



High-resolution mapping of compressed vis-NIR spectral data of Danish topsoils

Sebastian Gutierrez^{a,*}, Mogens H. Greve^a, Lucas Carvalho Gomes^a, Anders B. Møller^b, Anne-Cathrine S. Danielsen^a, Peter L. Weber^a, Charles Pesch^a, Maria Knadel^a, Deividas Mikstas^a, Amelie Beucher^c, Marzieh Zaresourmanabad^a, Per Møldrup^d, David A. Robinson^e, Lis W. de Jonge^a, Sarem Norouzi^a

^a Section of Soil Physics and Hydropedology, Department of Agroecology, Aarhus University, Tjele, Denmark

^b SEGES Innovation P/S, Aarhus, Denmark

^c Pioneer Center Land-CRAFT, Department of Agroecology, Aarhus University, Aarhus, Denmark

^d Division of Civil and Environmental Engineering, Department of the Built Environment, Aalborg University, Aalborg, Denmark

^e UK Centre for Ecology & Hydrology, Environment Centre Wales, Deiniol Road, Bangor, UK

ARTICLE INFO

Handling Editor: Dr. H. Neely

Keywords:

Soil spectroscopy

Latent variables

Machine learning

Environmental covariates

Soil-landscape modelling

ABSTRACT

Visible and near-infrared (vis-NIR) spectroscopy enables rapid and cost-effective soil characterization, supporting the estimation of soil chemical, physical, and biological properties. Most applications, however, focus on using spectroscopy to predict individual soil properties and augment conventional laboratory analyses, while the potential of spectral information to study soil variation in space remains largely unexplored. Latent variables derived from the dimensionality reduction of spectral data offer a compact representation of soil variability, capturing multiple soil properties simultaneously. When spatially predicted, these latent variables can provide a means to investigate soil-landscape relationships in a spatially explicit context.

This study aimed to (i) model the spatial variation of latent variables derived from compressed topsoil vis-NIR spectral data across Denmark at 10 m resolution using a digital soil mapping approach; (ii) identify the drivers, in relation to SCORPAN factors, controlling their spatial variation; and (iii) examine how the predicted latent variables represent the soil-landscape relationships across Denmark. An earlier study compressed the data using principal component analysis, kernel principal component analysis, shallow autoencoders, and convolutional autoencoders. To predict the latent variables in space, we used 32 predictors comprising harmonized environmental layers representing climate, relief, and parent material. A weighted machine learning algorithm was used to model each latent variable, with bootstrap resampling providing pixel-level uncertainty estimates. Overall, climate was the main driver across all methods, followed by topography and parent material, mainly the extent of clay till deposits.

By integrating compressed spectral data with diverse spatial predictors, we captured complex, non-linear soil-landscape relationships. The resulting high-resolution maps offer a spatial description of soil variation in relation to soil-forming factors and spectral signature. Our maps can be used to improve point-based predictions of soil properties or provide enhanced covariates to support digital soil mapping of soil properties and classes.

1. Introduction

High-resolution and up-to-date soil information is needed to support relevant scientific research and evidence-based decision-making in relation to soil health (Chen et al., 2023). Studies on soil spatial variation require intensive field and laboratory work, involving a significant

investment of resources and time. This creates the need for accurate and low-cost tools to analyze the variation of soil properties at different scales (Ramirez-Lopez et al., 2019).

Spectroscopy enables estimation of numerous chemical, physical and biological soil properties, and provides a cost-effective alternative to traditional laboratory analyses (Hu et al., 2026; Viscarra Rossel et al.,

* Corresponding author.

E-mail address: seb.gd@agro.au.dk (S. Gutierrez).

<https://doi.org/10.1016/j.geoderma.2026.118002>

Received 6 March 2026; Received in revised form 17 June 2026; Accepted 17 August 2026

Available online 19 August 2026

0016-7061/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

2010; Viscarra Rossel et al., 2006a). The underlying mechanisms for characterizing soil information using vis-NIR spectroscopy (400–2500 nm) are the relevant first, second, and third overtones, as well as the stretch combination vibrations (Stenberg et al., 2010). Absorptions in the visible region (400–780 nm) are primarily associated with minerals that contain iron (e.g., hematite, goethite), and chromophores associated with soil-organic matter. Absorptions in the NIR region (780–2500 nm) result from the overtones of OH, SO₄, and CO₃ groups, as well as combinations of fundamental features of H₂O and CO₂ (Clark, 1999). Clay minerals can show absorption in the vis-NIR region due to metal-OH bend plus O-H stretch combinations (Viscarra Rossel et al., 2006b). Carbonates exhibit absorption peaks in the near-infrared region (Clark, 1999). The dominant absorption bands of water in soil are around 1400–1900 nm (Liu et al., 2002). The vis-NIR spectroscopy can aid in characterizing soil properties over large areas with varying conditions and land uses (Hu et al., 2026; Yang et al., 2023b).

Spectral data is commonly used to increase sampling density by augmenting laboratory soil measurements with predicted values of soil properties (Camargo et al., 2015). These augmented datasets have been used to produce soil maps at different scales with a wide range of approaches. At the continental scale, Viscarra Rossel et al. (2006b) and Viscarra Rossel et al. (2015) produced accurate maps of soil properties with augmented data derived from vis-NIR spectroscopy (e.g., texture, bulk density, soil organic carbon content, pH, P, and N) for Australia. At the farm scale, vis-NIR spectroscopy has also been successfully applied to produce spatially denser datasets. For example, Brodský et al. (2013) used vis-NIR spectra to predict topsoil OC, and Ramirez-Lopez et al. (2019) demonstrated the utility of vis-NIR data for mapping texture and exchangeable calcium. These studies consistently show that augmenting laboratory measurements with vis-NIR-based predictions improves the accuracy and effective spatial resolution of soil maps, even though these predictions remain less precise than direct laboratory analyses.

Vis-NIR spectral data have also been used as input to produce environmental covariates and predict soil properties and classes within the digital soil mapping (DSM) context. Yang et al. (2023a) reported that environmental layers derived from spectral information can capture the soil-forming factors and improve the accuracy of the predicted soil properties. Viscarra Rossel et al. (2010), Viscarra Rossel and Chen (2011), and more recently Sun et al. (2023) showed that the vis-NIR reflectance is an integrative, multiparameter measure of the soil that can assist in predictive modeling and mapping of soil properties at a large scale.

In Denmark, several studies have shown the relationship between vis-NIR spectral information and soil properties, including soil water repellency (Blaesbjerg et al., 2022; Knadel et al., 2016), clay-to-SOC ratio (Hermansen et al., 2016), soil particle density (Weber et al., 2022), and soil texture (Hermansen et al., 2017; Knadel et al., 2013). In the DSM context, Knadel et al. (2013) modeled the spatial variation of 578 topsoil vis-NIR spectral measurements and related it to the topsoil texture class and glacial boundaries obtained from historic maps. They compressed the spectral information using principal component analysis (PCA) and spatialized the scores based on the spatial autocorrelation among samples. Peng et al. (2015) demonstrated the improvement in SOC predictions at regional scale when combining remote sensing products and maps of laboratory vis-NIR spectra. Likewise, Beucher et al. (2017) predicted the extent of acid sulfate soils in wetland areas of a catchment in Denmark using maps of laboratory vis-NIR spectral data. Further work is still needed to fully explore the potential of vis-NIR data to support national-scale soil mapping.

These earlier studies have shown that spatialized variables resulting from vis-NIR data compression, here referred to as latent variables, can be used to estimate soil properties. Despite these advances several gaps remain the use of vis-NIR spectral data for soil mapping. First, most studies have focused on using spectroscopy to augment laboratory soil data, while less attention has been given to exploiting spectral

information as spatially continuous environmental covariate within digital soil mapping frameworks. Second, existing studies rely on linear dimensionality reduction techniques such as PCA, potentially overlooking complex and non-linear relationships embedded in soil spectra. Third, there is still limited understanding of how spatially predicted latent variables derived from spectral data in Denmark relate to soil-landscape relationships and soil-forming factors.

We hypothesize that latent variables derived from vis-NIR spectral data can effectively represent soil-landscape relationships and reveal patterns of soil variation at the national scale. We tested the hypothesis and addressed the gaps by using a point dataset containing compressed vis-NIR spectra from topsoil samples (0–30 cm depth) collected all over Denmark (Norouzi et al., 2026). Norouzi et al. (2026) generated latent variables from vis-NIR spectral data using classical (PCA) and kernel PCA (KPCA) and deep learning approaches (autoencoders). They produced a dataset with the point-based latent variables for Denmark that we used as input for our study. We aimed to: (1) model the spatial variation of latent variables derived from compressed topsoil vis-NIR spectral data across Denmark at 10 m resolution using a digital soil mapping approach; (2) identify the drivers, in relation to SCORPAN factors, controlling their spatial distribution; and (3) examine how the predicted latent variables represent the soil-landscape relationships across Denmark. This study provides new insights into how we can understand the soil variation and soil-landscape relationships using continuous maps of vis-NIR-derived latent variables.

2. Data and methods

2.1. Study area

Denmark covers 45,000 km², and about 60% of the area is used for agriculture (Levin, 2025). Within these agricultural regions, around 87% is dedicated to intensive agriculture, roughly 1.5% is designated for permanent crops, and the remaining portion is dedicated to extensive agriculture, mainly focused on permanent grassland (Levin, 2025). During the winter, the mean temperature ranges from 1.5 °C in the interior of Jutland to 3 °C along the coasts, while the mean summer temperature varies from 15.5 °C to 17 °C. The annual precipitation ranges from 700 mm in the eastern islands to 850 mm in the western parts of the country (Wang, 2013). The elevation ranges from −9 to 170m above sea level, and the topography is generally flat. The present landscape is the product of glacial and glaciofluvial activity during the last two glaciations: The Saalian (350,000–130,000 years BP) and the Weichselian (115,000–9700 years BP) (Balstrøm et al., 2013). In the northern part of the country, the isostatic rebound following the ice age has lifted former marine sediments above sea level, and dunes cover many areas near the country's western coast. Our study area comprised the whole area of Denmark.

2.2. Point data

We mapped the latent variables generated by Norouzi et al. (2026), who applied dimensionality reduction (also known as data compression) to vis-NIR spectral measurements of 5607 topsoil samples (0–30 cm depth) collected across Denmark between 2018 and 2023. The spectra of sieved (<2 mm) samples were measured using a NIRS DS2500 spectrometer (FOSS), covering the wavelength range of 400–2500 nm at a resolution of 0.5 nm.

Norouzi et al. (2026) aimed to reduce this high-dimensional spectral dataset into a compressed latent space while preserving the key information needed to reconstruct the original spectra. To achieve this, they evaluated four methods: PCA, KPCA, Autoencoders, here referred to as *shallow* autoencoders (S-AE), and Convolutional Autoencoders (C-AE), representing both classical and deep learning approaches. For each method, latent representations of different sizes (5, 10, 15, and 20 variables) were generated.

All methods allow reconstruction of the original spectra from the latent space, enabling their evaluation based on information retention. Norouzi et al. (2026) calculated the Normalized Mean Squared Error (NMSE), which quantifies how accurately the latent variables reproduce the original spectral signal with minimal distortion or information loss (Mantripragada et al., 2022). Reconstruction errors for each method and latent dimensionality are presented in Supplementary Fig. S1. Overall, KPCA and the deep learning approaches outperformed PCA, with more pronounced differences at lower latent dimensionalities. Further details on spectral acquisition, preprocessing, and compression methods are provided in Norouzi et al. (2026).

Based on these results, five latent variables per method were selected for mapping in this study (Fig. 1). This choice was guided by NMSE, which indicated that this dimensionality captures significant spectral variation (with decreasing improvements when increasing the number of variables), while maintaining model simplicity (Supplementary Fig. S1). It should be noted that the same number of latent variables across methods does not imply equivalent information content. The latent variables are not comparable across methods as each compression technique captures spectral variation differently. We did not attempt to compare individual latent variables, but to assess how different compression approaches spatially represent soil spectral variation. Future work could explore selecting method-specific dimensionalities based on equivalent reconstruction thresholds.

