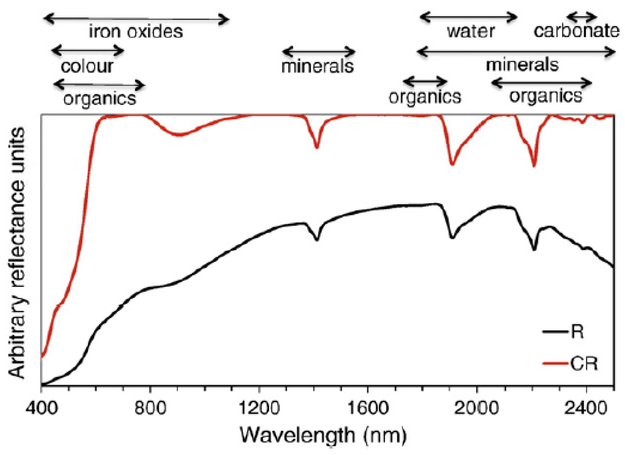
Dimensionality Reduction of Soil Vis–NIR Spectra: Implications for Soil Health Assessment and Mapping

Sarem Norouzi, Lis Wollesen de Jonge, Per Moldrup, Mogens Humlekrog Greve, Sebastian Gutierrez

AARHUS UNIVERSITY

Introduction: Measured soil spectra in the vis–NIR range contain information on soil physical, chemical, and mineralogical properties, offering a rapid and cost-effective approach for soil health assessment. However, raw spectra are high-dimensional and often unsuitable for direct modeling. We compared four dimensionality reduction methods—PCA, kernel PCA, autoencoders, and convolutional autoencoders—using reconstruction error as a direct measure of information loss. The analysis used 7,009 Danish soil spectra representing diverse land uses and soil types. Reconstruction error decreased with increasing numbers of latent variables, but nonlinear methods consistently outperformed PCA across all dimensionalities. At low dimensionality, nonlinear methods reduced reconstruction error by approximately 45–55% compared with PCA. We further mapped the learned latent variables across Denmark using spatial predictors, showing that they captured complex soil–landscape relationships linked to mineral and organic soil components.

Background and objective

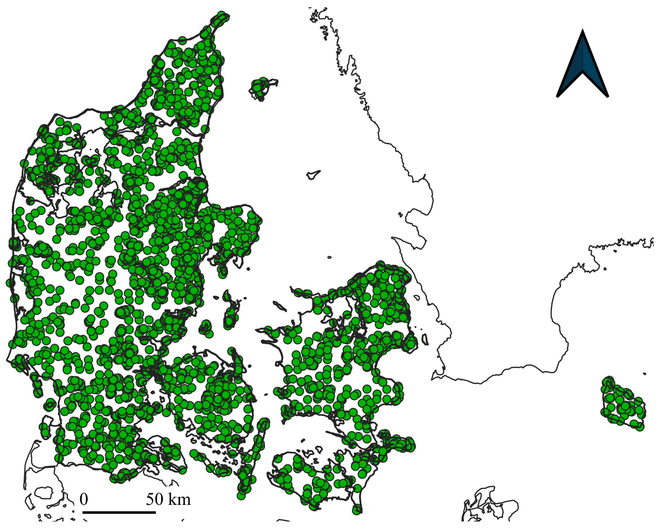


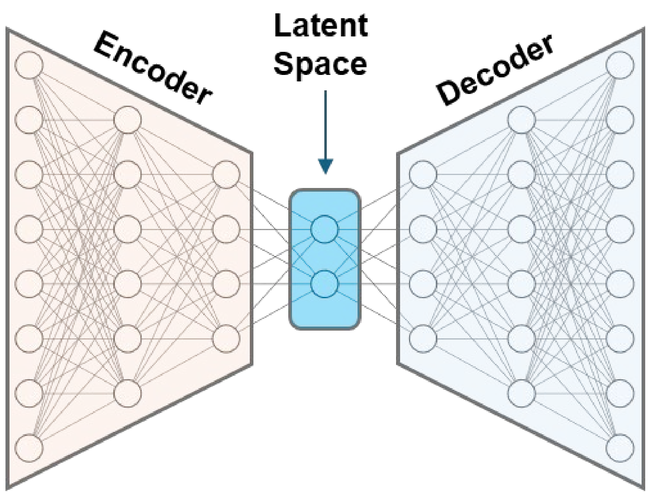
Key vis–NIR absorption features in soil spectra (Rossel & Chen (2011)

Objective: To compare linear and nonlinear dimensionality reduction methods for compressing soil vis–NIR spectra, evaluate how well they preserve spectral information, and investigate the spatial representation of the latent space.

