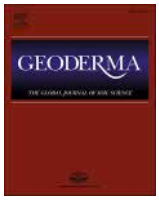
Geoderma 450 (2024) 117056

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma



Self-supervised learning of Vision Transformers for digital soil mapping using visual data

Paul Tresson [a,b,c,d,*,1], Maxime Dumont [d,e,f], Marc Jaeger [b,c], Frédéric Borne [b,c], Stéphane Boivin [d], Loïc Marie-Louise [d], Jérémie François [d], Hassan Boukcim [d], Hervé Goëau [b,c]

[a] AMAP, Université de Montpellier, IRD, CIRAD, CNRS, INRAE, Montpellier, France
[b] CIRAD, UMR AMAP, Montpellier, France
[c] AMAP, CIRAD, CNRS, INRAE, IRD, Université de Montpellier, Montpellier, France
[d] Valorhiz, Montpellier, France
[e] Univ. Montpellier, ITAP, Montpellier, France
[f] UMR ITAP, Institut Agro, INRAE, Montpellier, France

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Self-supervised learning
Vision transformers
Digital soil mapping
Arid lands

ABSTRACT

In arid environments, prospecting cultivable land is challenging due to harsh climatic conditions and vast, hard-to-access areas. However, the soil is often bare, with little vegetation cover, making it easy to observe from above. Hence, remote sensing can drastically reduce costs to explore these areas. For the past few years, deep learning has extended remote sensing analysis, first with Convolutional Neural Networks (CNNs), then with Vision Transformers (ViTs). The main drawback of deep learning methods is their reliance on large calibration datasets, as data collection is a cumbersome and costly task, particularly in drylands. However, recent studies demonstrate that ViTs can be trained in a self-supervised manner to take advantage of large amounts of unlabelled data to pre-train models. These backbone models can then be finetuned to learn a supervised regression model with few labelled data.

In our study, we trained ViTs in a self-supervised way with a 9500 km^2 satellite image of dry-lands in Saudi Arabia with a spatial resolution of 1.5 m per pixel. The resulting models were used to extract features describing the bare soil and predict soil attributes (pH H_2O, pH KCl, Si composition). Using only RGB data, we can accurately predict these soil properties and achieve, for instance, an RMSE of 0.40 $\pm$ 0.03 when predicting alkaline soil pH. We also assess the effectiveness of adding additional covariates, such as elevation. The pretrained models can as well be used as visual features extractors. These features can be used to automatically generate a clustered map of an area or as input of random forests models, providing a versatile way to generate maps with limited labelled data and input variables.

1. Introduction

Digital soil mapping is a crucial task in environmental sciences, since it serves as a basis for the understanding of soil properties and their spatial distribution. However, gathering data involves extensive field-work, and it can be challenging to collect samples across large and remote areas, such as deserts and arid lands.

Addressing this challenge, remote sensing data has been proved over the years to be a reliable way to monitor vast geographical areas. Furthermore, Multi spectral Optical Imagery serves as a reliable, cost effective way to map land cover due to the high spatial and temporal resolution of recent sensors such as Sentinel or Pleiades. Satellite data have therefore been used for digital soil mapping for a variety of tasks such as soil organic carbon prediction, soil water content measurement or soil/vegetation relationship analysis (*e.g.* Boettinger et al., 2008; Huisman et al., 2003; Peng et al., 2015; Maynard and Levi, 2017).

However, the analysis of such data can be complex and resource-intensive task, which has led to the use of machine learning for large-scale applications. Methods such as random forests, although yield the best results on tabular data, such as surveys or measurements tables (Grinsztajn et al., 2022), lack the ability to extract spatial patterns, unlike more recent models such as Convolutional Neural Networks

* Corresponding author.
E-mail address: paul.tresson@ird.fr (P. Tresson).
[1] Work done while at CIRAD and Valorhiz. Current affiliation is IRD.

https://doi.org/10.1016/j.geoderma.2024.117056
Received 1 February 2024; Received in revised form 4 October 2024; Accepted 4 October 2024
Available online 10 October 2024

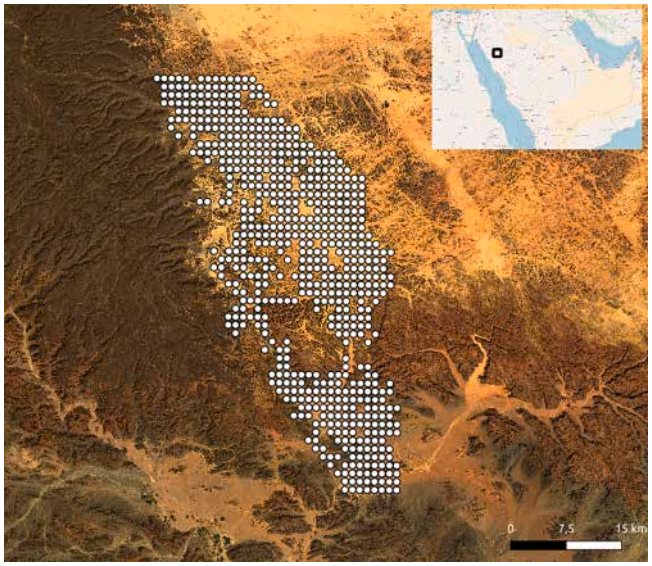


Fig. 1. Sampling sites of the kilometre grid with Spot 6 satellite image as background.