In PCA and KPCA, the resulting latent variables are typically ordered according to the amount of variance that they explain in the original data. In contrast, deep learning methods do not provide directly comparable metrics for ordering latent variables. To ensure consistency in data processing and presentation, we ordered the latent variables for all methods based on their calculated variance within the dataset reported by Norouzi et al. (2026). Supplementary Fig. S2 shows biplots of latent variables 1 and 2 for each compression method, while Supplementary Fig. S3A presents the correlations between the latent variables generated by the four methods.

2.3. Environmental covariates

The environmental covariates comprised 32 spatial raster layers at 10 m spatial resolution containing climate variables, terrain parameters, and geological maps that represent SCORPAN factors such as climate, geomorphology, and parent material, respectively. The layers were masked to the extent of the land area in Denmark and projected to the World Geodetic System 1984 ensemble UTM zone 32N (EPSG: 25832).

The annual precipitation (Fig. 2A) and mean temperature were obtained from the CHELSA database (Karger et al., 2017) over the 1981–2010 period. The CHELSA climate variables used in this study represent long-term averages, also known as “climate normal” (1981–2010), which were selected to capture the prevailing climatic conditions that shape soil-forming processes rather than short-term variability. A more recent climatological normal was not available at comparable spatial resolution.

As terrain parameters, we used geomorphological attributes derived from a digital elevation model shown in Fig. 2B. As parent material information, we included a map of geo-regions (Adhikari et al., 2013) (Fig. 2C), and the parent material map (Jakobsen et al., 2015) (Fig. 2D). The categorical layers (geo-regions and parent material) were converted into indicator variables, where each layer indicated the presence/absence of each category. Supplementary Table S1 contains a detailed description of the environmental layers. Supplementary Fig. S4 shows the correlation plot calculated for the environmental covariates.

2.4. Methodological workflow

The modelling workflow in Fig. 3 follows a DSM approach, where environmental covariates were used as predictors to explain the spatial variation of the latent variables. This study focuses on covariate-based

modelling to better understand the environmental drivers of the latent variables and interpret the results in terms of soil-landscape relationships.

2.4.1. Step 1. Regression matrix generation

The point dataset from Norouzi et al. (2026) was integrated with the environmental covariates in Section 2.3 to construct a regression matrix used for subsequent modelling.

2.4.2. Step 2. Point declustering

The sampling points were densely concentrated in some areas due to the inclusion of intensive surveys in forested and agricultural areas. This represents an issue, as clustered samples might also be spatially correlated and would introduce bias in the accuracy assessment of predictive models (de Bruin et al., 2022; Just et al., 2020; Muhammad and Glass, 2023). To address spatial clustering in the dataset, we applied a sample declustering (weighting) strategy in combination with repeated cross-validation when training the model. This approach reduces the influence of densely sampled areas and bias associated with spatially auto-correlated observations, while preserving representativeness across the full range of environmental conditions. In contrast to strict spatial cross-validation approaches, which may lead to overly-pessimistic estimates due to strong spatial and environmental separation between training and validation data (Tziachris et al., 2023; Wadoux et al., 2021; Yoo and Koo, 2024). Our strategy can be viewed as an intermediate approach that balances the trade-offs between over-optimistic error estimates in random cross-validation and over-pessimistic error estimates in spatially constrained schemes.

Weighted machine learning (ML) may be used to reduce the adverse effects of spatial correlation in clustered points (Hashemi and Karimi, 2018; Hengl et al., 2022). In weighted machine learning, the bias in the sampling intensity is incorporated into the model training by giving more weight to observations in sparsely sampled areas and less weight to observations in densely sampled areas.

One approach to assigning weights divides the study area into regular square cells, and each measurement receives a weight that is inversely proportional to the number of measurements within the same cell (Deutsch, 1989). However, the grid position and size are arbitrarily defined over the landscape, and that has a strong effect on the location and number of points within cells (Lechner et al., 2009). Alternatively, a second approach employs a set of Voronoi regions around each sample point (Isaaks and Srivastava, 1989), which results in small areas for closely spaced points and large areas for sparsely arranged sample points. The weights applied are directly related to these areas. In this approach, the polygon may suffer from significant edge-effect problems, depending on how the Voronoi regions are computed (de Smith et al., 2024).

We used a hybrid method in which the areas were defined by grouping the 32 environmental covariates (described in Section 2.3) plus the X and Y-coordinates raster layers into clusters using the k-means algorithm (Hartigan and Wong, 1979). The k-means clustering was implemented using the *ClusterR* package in R (Mouselimis, 2024) with the covariate values extracted from a regular grid of 160,000 pixels. The algorithm was repeated 10 times with different centroid seeds, 5000 as the maximum number of clustering iterations, and *k-means++* as the method of initialization. The number of clusters tested ranged from 8 to 100. We selected eight clusters as a starting point, based on the proportion of variance explained (PVE) by clustering (Supplementary Fig. S4), which reached 90 % at eight clusters and was calculated as follows:

$$PVE = TSS - WCSS/TSS \quad (1)$$

where, TSS is the total sum of squares and WCSS is the within-clusters sum of squares. The closer to 1, the better the clustering.

After obtaining cluster centroids for 8 to 100 clusters, we calculated

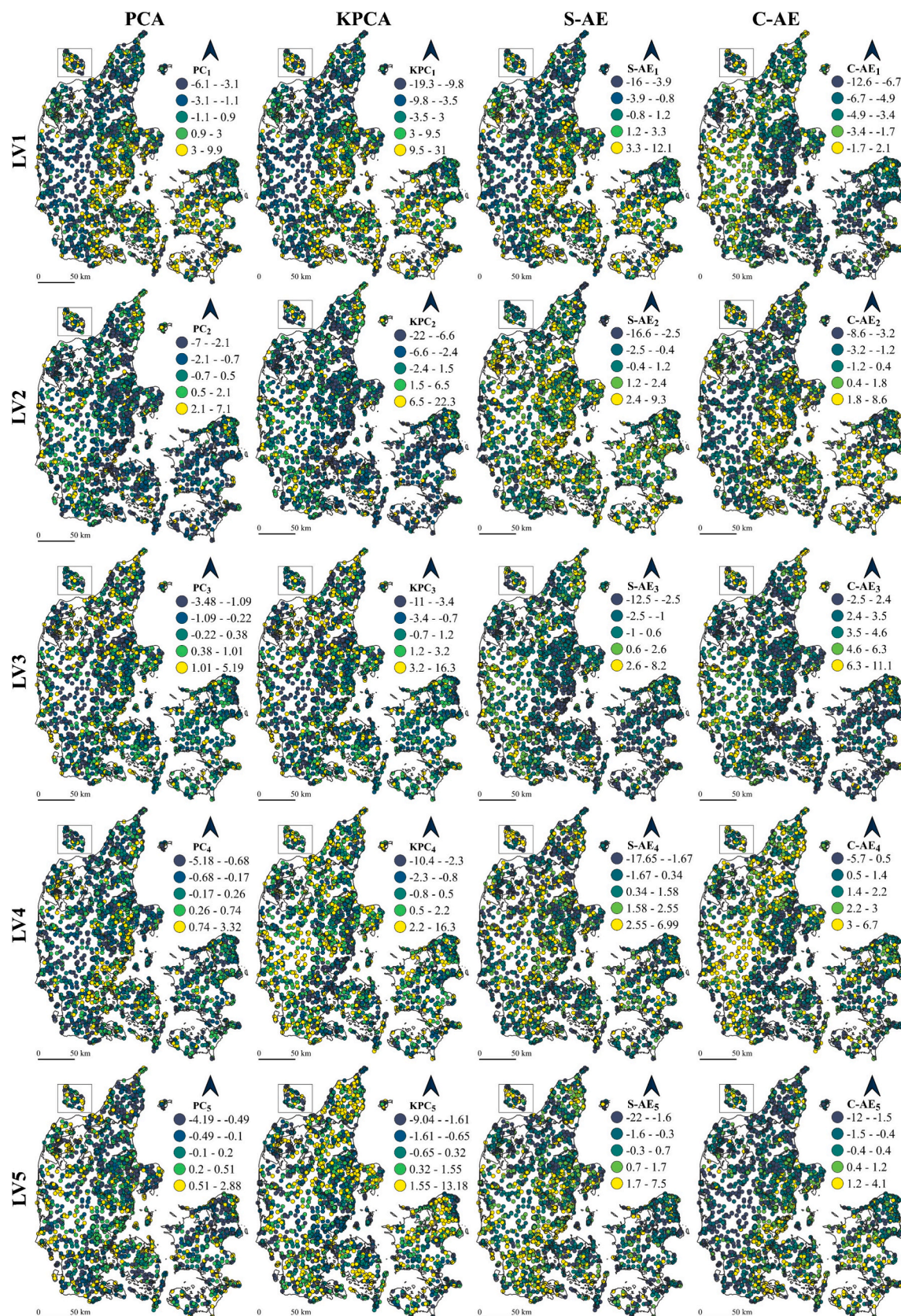


Fig. 1. Point distribution of the compressed vis-NIR spectral measurements. LV: latent variable. PC: principal component. KPC: kernel principal component. S-AE: shallow autoencoder. C-AE: convolutional autoencoder.

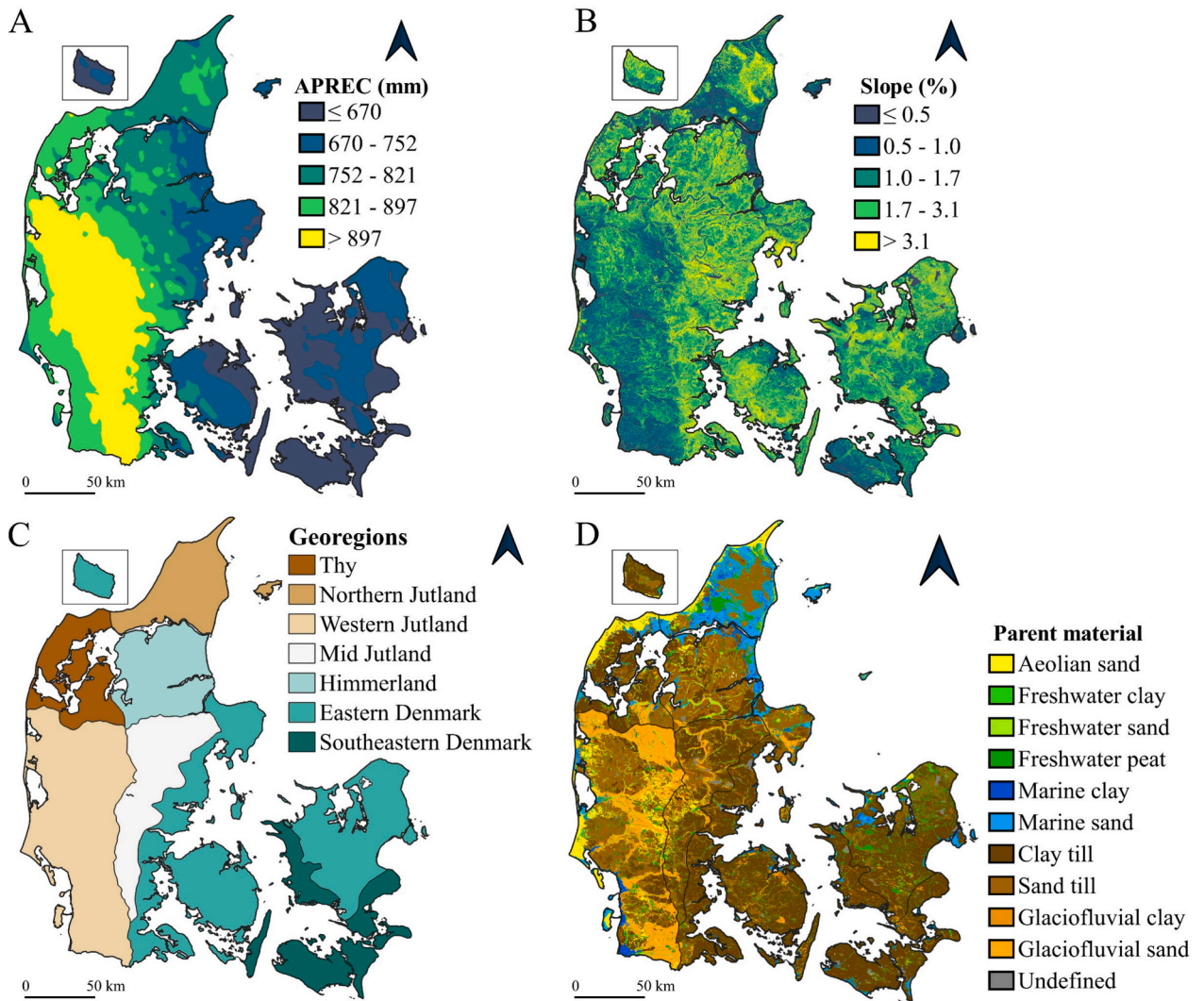


Fig. 2. Examples of covariates included in the study to map latent variables derived from vis-NIR compressed data. A) Annual precipitation in mm (APREC), B) slope in %, C) geo-regions, and D) parent material.