Dataset and methodology

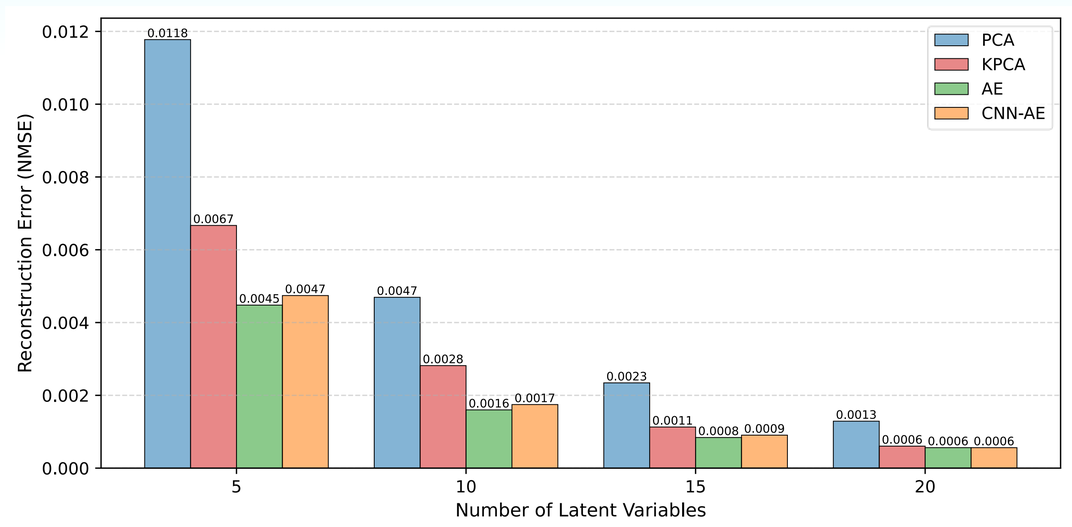
- Points with compressed vis-NIR measurements (n = 7,009)

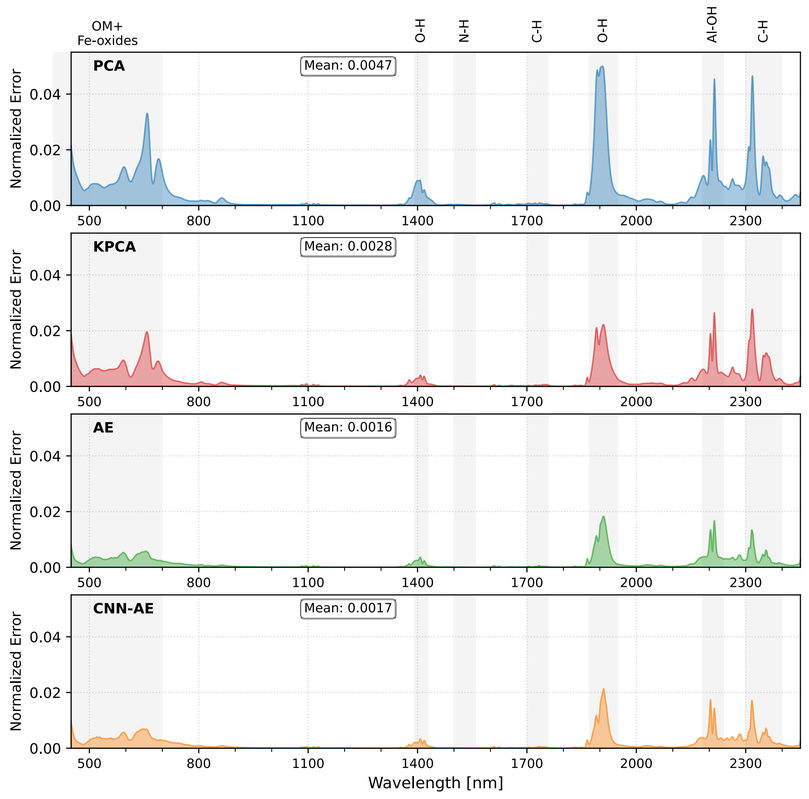




Schematic representation of an autoencoder architecture.