(CNN) and more recently, Vision Transformers (ViTs).

Consequently, using automated image analysis algorithms allows researchers to process larger datasets and study broader areas. Deep learning models have demonstrated remarkable capabilities in extracting meaningful information from images on almost all state of the art datasets for image analysis for nearly a decade (LeCun et al., 2015; Krizhevsky et al., 2017). This has been translated as well for remote sensing imagery for a variety of input data and tasks (Yuan et al., 2020). Nevertheless, deep learning methods typically demand large amounts of high quality labelled data to train effectively, which is prohibited in cases where ground truth data is difficult to obtain, such as soil mapping of arid areas.

Indeed, deep learning has been used for digital soil mapping, for instance using CNN and LSTM models (Wadoux, 2019; Padarian et al., 2019; Li et al., 2022; Zhang et al., 2022). However, these studies employed supervised learning methods, which depend on relatively large datasets or extensive covariate measurements. Unfortunately, large soil datasets and comprehensive covariate measure ments are often unavailable in many real-world applications.

Nonetheless, the state of the art in image analysis has recently shifted to Self-Supervised Learning methods (SSL), a method that offers a promising solution to this labelling conundrum. In self-supervised learning, the model starts by learning properties of the dataset in an unsupervised way via pretext tasks, such as the reconstruction of masked data (Bao et al., 2021; He et al., 2022) or self-distillation (Caron et al., 2021). By solving these pretext tasks, the model learns to capture essential features and patterns in the data. When a SSL pre-trained model is fine-tuned for a target task, it is already able to produce rich descriptors of the dataset's features. SSL demands for more unlabelled data but on the other hand, finetuning is less prone to overfitting and less data-hungry (Ericsson et al., 2021).

In this study, we explore the synergy between remote sensing optical data and SSL techniques for digital soil mapping. Our case study is located in the region of AlUla in the Medina Province of Saudi Arabia. This region is characterised by desert and arid landscapes, making the ground visible from above. Our underlying hypotheses are that an expert knowing these landscapes could infer soil properties from high resolution satellite data, and that a well trained deep learning model could replicate this knowledge. We train a Vision Transformer model (Dosovitskiy et al., 2020) in a SSL manner on satellite images and measure its ability to predict soil properties measured on the ground. To quantify the usefulness of the proposed method, we compare our results to more conventional machine learning, namely random forests.

Table 1
Layer details of the Vision Transformer used. Aside from the regression head, the architecture follows Caron et al. (2021). The projection and transformer blocks are commonly referred to as"backbone".

Module	Layer (type)	Output Dimension	Param #
Input		[224,224,3]	
Projection	Conv2d	[768, 14, 14]	590,592
and embedding	PatchEmbed	[196, 768]	0
Transformer	Linear	[197, 768]	590,592
Block	Dropout	[197, 768]	0
(×12)	Attention	[197, 768]	0
	LayerNorm	[197, 768]	1,536
	Linear	[197, 3072]	2,362,368
	GELU	[197, 3072]	0
	Dropout	[197, 3072]	0
	Linear	[197, 768]	2,360,064
	Dropout	[197, 768]	0
	LayerNorm	[197, 768]	1,536
	Linear	[197, 2304]	1,771,776
Layer Normalisation	LayerNorm	[197, 768]	1,536
and	Linear	[512]	393,728
Regression Head	ReLU	[512]	0
	Linear	[512]	262,656
	ReLU	[512]	0
	Linear	[1]	513

2. Material and methods

2.1. Ground truth soil data

The foundation of our study is the ground truth soil data obtained from the SoFunLand (SFL) Project. This project maintains a comprehensive soil monitoring network based on a 1 km regular grid across the AlUla region, covering 1,069 km^2. The SFL network consists of 663 monitoring sites, each located at the center of a 1 x 1 km cell Fig. 1. These sites provide detailed soil profile, soil physico-chemical properties, as well as information on site environment, location, vegetation, and land management. The soil samples analysis follow the method used by Maurice et al. (2023). The samples in the present study were collected in 2019. They are composites of 5 samples collected from a depth of 30–40 cm across a 1 x 1 m plot. Our dataset includes pH measurements obtained using both H_2O and KCl, as well as X-ray fluorescence spectrometry of elemental concentration (Si percentage) to demonstrate the versatility of our method.

2.2. Remote sensing data

Our remote sensing data comprise RGB Spot 6 images. The original multispectral Spot images were pan-sharpened on 1 m resolution panchromatic images. The final dataset is then of very high spatial resolution (1.5 m) but containing only the RGB bands. The images were captured during the soil sampling campaign and cover the entire kilometre grid.