the Euclidean distance between each observation in the dataset and the centroids. Each observation was then assigned to the cluster with the nearest centroid. We then assigned each observation a weight inversely proportional to the number of points within its cluster. The observation weights from all clustering results were then averaged, and these values were incorporated into the model fitting process. This approach allowed us to weight observations according to their proximity in both geographic and environmental-covariate space. Our approach represents a compromise between completely random and spatial cross-validation. While this reduces some spatial dependence, it does not fully resolve the spatial autocorrelation among samples.

2.4.3. Step 3. Weighted machine learning modelling

We trained an eXtreme Gradient Boosting (XGBoost) model using weighted observations to predict our response variables with the *xgboost* package in R (Chen and Guestrin, 2016). The response variables were the latent variables obtained from the dimensionality reduction of the vis-NIR spectral data (Norouzi et al., 2026). For classical methods, these represent component scores, and encoded features for deep learning-based models. We scaled the latent variables to a range between 0 and 1. A separate model was developed for each latent variable from all compression methods. XGBoost builds a robust learner by combining multiple decision trees through a gradient-boosting framework (Chen and Guestrin, 2016). The algorithm iteratively improves predictions

by building new trees based on previous errors (Kumar et al., 2023; Rahbar et al., 2024; Zhang et al., 2022).

To optimize the model's hyperparameters (i.e., algorithm type, tree depth, observation weights, sample size, variables used, and learning rate), we performed a grid search using a ten-fold cross-validation repeated 10 times using the *caret* R package (Kuhn, 2008). The best-performing models were used to generate 10-m resolution maps for each latent variable.

Variable importance was derived from the trained models to assess the relative influence (percentage) of each predictor in reducing the error estimation (Friedman, 2001). We used the *vip* package in R (Brandon, 2020).

2.4.4. Step 4. Weighted accuracy evaluation

We calculated the model accuracy as the weighted RMSE ($RMSE_w$), coefficient of determination (R^2_w), and Lin's concordance correlation coefficient (CCC_w), as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n w_i (\hat{x}_i - x_i)^2}{\sum_{i=1}^n w_i (x_i - \mu_x^w)^2} \quad (2)$$

$$CCC_w = \frac{2\rho_w \sigma_x^w \sigma_x^w}{(\sigma_x^w)^2 + (\sigma_x^w)^2 + (\mu_x^w - \mu_x^w)^2} \quad (3)$$

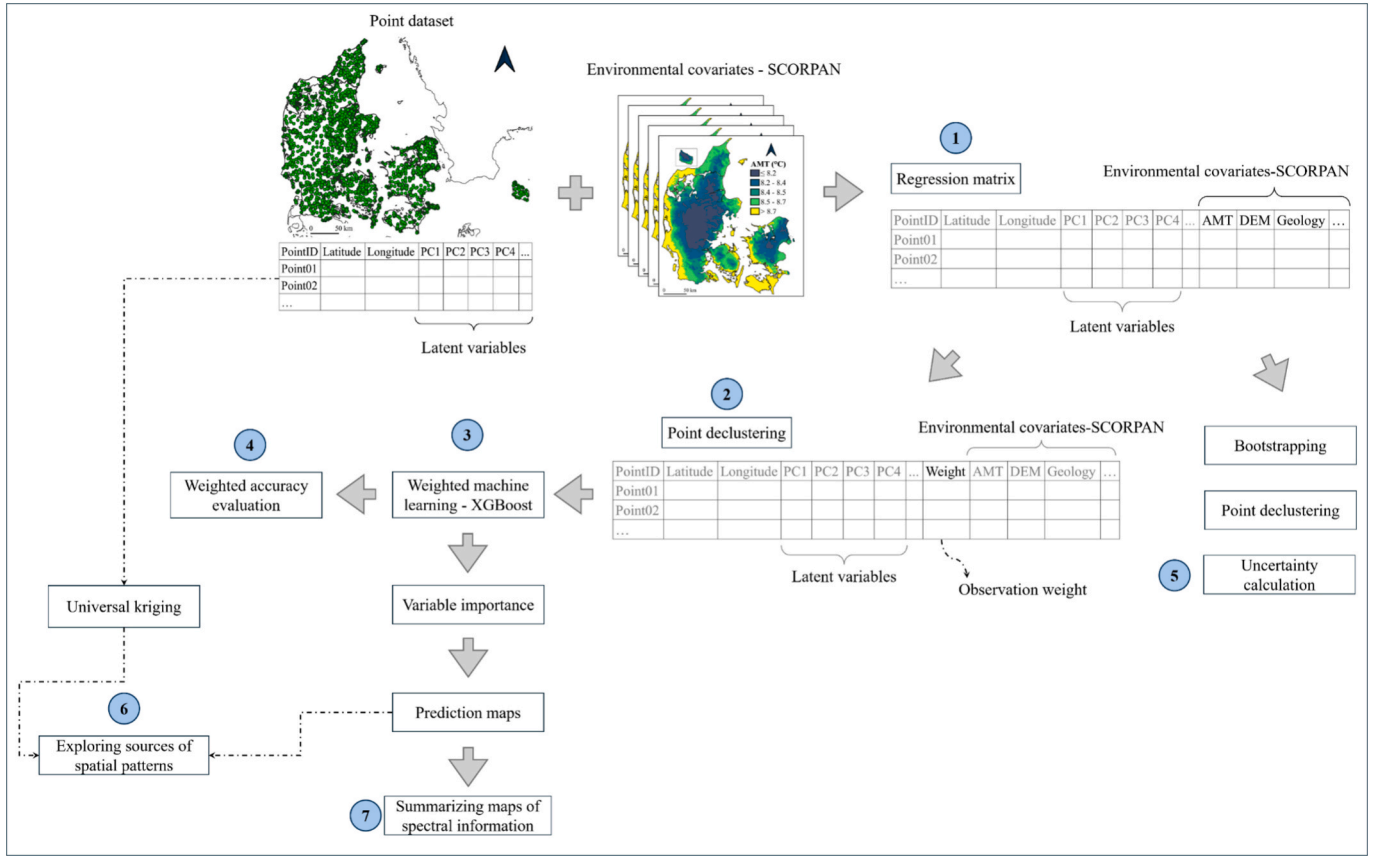


Fig. 3. Methodological workflow for modelling latent variables derived from vis-NIR compressed data.

$$RMSE_w = \sqrt{\frac{\sum_{i=1}^n w_i (\hat{x}_i - x_i)^2}{\sum_{i=1}^n w_i}} \quad (4)$$

where x_i is an observed value, $\hat{x}_i$ is a predicted value, x_i is an observed value, w_i is the weight for the observation i , μ_x^w is the weighted mean of the observed values, ρ_w is the weighted Pearson correlation between observed and predicted values, σ_x^w is the weighted standard deviation of predicted values, σ_x^w is the weighted standard deviation of the observed values, σ_x^w is the weighted standard deviation of the predicted values.

2.4.5. Step 5. Uncertainty maps calculation

XGBoost does not have the capacity to provide uncertainties for the predicted values, so we used a non-parametric bootstrap approach to estimate uncertainties for the target variables. This procedure generates a prediction distribution by repeatedly fitting models to randomly selected subsets of the training data (Malone et al., 2011; Viscarra Rossel et al., 2015). We took 50 bootstrap samples of 50% of the dataset for each variable, declustered the observations applying the procedure from Section 2.4, tuned the hyperparameters, and generated 50 spatial predictions using the trained XGBoost model. The choice of 50 bootstrap iterations and a sampling proportion of 50% was made as a trade-off between computational efficiency and stability of the uncertainty estimates, as preliminary testing indicated limited changes in uncertainty patterns with additional iterations.

By repeatedly sampling with replacement and performing spatial modeling on each bootstrap sample, we obtained probability distributions for the outcomes at each pixel. We estimated the 5th and 95th percentiles to calculate the 90% prediction interval as an uncertainty measure of our predictions. The prediction interval estimates presented in this study reflect variability associated with data availability (Zhu et al., 2025). We did not account for other sources of uncertainty,

such as positional error, model structural uncertainty, or uncertainty in the environmental covariates.

2.4.6. Step 6. Exploring sources of spatial patterns

We obtained prediction maps of the latent variables using universal kriging (UK) as well. The UK is a geostatistical method mostly used for mapping when there is a spatial trend in the data, and the dependent variable does not fulfill the criterion of weak stationarity. In UK, the mean of the response variable is a function of the spatial coordinates in linear, quadratic, or higher form (Goovaerts, 1997). We used ArcGIS Pro v.3.7.0 to fit semi-variogram models of the latent variables assuming a second-order trend of the spatial coordinates, calculate accuracy measurements by using a leave-one-out cross-validation, and generate prediction and standard error maps at 10-m resolution.

In this study, the primary focus was on the covariate-driven DSM framework to map the latent variables. UK approach was implemented as a complementary and exploratory analysis to evaluate the role of spatial autocorrelation. We did not intend to establish either the UK or DSM as a benchmark. Instead, we aimed to provide complementary perspectives on the spatial variability of the latent variables as the two approaches rely on different assumptions. UK explores the spatial autocorrelation of the measured latent variables through variogram modelling, whereas the ML approach incorporates environmental covariates to explain and predict spatial variation. As a result, ML predictions are driven by soil-landscape relationships, while UK reflects spatial structure inherent in the sampled data.

By having both sets of maps, we assess the extent to which spatial patterns are consistently captured across methods and identify areas of agreement and divergence. To further explore these differences, we computed the spatial difference between ML and UK predictions. This comparison provides insight into where environmental covariates effectively explain the spatial variability of the latent variables and

where spatial dependence not captured by the covariates may play a stronger role.

2.4.7. Step 7. Summarizing maps of spectral information in Danish soils

We created a red–green–blue (RGB) composite using the first three latent variable maps of each compression method and clustered the latent variable maps of the method with the highest accuracy to further investigate how the soil variation is represented in our prediction maps.

The RGB composites were created using the *Build Virtual Raster* tool (Raster Miscellaneous menu) in QGIS v.3.44.3. For each compression method, the first three latent variable maps were selected as input layers and assigned to separate bands using the “Place each input file into a separate band” option. We investigated how spatialized latent variables derived from vis-NIR spectral data can represent variation in soils at the national scale and how they relate to legacy soil maps from Denmark. We used the k-means clustering algorithm (Hartigan and Wong, 1979) with the values extracted from the latent variable maps derived from compression using CA-E, using the same regular grid of 160,000 pixels as in section 2.4.

