaz Olmo, F., Artigas, R. C., Recio Espejo, J. M. & Barrera, C. B. Geoarchaeological alluvial terrace system in Tarazona: Chronostratigraphical transition of Mode 2 to Mode 3 during the middle-upper pleistocene in the Guadalquivir River valley (Seville, Spain). *Quat. Int.* **243**, 143–160. <https://doi.org/10.1016/j.quaint.2011.04.022> (2011).
55. Schaller, M. *et al.* Spatial and temporal variations in denudation rates derived from cosmogenic nuclides in four European fluvial terrace sequences. *Geomorphology* **274**, 180–192. <https://doi.org/10.1016/j.geomorph.2016.08.018> (2016).
56. Finné, M., Holmgren, K., Sundqvist, H. S., Weiberg, E. & Lindblom, M. Climate in the eastern Mediterranean, and adjacent regions, during the past 6000 years—A review. *J. Archaeol. Sci.* **38**, 3153–3173. <https://doi.org/10.1016/j.jas.2011.05.007> (2011).
57. Vogel, H.-J. *et al.* Quantitative evaluation of soil functions: Potential and state. *Front. Environ. Sci.* <https://doi.org/10.3389/fenvs.2019.00164> (2019).
58. Amundson, R. Factors of soil formation in the 21st century. *Geoderma* **391**, 114960. <https://doi.org/10.1016/j.geoderma.2021.114960> (2021).
59. Diacono, M. & Montemurro, F. Long-term effects of organic amendments on soil fertility. A review. *Agron. Sustain. Dev.* **30**, 401–422. <https://doi.org/10.1051/agro/2009040> (2010).
60. Amundson, R., Heimsath, A., Owen, J., Yoo, K. & Dietrich, W. E. Hillslope soils and vegetation. *Geomorphology* **234**, 122–132. <https://doi.org/10.1016/j.geomorph.2014.12.031> (2015).
61. Pike, R. J. The geometric signature: Quantifying landslide-terrain types from digital elevation models. *Math. Geol.* **20**, 491–511. <https://doi.org/10.1007/BF00890333> (1988).
62. Lark, R. M. & Webster, R. Analysing soil variation in two dimensions with the discrete wavelet transform. *Eur. J. Soil Sci.* **55**, 777–797. <https://doi.org/10.1111/j.1365-2389.2004.00630.x> (2004).
63. Schmidt, K., Behrens, T., Friedrich, K. & Scholten, T. A method to generate soilscape maps from soil maps. *J. Plant Nutr. Soil Sci.* **173**, 163–172. <https://doi.org/10.1002/jpln.200800208> (2010).
64. Scholten, T. *et al.* On the combined effect of soil fertility and topography on tree growth in subtropical forest ecosystems—a study from SE China. *J. Plant Ecol.* **10**, 111–127. <https://doi.org/10.1093/jpe/rtw065> (2017).
65. Mar Delgado-Serrano, M. & Ángel Hurtado-Martos, J. Land use changes in Spain. Drivers and trends in agricultural land use. *EU Agrar. Law* **7**, 1–8. <https://doi.org/10.2478/eual-2018-0006> (2018).
66. Kühn, P., Lehdorff, E. & Fuchs, M. Lateglacial to Holocene pedogenesis and formation of colluvial deposits in a loess landscape of Central Europe (Wetterau, Germany). *CATENA* **154**, 118–135. <https://doi.org/10.1016/j.catena.2017.02.015> (2017).
67. Scherer, S. *et al.* Middle Bronze Age land use practices in the northwestern Alpine foreland—a multi-proxy study of colluvial deposits, archaeological features and peat bogs. *Soil* **7**, 269–304. <https://doi.org/10.5194/soil-7-269-2021> (2021).
68. Pickett, S. T. A. Space-for-time substitution as an alternative to long-term studies. In *Long-Term Studies in Ecology* (ed. Likens, G. E.) 110–135 (Springer, New York, 1989). https://doi.org/10.1007/978-1-4615-7358-6_5.
69. Blois, J. L., Williams, J. W., Fitzpatrick, M. C., Jackson, S. T. & Ferrier, S. Space can substitute for time in predicting climate-change effects on biodiversity. *Proc. Natl. Acad. Sci.* **110**, 9374–9379. <https://doi.org/10.1073/pnas.1220281110> (2013).

Acknowledgements

We thank the German Research Foundation (DFG) for supporting this research through the Collaborative Research Centre SFB 1070 RESOURCECULTURES (project number 215859406; subprojects A02, S, and Z), the Germany's Excellence Strategy Cluster EXC 2064 Machine Learning: New Perspectives for Science (project number 390727645), the eScience Center of the University of Tübingen, and the Open Access Publishing Fund of the University of Tübingen. Further, we are indebted to Philipp Baumann, Laura Bindereif, Matthias Czechowski, Philipp Gries, Susan Mentzer, Renate Nitschke, and Carolin Walper for their advice, help, and assistance with field work, lab work, and data analysis.

Author contributions

T.R., T.S., and K.S. conceived and designed the study, lead, and conducted the data analyses, interpretation, and writing; T.B. and S.T. contributed to data analysis, interpretation, and writing; M.B., M.D.-Z.B., and T.S. contributed to project administration, interpretation, and writing.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-13514-5>.

Correspondence and requests for materials should be addressed to T.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022