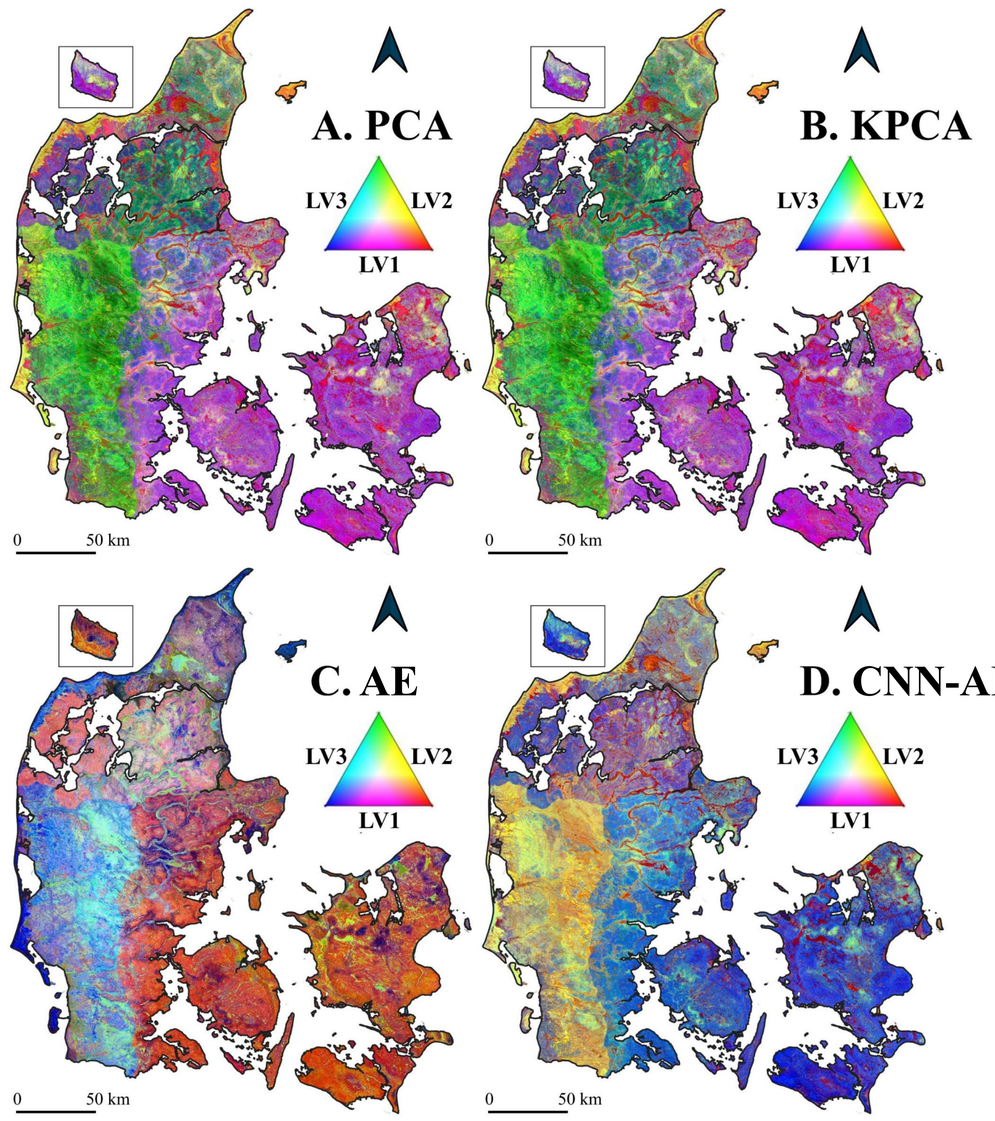
Findings





- AE and CNN-AE reduced reconstruction error by up to ~60% at low dimensionality and ~54% with 20 latent variables compared with PCA.
- Reconstruction error was concentrated in key spectral regions, especially around 500–800, 1400, 1900, and 2200–2300 nm, where major soil absorption features occur.

Each dimensionality reduction method captured soil variation across Denmark differently, reflecting distinct soil–landscape relationships.



Conclusion:

- Nonlinear methods preserved soil vis–NIR information better than PCA, showing that explained variance alone is insufficient for selecting latent variables..
- Latent maps provide valuable input for modeling soil health indicators. For example, they could be used to define Soil Monitoring Units for soil health assessment.