We selected the number of clusters based on the elbow method over the curve of the sum of within-cluster variance with respect to the number of clusters (Han et al., 2012). The elbow, or optimal number of clusters, is where the within-cluster dispersion does not change drastically with the addition of another cluster (Roell et al., 2020).

After the k-means algorithm, we applied an agglomerative hierarchical cluster algorithm to group the centroids of the clusters obtained before to further assess their similarities, as proposed in Román Dobarco et al. (2023) and Styc et al. (2025). The similarity between cluster

centroids was estimated with Ward’s method, which minimizes the total within-cluster variance. Ward’s method merges the pair of clusters with the minimum between-cluster distance at each successive step. The resulting dendrogram was divided into branches, and different color ramp scales were assigned to enable the interpretation of the spatial distribution of similar classes.

To support the interpretation of spatial patterns in the latent variable maps, we compared the resulting RGB composites and cluster maps with existing reference datasets, including maps of glacial deposits (Houmark-Nielsen, 2011), landscape elements (Madsen et al., 1992), and soil texture fractions (Møller et al., 2025) (Fig. 4). This comparison enabled an evaluation of how the latent variables relate to known soil-forming factors and soil properties.

3. Results

3.1. Latent variable maps and accuracy measures

Overall, each compression method shows distinct regional patterns across Denmark, reflecting both broad national-scale gradients and localized spatial heterogeneity (Fig. 5). Notably, Western Jutland consistently exhibited greater local variability across all methods. Prediction uncertainty was generally low for all methods (Supplementary Fig. S5), although the widest prediction intervals and strongest patterns of local variation were observed in the Western Jutland and South-eastern Denmark regions. Latent variables derived from C-AE were characterized by larger prediction intervals and greater local variation compared to those obtained with the other methods.

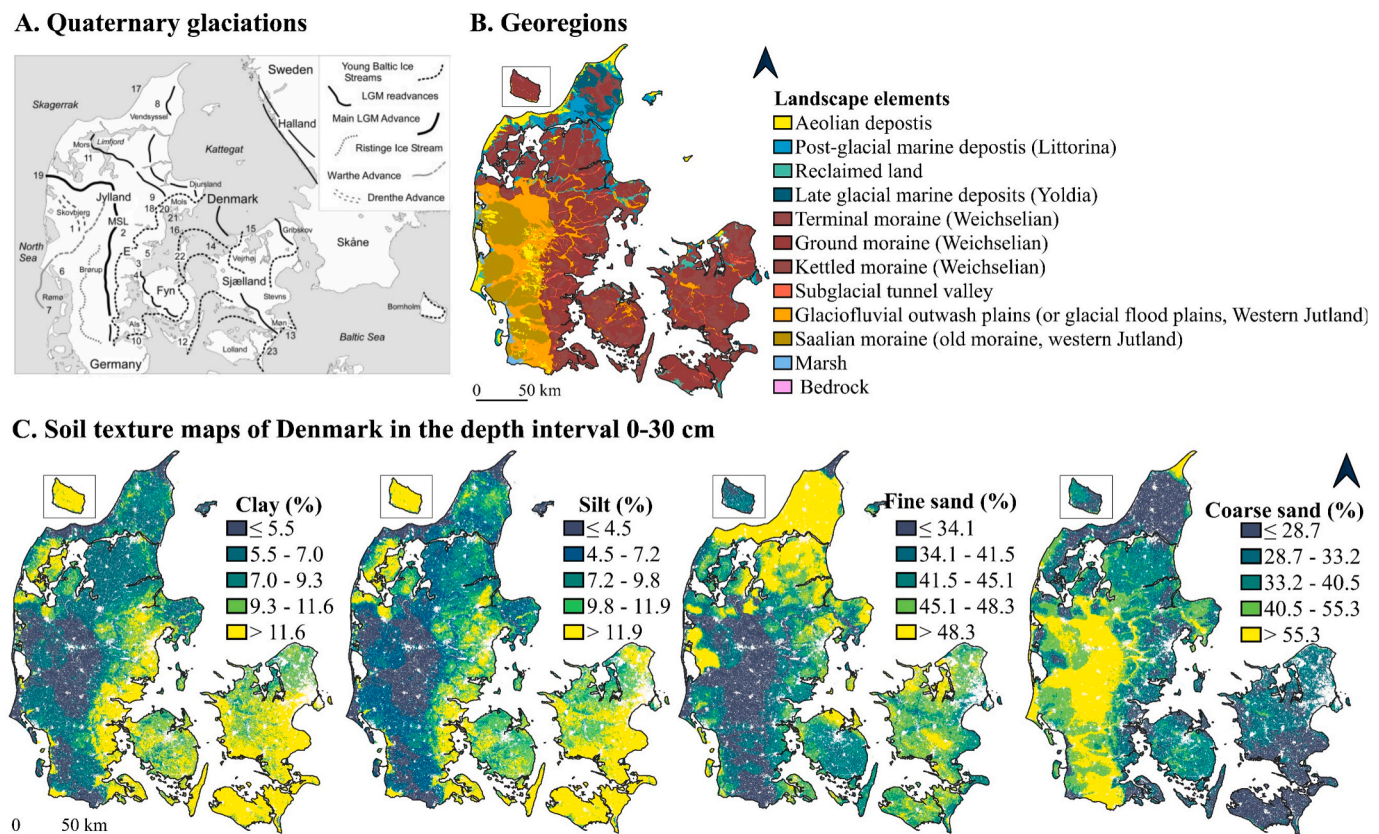


Fig. 4. Legacy maps and published soil data from Denmark. A) Extent and chronology of Quaternary glaciations in Denmark, adapted from Houmark-Nielsen (2011). The numbers refer to key sites; 1: Harreskov; 2: Sønder Vissing; 3: Treldehø; 4: Røgle; 5: Ashoved; 6: Esbjerg; 7: Lister Deep; 8: Skærumhede; 9: Hinnerup; 10: Stensigmoose-Gammelmark; 11: Sundsøre; 12: Ristinge; 13: Klintholm; 14: Sejersø; 15: Hundested; 16: North Samsø; 17: Lønstrup; 18: Fårup; 19: Bovbjerg; 20: Højvang; 21: Mols Hoved; 22: Visborg and 23: Korselitte. For more detailed information on this figure, the reader is referred to Houmark-Nielsen (2011). B) Landscape elements map adapted from Madsen et al. (1992). C) High-resolution (10 m) digital soil texture, showing the spatial distributions of the predicted fractions in the depth interval 0–30 cm, adapted from Møller et al. (2025).

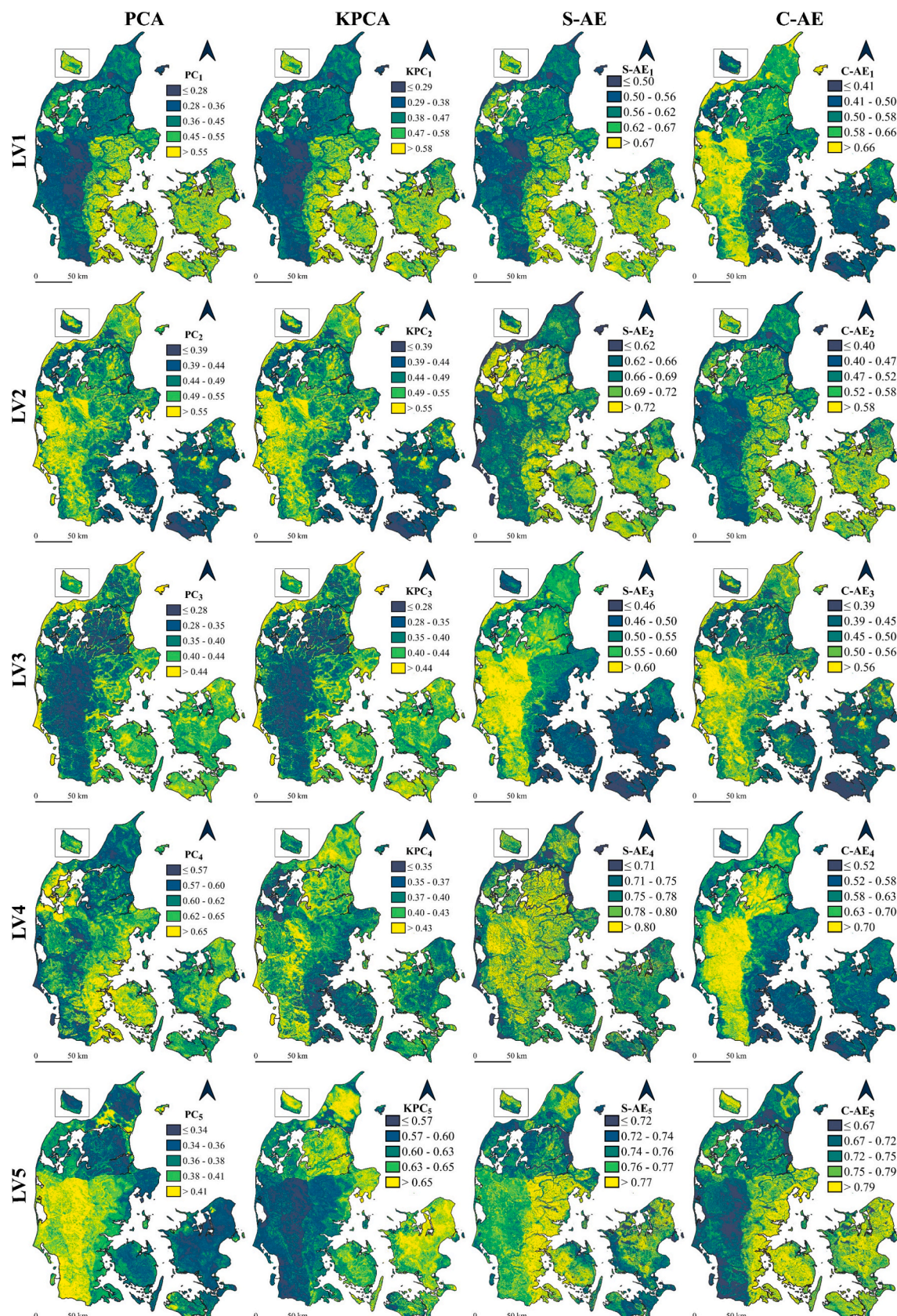


Fig. 5. Prediction maps of the compressed vis-NIR spectral measurements. LV: latent variable. PC: principal component. KPC: kernel principal component. S-AE: shallow autoencoder. C-AE: convolutional autoencoder.

For the first three latent variables, all methods achieved comparable weighted accuracy values (Table 1). Differences became more visible when examining the fourth and fifth latent variables. Models for latent variables from PCA and KPCA had a sharp decline in performance for the fourth and fifth latent variables, while the predictions for the same latent variables from S-AE and C-AE had higher performance.

On average, the accuracy was higher for latent variables obtained by compressing vis-NIR spectral data using deep learning methods.

It should be noted that the accuracy measures reflect the adopted strategy combining sample declustering with repeated cross-validation. As presented in Section 2.4, this approach represents an intermediate point between purely random and strictly spatial validation schemes. Consequently, the reported accuracy metrics should be interpreted with this context in mind.

3.2. Driving factors

We examined the importance of the predictors grouped by SCORPAN factor (climate, relief, and parent material) across all the latent variables within each compression method (Fig. 6). Climate variables ranked among the most important factors across all four compression methods. Topographical parameters, including vertical distance to the channel network (VDtoCN), wetness index (WI), and cost distance (CostDist) also showed strong predictive power. Climate and topographic variables related to water movement in soil were followed in importance by parent material.

The relative ordering of variables and SCORPAN-factor importance was broadly consistent across models. We further examined how examples of individual numerical predictors influenced the latent variables within each compression method. The partial dependence plots (PDPs) (Fig. 7) show the marginal effects of annual precipitation and vertical distance to the channel network on four representative latent variables (PC5, KPC1, S-AE4, and C-AE3). These predictors were not correlated as shown in the correlation plot in Supplementary Fig. S4. These latent variables were selected because their prediction maps (Fig. 5) exhibit contrasting spatial patterns and because they show low mutual correlation (Supplementary Fig. S3B). The full set of PDPs for the same predictors across all latent variables and compression methods is presented in Supplementary Fig. S7.

The PDPs reveal marked interactions between the two predictors, with each latent variable responding differently to precipitation and vertical distance to the channel network. PC5 increases under high annual precipitation and low vertical distance to the channel network. In contrast, KPC1 shows higher predicted values under low precipitation. C-AE3 displays the opposite pattern, with its highest values associated with high precipitation. S-AE4 also increases with annual precipitation, while low vertical distance to the channel network is consistently linked to low predicted values, regardless of precipitation level. Overall, we observed that these examples of numerical predictors differently influenced the predicted values of latent variables across the compression methods.

Similarly, we explored examples of categorical predictors (extent of

clay till and freshwater deposits, and the geological settings of Western Jutland-Georegion 03) and how these influenced the latent variables for each compression method (Fig. 8). The influence of these categorical predictors varied across latent variables and compression approaches. For example, soils derived from clay till deposits were associated with higher predicted values for PC1, KPC1, and S-AE1, whereas C-AE1 showed the opposite pattern. In contrast, latent variables 3 and 5 across all compression methods consistently showed lower predicted values for soils developed in the geo-region corresponding to Western Jutland.

3.3. Exploring complementary sources of spatial patterns

Overall, the differences between the UK- and ML-based maps were small across all latent variables and compression methods (Fig. 9), and we found no systematically distributed differences. An exception was the fifth latent variable from PCA and KPCA, which showed a distinct spatial structure, with generally smaller differences in western Jutland. Cross-validation statistics for the UK models are shown in Supplementary Table S2.

Regarding uncertainty, the two approaches reflect different sources. In the ML models, uncertainty is derived from bootstrap resampling and represents model variability related to the input data (Supplementary Fig. S5). In contrast, UK uncertainty is expressed as kriging standard error, linked to the fitted variogram model (Supplementary Fig. S8). Despite these differences, both approaches showed similar patterns, with higher uncertainty in western Jutland and lower values in the eastern part of the Jutland peninsula.

3.4. Summary of the spectral information in Danish soils

Fig. 10A-D presents RGB composite maps integrating the spatial distribution of the first three latent variables derived from each compression method. In each composite, the three latent variables were assigned to the red, green, and blue channels, respectively, highlighting multidimensional patterns in soil spectral variation. Pixels with similar colors share similar characteristics in terms of values derived from the compressed vis-NIR spectral data, whereas color differences indicate dissimilarities in the underlying spectral features.

The optimal number of clusters is shown in Supplementary Fig. S9. Using the hierarchical clustering approach, we grouped the similar cluster centroids into larger clusters, referred in the figure as to *super-clusters*, as shown in the dendrogram in Supplementary Fig. S10.

Maps in Fig. 10A-D highlight different aspects of spectral variation, from broad gradients to local patterns. Despite these differences, a consistent structure distinguishes Western and Eastern Denmark. This is also reflected in the cluster map (Fig. 11E), where 13 clusters form three larger clusters: western Denmark, eastern Denmark, and transitional areas.

These patterns align well with existing soil and geological maps (Fig. 4). The sharp west-east transition corresponds to the Main Stationary Line (MSL; Fig. 4A), marking the extent of the Weichselian glaciers. Supercluster 1 is associated with coarse-textured soils and

Table 1

Weighted accuracy measures of the latent variable models for the three compression methods. LV: latent variable. PCA: principal component analysis. KPCA: kernel principal component analysis. S-AE: shallow autoencoder. C-AE: convolutional autoencoder. R_w^2 : weighted coefficient of determination. CCC_w : weighted Lin's concordance correlation coefficient. $RMSE_w$: weighted root mean squared error.

Latent variable	R_w^2				CCC_w				$RMSE_w$			
	PCA	KPCA	S-AE	C-AE	PCA	KPCA	S-AE	C-AE	PCA	KPCA	S-AE	C-AE
LV ₁	0.45	0.42	0.32	0.46	0.62	0.61	0.49	0.63	0.15	0.16	0.07	0.13
LV ₂	0.4	0.39	0.46	0.43	0.58	0.56	0.63	0.61	0.13	0.13	0.08	0.12
LV ₃	0.43	0.43	0.30	0.42	0.6	0.6	0.47	0.60	0.11	0.11	0.09	0.13
LV ₄	0.18	0.17	0.43	0.37	0.3	0.31	0.61	0.56	0.1	0.1	0.08	0.10
LV ₅	0.21	0.19	0.41	0.39	0.36	0.34	0.60	0.57	0.08	0.09	0.08	0.08
Mean value	0.31	0.32	0.38	0.41	0.49	0.48	0.56	0.59	0.11	0.12	0.08	0.11

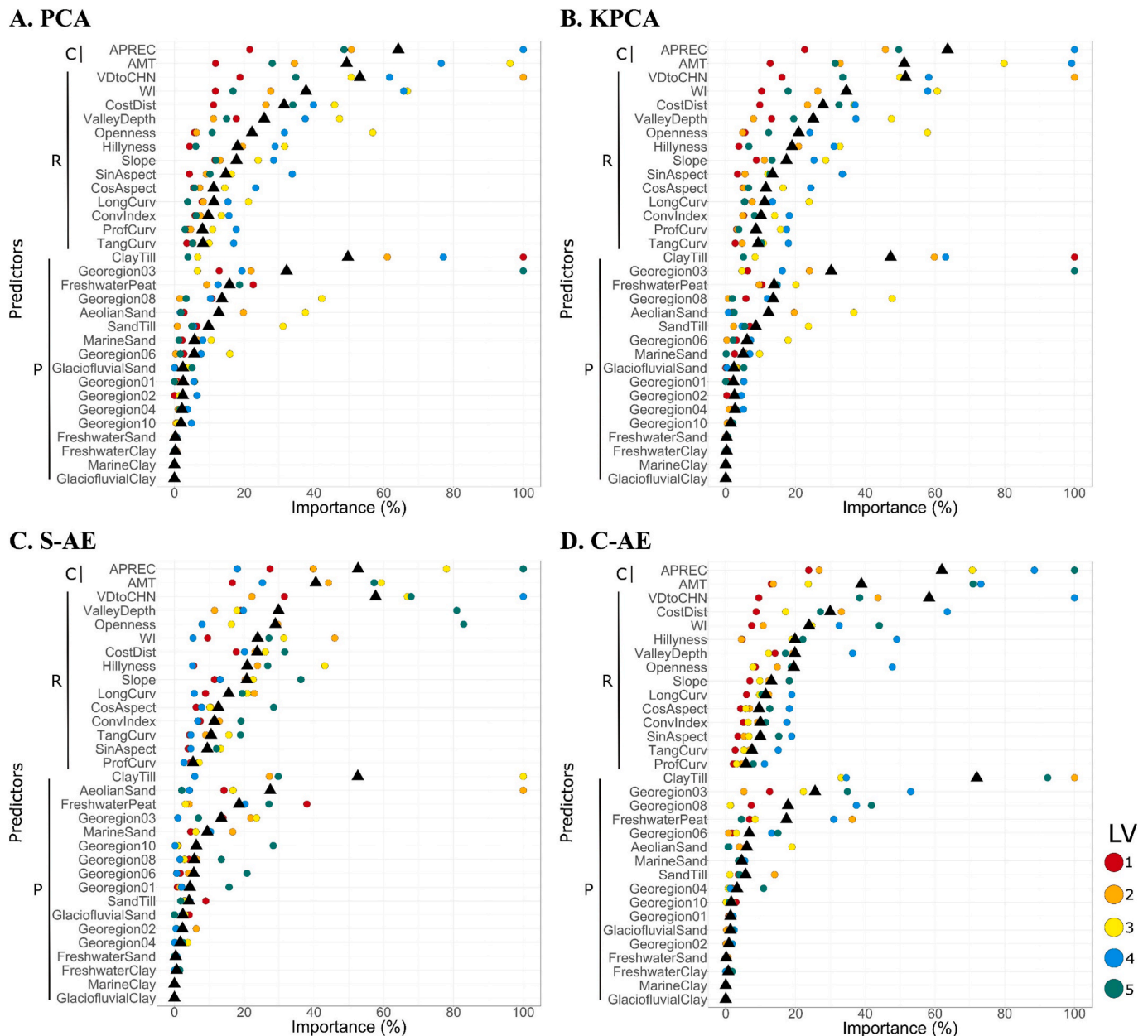


Fig. 6. Summary of predictor importance grouped by SCORPAN factor for modeling latent variables derived from A) PCA, B) KPCA, C) S-AE, and D) C-AE. Letters next to the predictor names represent SCORPAN factors (C: climate, R: relief, P: parent material). Each point represents the importance of a predictor for an individual latent variable represented in a color scheme. The triangles indicate the mean importance of each predictor, calculated across the five latent variables for each method. LV: latent variable. PCA: principal component analysis. KPCA: kernel principal component analysis. S-AE: shallow autoencoder. C-AE: convolutional autoencoder. Names and descriptions of predictors can be seen in the table listing the environmental layers used in the study (Supplementary Table S1).

glaciofluvial outwash plains (Fig. 4B), whereas supercluster 2 includes both sandy and finer-textured soils similar in extent to maps in Fig. 4C. Supercluster 3 represents sparse areas linked to subglacial tunnel valleys (Fig. 4B).

4. Discussion

4.1. Importance of the latent variable maps

Our approach incorporated environmental layers to explain the spatial variation of compressed vis-NIR spectral data across Denmark. The resulting maps are valuable as they provide an integrative representation of soil spectral variation in relation to climate, relief, and parent material. Moura-Bueno et al. (2021) showed that integrating

environmental variables with soil spectral data improves understanding of soil-landscape systems by accounting for factors not captured in spectral data alone.

The maps produced with a weighted ML algorithm preserved local patterns, and the differences between the UK and ML prediction maps were minimal and showed no clear spatial patterns. In this context, the consistency of spatial patterns observed across modelling approaches, including the comparison between ML and UK maps, suggests that the dominant spatial structures captured by the models are robust. This supports the use of the latent variables as meaningful representations of soil spectral variation at the landscape scale. Future work should evaluate the sensitivity of model performance to alternative validation strategies to better understand their implications for spatially structured soil datasets.

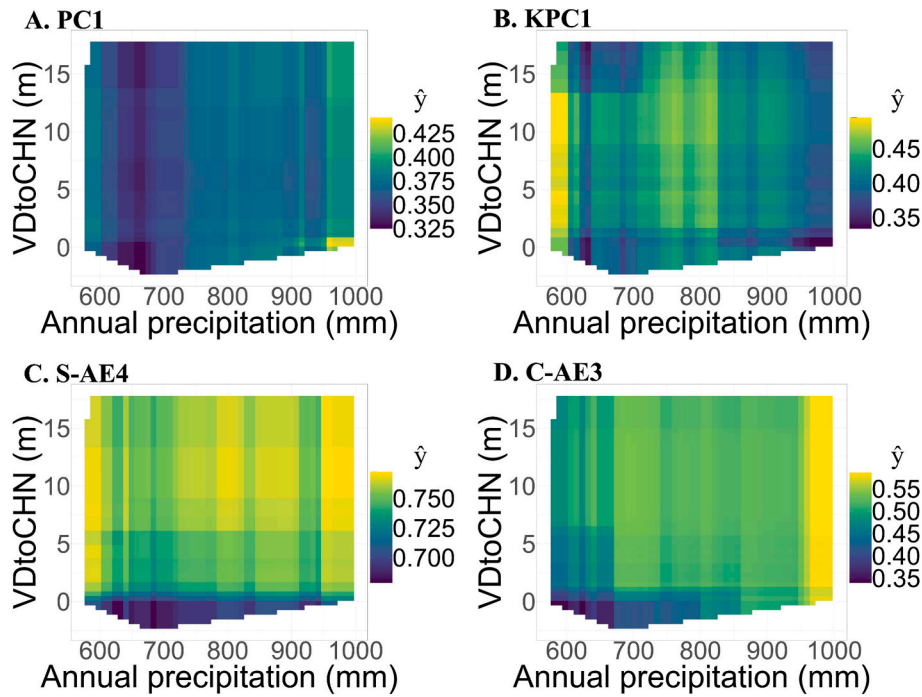


Fig. 7. Partial dependency plots for the most important numeric predictors to model the latent variables ($\hat{y}$) for each compression method.

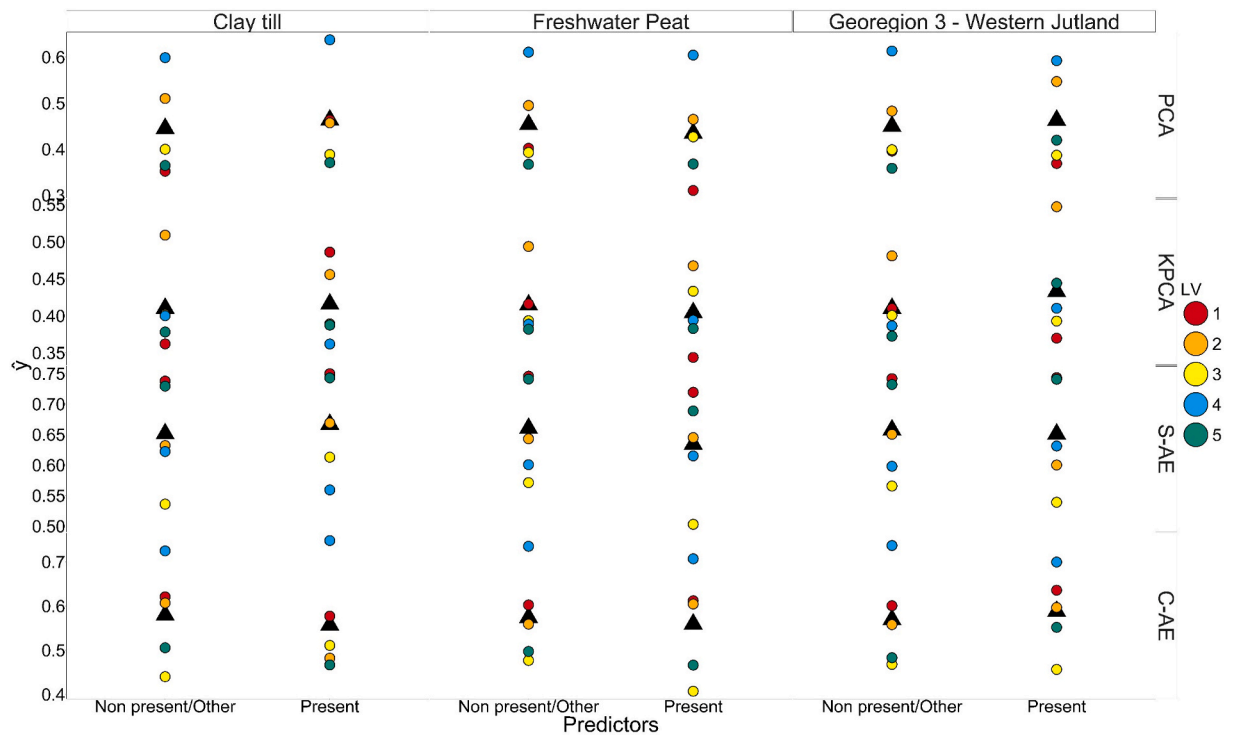


Fig. 8. Partial dependency plots showing the influence of the most important binary predictors on the latent variables ($\hat{y}$) for each compression method. Dots represent the dependency of individual categorical predictors on the latent variables, while triangles indicate the average predicted response across the five models within a given compression method. LV: latent variable.

4.2. Driving factors of soil vis-NIR spectral variation

Knadel et al. (2013) interpreted PCA-based latent variables using loadings to link spectral bands to soil properties. However, such interpretation is not directly feasible for KPCA or autoencoder-based methods due to their non-linear transformations. Instead, we used

variable importance from the XGBoost model to examine how environmental predictors explain the latent variables.

Across all methods, climate, topographical variables related to water movement, and parent material related to the latent variables (Fig. 5). Parent material influences spectral responses through its mineral composition and degree of weathering, which are reflected in vis-NIR

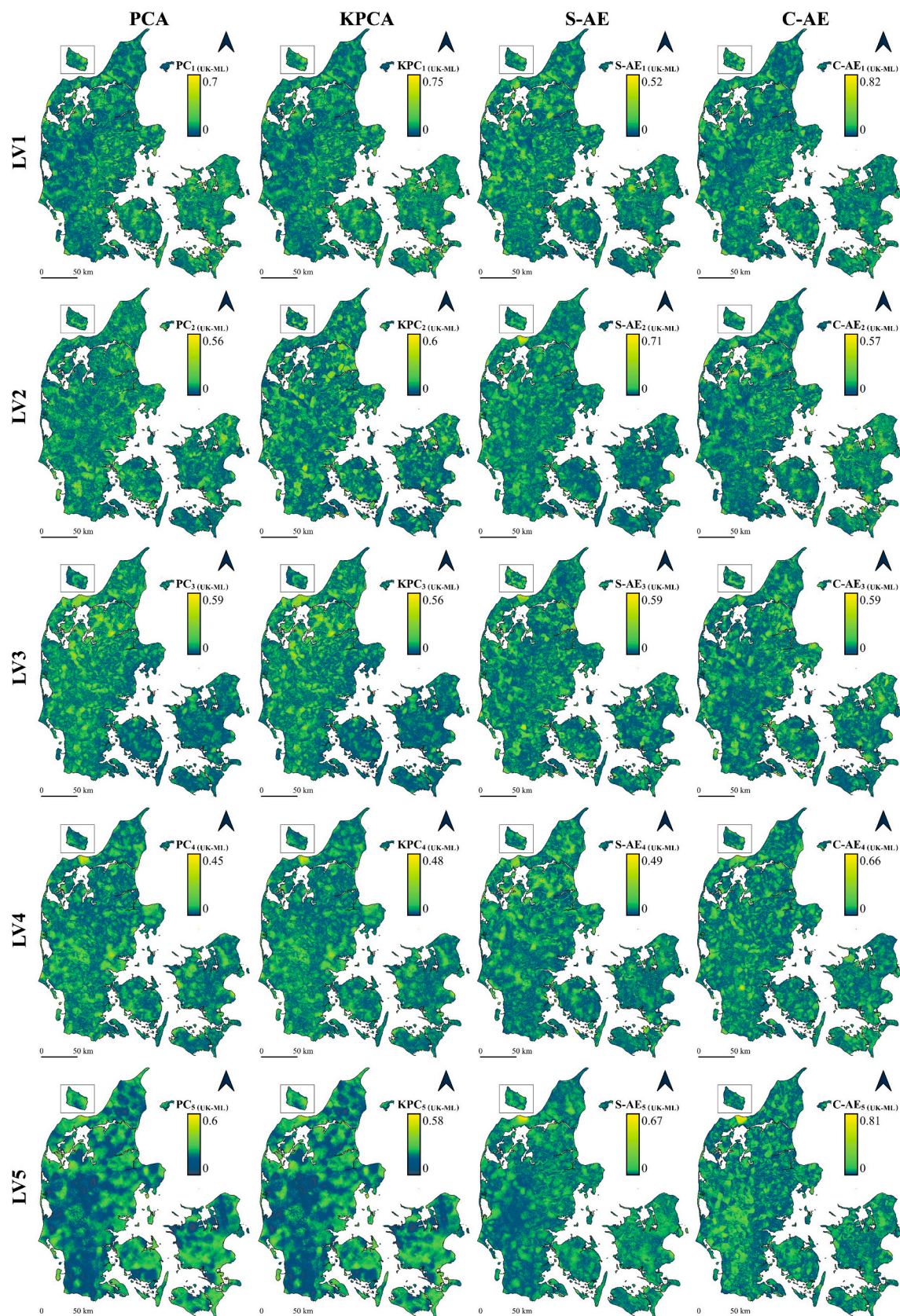


Fig. 9. Maps showing the absolute difference between prediction maps obtained using universal kriging and prediction maps obtained using the weighted machine learning model. LV: Latent Variable. UK: Universal Kriging. ML: Machine Learning. PC: Principal Component. KPC: Kernel Principal Component. S-AE: Shallow Autoencoder. C-AE: Convolutional Autoencoder.

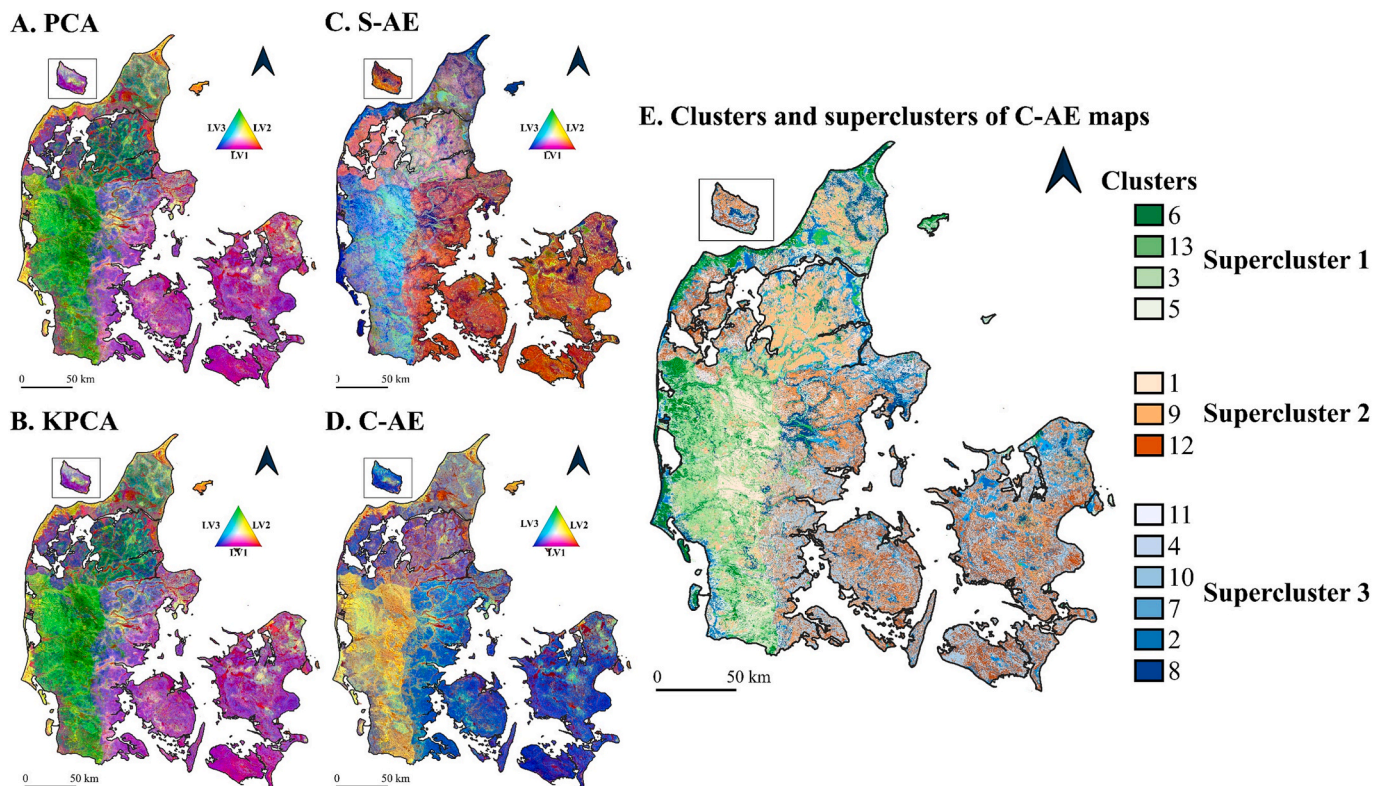


Fig. 10. RGB composite maps of Denmark showing the spatial distribution of the first three latent variables (LV1 = red, LV2 = green, LV3 = blue) derived from the compressed vis-NIR spectral data using the four methods: A) PCA, B) KPCA, C) S-AE, and D) C-AE. Similar colors between pixels in the RGB maps indicate similarity in the corresponding latent space. E) Cluster of maps produced with the five latent variables generated with C-AE. LV = latent variable. PCA: principal component analysis. KPCA: kernel principal component analysis. S-AE: shallow autoencoder. C-AE: convolutional autoencoder.

absorption features. Danish landscapes are shaped by glacial and glaciofluvial processes during the Saalian and Weichselian glaciations, resulting in heterogeneous deposits with contrasting lithology, texture, and mineralogy (Balström et al., 2013). Subsequent pedogenic processes, climate, and land use have further modified these materials, reinforcing their influence on soil spectral variation. Similar relationships between soil spectra, texture, and glacial history have been reported by Knadel et al. (2013) and Demattè and da Silva Terra (2014).

The large-scale spatial patterns observed in Fig. 10 reflect this glacial history, particularly the contrast between western and eastern Denmark. This transition aligns with the Main Stationary Line (MSL; Fig. 9A), marking the maximum extent of the Weichselian glaciers and the shift in depositional environments that drives differences in soil texture.

At finer scales, the spatial variability and clustering patterns (Fig. 10) correspond well with landscape elements (Fig. 4B). Western Jutland is dominated by coarse-textured soils derived from glaciofluvial sands and sandy tills, forming extensive outwash plains. Coastal areas show additional differentiation linked to aeolian and marine sands. In contrast, eastern Denmark is characterized by loamy and finer-textured deposits associated with Weichselian glaciation. Narrow linear features correspond to subglacial tunnel valleys, reflecting localized geomorphological control.

These geological and environmental factors also influence soil development. In western Denmark, higher precipitation and sandy deposits promote the formation of Arenosols and Podzols through enhanced leaching. In contrast, the warmer temperatures and more seasonally variable moisture conditions in eastern Denmark favor the development of Luvisols on finer-textured soils (Adhikari et al., 2014; Madsen et al., 1992).

4.3. Potential use for soil health assessment

Advances in sensing technologies and computational methods are rapidly expanding soil spectral datasets, enabling new applications in DSM. Considering that spectral information integrates chemical, physical, and hydrological properties, our maps can support the prediction of soil properties as these reflect both intrinsic soil characteristics and environmental controls. Although, Peng et al. (2015) modeled SOC using spatialized spectral data at regional scale in Denmark. Further studies are needed to explore the utility of our layers in predicting soil properties at multiple scales and across different land uses.

The spatial patterns represented in the clustering map reflect the combined influence of soil-forming factors, including climate, relief, and parent material, as captured through spectral signatures. From a conceptual perspective, they can potentially be interpreted as spatially continuous entities (pedo-spectral units) derived from compressed soil spectral information that may capture the joint variation of multiple soil attributes. Previous studies in Denmark have used spatial layers derived from vis-NIR spectral data (Knadel et al., 2013), combined with environmental covariates, to derive spectral-wise homogeneous soil zones (Peng et al., 2020; Peng et al., 2021; Roell et al., 2020). This might be relevant for initiatives such as the EU's Soil Monitoring and Resilience Law (European Commission, 2023), which aims to assess soil health using multiple indicators within soil units. Future research could explore the pedo-spectral units to support soil unit delineation for Denmark in compliance with the Soil Monitoring Law.

Despite these opportunities, challenges remain. Although spatialized vis-NIR data might improve predictive performance in soil properties modelling, this often comes at the cost of reduced interpretability. Deep learning-based compression approaches and the machine learning models used for spatial prediction introduce additional complexity, making it difficult to fully explain the observed spatial patterns. Future

work should therefore focus on enhancing model interpretability and developing methods to better link latent representations to specific soil properties and processes.

Finally, combining different spectral regions may hold additional value. Portable X-ray fluorescence (pXRF) has been used to predict CEC and EC (Silva et al., 2021) and to characterize parent material and weathering stages (Gozukara et al., 2021). Mid-infrared (MIR) spectroscopy can offer superior accuracy for certain properties (Gozukara et al., 2025). For instance, Helfenstein et al. (2021), employed MIR to improve SOC prediction in temperate peatlands. Laser-induced breakdown spectroscopy (LIBS) has outperformed vis-NIR spectroscopy for SOC and sand content prediction at continental scales (Wangeci et al., 2024).

5. Conclusions

This study spatially modelled compressed vis-NIR spectral data of Danish topsoils with environmental covariates at 10-m resolution. Using an XGBoost model, we produced prediction and uncertainty maps of latent variables.

Each compression method highlighted different aspects of soil-landscape relationships, as reflected in the RGB composites, cluster map, and predictor importance analyses. Climate, topography-related variables controlling soil water movement, and parent material extent consistently emerged as important drivers of spectral variation, distinguishing coarse-textured and aeolian sands in Western Jutland and coastal areas from the finer loamy and clayey soils of Eastern and Southeastern Denmark, as well as organic carbon-rich soils in Northern Jutland.

Beyond producing spatial datasets, this work provides a methodological and conceptual framework for incorporating compressed spectral information into DSM. By demonstrating how latent variables can act as integrative, spatially explicit descriptors of soil variation, the study advances the understanding of how vis-NIR spectral data can be used to capture soil-landscape relationships. This places the mapping of compressed vis-NIR spectral as a promising tool for future research and soil monitoring applications.

6. Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Ahrefs (a Free AI Sentence Rewriter Tool) in order to enhance the consistency of language and improve the readability of the Introduction section. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Sebastian Gutierrez: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Mogens H. Greve:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Lucas Carvalho Gomes:** Writing – review & editing, Visualization, Methodology, Data curation. **Anders B. Møller:** Writing – review & editing, Visualization, Methodology, Data curation. **Anne-Cathrine S. Danielsen:** Writing – review & editing, Visualization, Data curation. **Peter L. Weber:** Writing – review & editing, Visualization, Data curation. **Charles Pesch:** Writing – review & editing, Data curation. **Maria Knadel:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Deividas Mikstas:** Writing – review & editing, Visualization, Data curation. **Amelie Beucher:** Writing – review & editing, Visualization, Methodology, Conceptualization. **Marzieh Zaresourmanabad:** Writing – review & editing, Visualization, Data curation. **Per Møldrup:** Writing – review & editing, Visualization, Conceptualization. **David A. Robinson:** Writing – review & editing, Visualization, Conceptualization. **Lis W. de Jonge:**

Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Sarem Norouzi:** Writing – review & editing, Supervision, Methodology, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This study is funded by the European Union Horizon Europe research and innovation program under grant agreement No. 101086179 (AI4SoilHealth). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Appendix A. Supplementary data

This file contains supplementary figures and tables supporting the findings presented in the main manuscript. The online version can be found at <https://doi.org/10.1016/j.geoderma.2026.118002>.

Data availability

Data will be made available on request.

References

- Adhikari, K., Kheir, R.B., Greve, M.B., Bøcher, P.K., Malone, B.P., Minasny, B., McBratney, A.B., Greve, M.H., 2013. High-Resolution 3-D Mapping of Soil Texture in Denmark. *Soil Science Society of America Journal* 77 (3), 860–876.
- Adhikari, K., Minasny, B., Greve, M.B., Greve, M.H., 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. *Geoderma* 214–215, 101–113.
- Balstrøm, T., Breuning-Madsen, H., Krüger, J., Jensen, N.H., Greve, M.H., 2013. A statistically based mapping of the influence of geology and land use on soil pH. *Geoderma* 192, 453–462.
- Beucher, A., Peng, Y., Knadel, M., Greve, H.M., 2017. Supplementing predictive mapping of acid sulfate soil occurrence with Vis-NIR spectroscopy, *Pedometrics Conference*, Wageningen, The Netherlands.
- Blaesbjerg, N.H., Weber, P.L., de Jonge, L.W., Møldrup, P., Greve, M.H., Arthur, E., Knadel, M., Hermansen, C., 2022. Water repellency prediction in high-organic Greenlandic soils: comparing vis-NIRS to pedotransfer functions. *Soil Sci. Soc. Am. J.* 86 (3), 643–657.
- Brandon, M., 2020. Variable importance plots—an introduction to the vip package. *The R Journal* 12 (1).
- Brodský, L., Vašát, R., Klement, A., Zádorová, T., Jaksík, O., 2013. Uncertainty propagation in VNIR reflectance spectroscopy soil organic carbon mapping. *Geoderma* 199, 54–63.
- Camargo, L.A., Marques, J., Barrón, V., Alleoni, L.R.F., Barbosa, R.S., Pereira, G.T., 2015. Mapping of clay, iron oxide and adsorbed phosphate in Oxisols using diffuse reflectance spectroscopy. *Geoderma* 251–252, 124–132.
- Chen, S., Saby, N.P.A., Martin, M.P., Barthès, B.G., Gomez, C., Shi, Z., Arrouays, D., 2023. Integrating additional spectroscopically inferred soil data improves the accuracy of digital soil mapping. *Geoderma* 433.
- Chen, T., Guestrin, C., 2016. XGBoost: A Scalable tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.
- Clark, R.N., 1999. *Spectroscopy of Rocks and Minerals, and Principles of Spectroscopy*. John Wiley & Sons, Chichester, pp. 3–58.
- de Bruin, S., Brus, D.J., Heuvelink, G.B.M., van Ebbenhorst Tengbergen, T., Wadoux, A. M.J.C., 2022. Dealing with clustered samples for assessing map accuracy by cross-validation. *Eco. Inform.* 69.
- de Smith, M., Goodchild, M., Longley, P., 2024. Data exploration and spatial statistics. In: M. de Smith, M. Goodchild, P. Longley (Eds.), *Geospatial analysis. A comprehensive guide to principles techniques and software tools*. London, pp. 236–329.
- Dematté, J.A.M., da Silva Terra, F., 2014. Spectral pedology: a new perspective on evaluation of soils along pedogenetic alterations. *Geoderma* 217–218, 190–200.
- Deutsch, C., 1989. DECLUS: a fortran 77 program for determining optimum spatial declustering weights. *Comput. Geosci.* 15 (3), 325–332.
- Commission, E., 2023. Proposal for a directive of the European parliament and of the council on Soil monitoring and Resilience (Soil monitoring Law). European Commission, Brussels.

- Friedman, J., 2001. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* 29 (5), 1189–1232.
- Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*. Oxford University Press, New York.
- Gozukara, G., Hartemink, A.E., Huang, J., Demattè, J.A.M., 2025. Prediction accuracy of pXRF, MIR, and Vis-NIR spectra for soil properties—A review. *Soil Sci. Soc. Am. J.* 89 (2).
- Gozukara, G., Zhang, Y., Hartemink, A.E., 2021. Using vis-NIR and pXRF data to distinguish soil parent materials – an example using 136 pedons from Wisconsin. *Geoderma*, USA, p. 396.
- Han, J., Kamber, M., Pei, J., 2012. Cluster analysis: Basic concepts and methods. In: M. Kaufmann (Ed.), *Data mining*, pp. 443–495.
- Hartigan, J.A., Wong, M.A., 1979. A K-Means Clustering Algorithm. *J. Roy. Stat. Soc.: Ser. C (Appl. Stat.)* 28 (1), 100–108.
- Hashemi, M., Karimi, H.A., 2018. *Weighted Machine Learning. Statistics, Optimization & Information Computing* 6 (4).
- Helfenstein, A., Baumann, P., Viscarra Rossel, R., Gubler, A., Oechslin, S., Six, J., 2021. Quantifying soil carbon in temperate peatlands using a mid-IR soil spectral library. *Soil* 7 (1), 193–215.
- Hengl, T., Parente, L., Wheeler, L., 2022. *Spatial sampling and resampling for Machine Learning* (v0.1).
- Hermansen, C., Knadel, M., Moldrup, P., Greve, M.H., Gislum, R., de Jonge, L.W., 2016. Visible-near-infrared spectroscopy can predict the clay/organic carbon and mineral fines/organic carbon ratios. *Soil Sci. Soc. Am. J.* 80 (6), 1486–1495.
- Hermansen, C., Knadel, M., Moldrup, P., Greve, M.H., Karup, D., de Jonge, L.W., 2017. Complete soil texture is accurately predicted by visible near-infrared spectroscopy. *Soil Sci. Soc. Am. J.* 81 (4), 758–769.
- Houmark-Nielsen, M., 2011. Pleistocene Glaciations in Denmark: A Closer Look at Chronology, Ice Dynamics and Landforms. In: J. Ehlers, P. Gibbard, P. Hughes (Eds.), *Quaternary Glaciations - Extent and Chronology - A Closer Look. Developments in Quaternary Sciences*, pp. 47–58.
- Hu, Y., Cross, A., Shen, Z., Bouma, J., Viscarra Rossel, R.A., 2026. On soil health and the pivotal role of sensing. *Soil* 12 (1), 227–252.
- Isaaks, E.H., Srivastava, R.M., 1989. *An Introduction to Applied Geostatistics*. Oxford University Press, New York, NY, USA.
- Just, A.C., Arfer, K.B., Rush, J., Dorman, M., Shtein, A., Lyapustin, A., Kloog, I., 2020. Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (PM_{2.5}) using satellite data over large regions. *Atmos Environ* (1994) 239.
- Jakobsen, P.R., Tougaard, L., Anthonen, K.L., 2015. Danmarks Digitale Jordartskort, 1: 25,000. Version 4.0..
- Karger, D.N., Conrad, O., Böhrner, J., Kewhl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. *Scientific Data* 4 (1), 1–20.
- Knadel, M., Masís-Meléndez, F., de Jonge, L.W., Moldrup, P., Arthur, E., Greve, M.H., 2016. Assessing Soil Water Repellency of a Sandy Field with Visible near Infrared Spectroscopy. *J. Near Infrared Spectrosc.* 24 (3), 215–224.
- Knadel, M., Viscarra Rossel, R.A., Deng, F., Thomsen, A., Greve, M.H., 2013. Visible-Near Infrared Spectra as a Proxy for Topsoil Texture and Glacial Boundaries. *Soil Sci. Soc. Am. J.* 77 (2), 568–579.
- Kuhn, M., 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* 28 (5), 1–26.
- Kumar, A., Moharana, P.C., Jena, R.K., Malyan, S.K., Sharma, G.K., Fagodiya, R.K., Shabnam, A.A., Jigyasa, D.K., Kumari, K.M.V., Doss, S.G., 2023. Digital mapping of soil organic carbon using machine learning algorithms in the upper brahmaputra valley of Northeastern India. *Land* 12 (10).
- Lechner, A.M., Stein, A., Jones, S.D., Ferwerda, J.G., 2009. Remote sensing of small and linear features: quantifying the effects of patch size and length, grid position and detectability on land cover mapping. *Remote Sens. Environ.* 113 (10), 2194–2204.
- Levin, G., 2025. Basemap 05. Documentation of the data and method for elaboration of a land use and land cover map for Denmark, Aarhus University, DCE – Danish Centre for Environment and Energy.
- Liu, W., Baret, F., Xingfa, G., Qingxi, T., Lanfen, Z., Bing, Z., 2002. Relating soil surface moisture to reflectance. *Remote Sens. Environ.* 81, 238–246.
- Madsen, H.B., Nørr, A.H., Aa Holst, K., 1992. *Atlas of Denmark. The Danish Soil Classification, 3. The Royal Danish Geographical Society*.
- Malone, B.P., McBratney, A.B., Minasny, B., 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma* 160 (3–4), 614–626.
- Mantripragada, K., Dao, P., He, Y., Qureshi, F., 2022. The Effects of Spectral Dimensionality Reduction on Hyperspectral Pixel Classification: A Case Study. *arXiv e-prints arXiv:2104.00788v2*.
- Moura-Bueno, J.M., Dalmolin, R.S.D., Horst-Heinen, T.Z., Grunwald, S., ten Caten, A., 2021. Environmental covariates improve the spectral predictions of organic carbon in subtropical soils in southern Brazil. *Geoderma* 393.
- Mouselimis, L., 2024. ClusterR: Gaussian Mixture Models, K-Means, Mini-Batch-Kmeans, K-Medoids and Affinity Propagation Clustering. *R Package Version* 1 (3), 3.
- Muhammad, K., Glass, H.J., 2023. A fuzzy logic based explicit declustering technique. *Heliyon* 9 (6), e16817.
- Møller, A.B., Nyborg, L., Grogan, K., Druce, D., Filil, S., Greve, M.B., Gutierrez, J.S., Styczen, M., Greve, M., H., Knudsen, L., Beucher, A.M., 2025. High-resolution 3D soil texture mapping in Denmark using satellite time series and bare soil composites. *Advances in Agronomy*.
- Norouzi, S., de Jonge, L.W., Marzieh Zaresourmanabad, Knadel, M., Danielsen, A.C., Mikstas, D., Robinson D., Gomes, L., Weber, P., Møldrup, P., Greve, M., Gutierrez-Diaz, J., 2026. *Soil Spectral Compression: A Comparative Analysis of Classical and Deep Learning Methods. SSRN*.
- Peng, Y., Roell, Y.E., Bjørn Møller, A., Adhikari, K., Beucher, A., Greve, M.B., Greve, M. H., 2020. Identifying and mapping terrons in Denmark. *Geoderma* 363, 114174.
- Peng, Y., Roell, Y.E., Odgers, N.P., Møller, A.B., Beucher, A., Greve, M.B., Greve, M.H., 2021. Mapping and describing natural terroir units in Denmark. *Geoderma* 394.
- Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., Greve, M.H., 2015. Modeling Soil Organic Carbon at Regional Scale by Combining Multi-Spectral Images with Laboratory Spectra. *PLoS One* 10 (11), e0142295.
- Rahbar, F., Shahbazi, F., Rezaei, H., Biswas, A., 2024. Multi-property digital soil mapping at 30-m spatial resolution down to 1 m using extreme gradient boosting tree model and environmental covariates. *Remote Sensing Applications: Society and Environment* 33.
- Ramirez-Lopez, L., Wadoux, A.M.J.C., Franceschini, M.H.D., Terra, F.S., Marques, K.P.P., Sayão, V.M., Demattè, J.A.M., 2019. Robust soil mapping at the farm scale with vis-NIR spectroscopy. *Eur. J. Soil Sci.* 70 (2), 378–393.
- Roell, Y.E., Peng, Y., Beucher, A., Greve, M.B., Greve, M.H., 2020. Development of hierarchical terroir workflow based on gridded data – a case study in Denmark. *Comput. Geosci.* 138.
- Román Dobaro, M., Padarian Campusano, J., McBratney, A.B., Malone, B., Minasny, B., 2023. Genosol and phenosol mapping in continental Australia is essential for soil security. *Soil Secur.* 13.
- Silva, S., Teixeira, B., Braga, M., Pereira, H., Lopes, G., Santos, G., Guimaraes, L., Resende, M., Mancini, M., Curi, N., Borguete, R., Cardelli, V., Cocco, S., Corti, G., Chakraborty, S., Li, B., Weindorf, D., 2021. pXRF in tropical soils: Methodology, applications, achievements and challenges. *Advances in Agronomy* 1–62.
- Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. *Visible and Near Infrared Spectroscopy in Soil Science*, 107, first ed. Elsevier Inc.
- Styc, Q., Pachon, J., Ng, W., Padarian, J., McBratney, A., 2025. Creating soil districts for Australia based on pedogenon mapping. *Geoderma* 454.
- Sun, Z., Liu, F., Wang, D., Wu, H., Zhang, G., 2023. Improving 3D digital soil mapping based on spatialized lab soil spectral information. *Remote Sens. (Basel)* 15 (21).
- Tziachris, P., Nikou, M., Aschonitis, V., Kallioras, A., Sachsamanoglou, K., Fidelibus, M. D., Tziritis, E., 2023. Spatial or random cross-validation? The effect of resampling methods in predicting groundwater salinity with machine learning in Mediterranean Region. *Water* 15 (12).
- Viscarra Rossel, R.A., Chappell, A., McKenzie, N., 2010. Mapping the information content of Australian visible-near infrared soil spectra. In: Gilkes, R.J., Prakhongkep, N. (Eds.), *19th World Congress of Soil Science. Soil solutions for a changing world*, Brisbane, pp. 91–94.
- Viscarra Rossel, R.A., Chen, C., 2011. Digitally mapping the information content of visible-near infrared spectra of surficial Australian soils. *Remote Sens. Environ.* 115 (6), 1443–1455.
- Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. *Soil Res.* 53 (8).
- Viscarra Rossel, R.A., McGlynn, R.N., McBratney, A.B., 2006a. Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. *Geoderma* 137 (1–2), 70–82.
- Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006b. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* 131 (1–2), 59–75.
- Wadoux, A.M.J.C., Heuvelink, G.B.M., de Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecol. Model.* 457.
- Wang, P.R., 2013. Referenceværdier: Døgn-, måneds- og årsværdier for regioner og hele landet 2001 - 2010, Danmark for temperatur, relativ luftfugtighed, vindhastighed, globalstråling og nedbør, Danmarks Meteorologiske Institut.
- Wangeci, A., Adén, D., Nikolajsen, T., Greve, M.H., Knadel, M., 2024. Comparing laser-induced breakdown spectroscopy and visible near-infrared spectroscopy for predicting soil properties: a pan-European study. *Geoderma* 444.
- Weber, P.L., Hermansen, C., Nørgaard, T., Pesch, C., Moldrup, P., Greve, M.H., Arthur, E., de Jonge, L., 2022. Evaluating the particle densities of subarctic soils using pedotransfer functions and vis-NIR spectroscopy. *Soil Sci. Soc. Am. J.* 86 (4), 964–978.
- Yang, M., Chen, S., Guo, X., Shi, Z., Zhao, X., 2023a. Exploring the potential of vis-NIR Spectroscopy as a Covariate in Soil Organic Matter Mapping. *Remote Sens. (Basel)* 15 (6).
- Yang, N., Wang, Y., Liu, B., Zhang, J., Hua, J., Liu, D., Bhople, P., Zhang, Y., Zhang, H., Zhang, C., Ruan, H., Wang, W., 2023b. Exploration of soil microbial diversity and community structure along mid-subtropical elevation gradients in Southeast China. *Forests* 14 (4).
- Yoo, M., Koo, H., 2024. Exploring the impact of spatial autocorrelation on optimistic bias in cross-validation and assessing the effectiveness of spatial cross-validation. *Cartogr. Geogr. Inf. Sci.* 52 (5), 596–609.
- Zhang, J., Schmidt, M.G., Heung, B., Bulmer, C.E., Knudby, A., 2022. Using an ensemble learning approach in digital soil mapping of soil pH for the Thompson-Okanagan region of British Columbia. *Can. J. Soil Sci.* 102 (3), 579–596.
- Zhu, A.X., Ma, T., Zhao, F.-H., Yang, X., Xia, Y., 2025. Uncertainty quantification for digital soil mapping. An overview. *Pedosphere*.