We also used a Digital Elevation Model (DEM) at 1 m resolution (map of the DEM available in Appendix C) to evaluate the impact of the addition of input variables. To be fed into the neural network, the DEM was resampled at 1.5 m resolution like the RGB data.

2.3. Self-supervised learning

We used a ViT base deep learning model (Dosovitskiy et al., 2020) (see Table 1). This model is pre-trained using the DINO method (Caron et al., 2021) (see Appendix A for an overview of the method). SSL pre-training was done starting from the model trained on the ImageNet1K (Russakovsky et al., 2015) dataset provided by Caron et al. (2021). This starting model hasbeen therefore trained on natural images present in

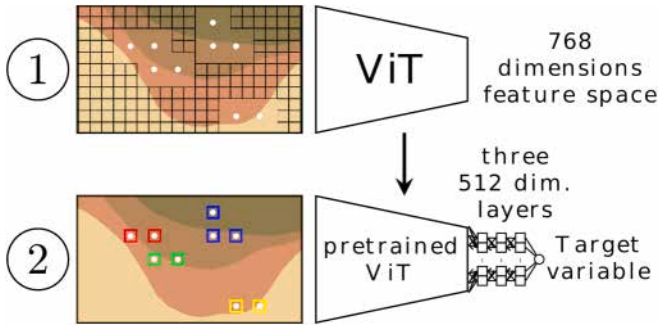


Fig. 2. Overview of the SSL training protocol. The backbone is first trained without label with the DINO method (1) before being trained in a supervised way to predict soil properties of sampling points with a geographically defined cross validation (2).

the ImageNet dataset, for instance pictures of cats, dogs, people or cars. To adapt the model to a new domain (satellite images over the desert), it is pre-trained on tiled satellite images, with each tile sized at 224 × 224 pixels (therefore corresponding to 336 × 336 m on the ground). This sampling size was chosen because it is the standard input size of a base ViT (Dosovitskiy et al., 2020). Then, input images are used at native resolution, without any resampling within a tile (aside from data augmentation, see Appendix B). This 336 x 336 m area seemed to be a suitable scale to study the soil properties of a given point. Indeed, this scale is able to convey local information but also features the surrounding of the sampling point (see Fig. 3). The entire image is sampled along a grid, except for zones around ground truth data with a buffer of 448 pixels around the ground truth point (*i.e.* 672 m). This sampling reduced the risk of bias with the model already learning features around ground truth sampling points. Then, the model was first fine-tuned without labels using the DINO method (see Fig. 2). Subsequently, we fine-tuned the model in a supervised way using tiles where ground truth data is available. As our purpose here is not to evaluate the accuracy of the map produced but rather the robustness and extrapolation capacity of the method, we chose to conduct the cross-validation using spatially defined folds, even if this can lead to over-pessimistic results (Wadoux et al., 2021). The supervised training is then performed with a 6-fold cross- validation strategy (see Fig. 4), with the folds being defined by applying a K-means to the geographical coordinates of the points.

As described in Table 1, our model is then constituted of two main parts: The "backbone" that was pre-trained and outputs descriptors of the tiles in a feature space, and a regression "head", finetuned on a downstream task, in our case the prediction of soil properties. Our regression head is composed of three fully connected layers with an intermediate dimension of 512 and ReLus that were added on top of the pre-trained backbone. In the case of the ViT base architecture, the feature space after the backbone is of 768 dimensions (see Fig. 2 and Table 1). During supervised training, backbone weights were frozen and only the regression head was trained. The loss criterion used was the Mean Square Error (MSE).

As described in Table 1, the backbone part of the model (projection and transformer blocks) constituted a total of 85,646,592 parameters. These parameters were only trained during the unsupervised phase of the training. During the last supervised phase, only the 656,897 parameters of the regression head were learned.

When training on RGB + DEM data, the weights corresponding to the DEM channel were initialized by cloning the weights corresponding to the R channel, thus making the projection head able to take (224,224,4) dimension input, with DEM as the fourth channel. Detailed training hyper-parameters can be found in the appendix (see Supplementary Materials B).

2.4. Comparison with other machine learning methods

To evaluate our method, we compare it to random forests trained using the same method as Dumont et al. (2024) but using spatial cross-validation rather than random folds and only RGB and DEM as covariates. Following this method, the random forest takes only the value of the pixel on the point as input. During the fit, we tested a range of hyper-parameters and selected the best models (see Appendix B).

We also compared with random forests fitted using the descriptors produced by the ViT pre-trained backbone in place of the deep learning regression head.

To asses the impact on SSL pre-training, we also trained the model in a simple transfert learning modality using the starting ImageNet weights.

2.5. Evaluation metrics

To quantify the accuracy of the predictions, we use the Root Mean Square Error (RMSE) and R^2 as the evaluation metrics, which can be used for our different target variables, pH H_2O, pH KCl and Si composition.

3. Results

Tables 2 and 3 summarize RMSE and R^2 obtained for our target variables with different methods.

3.1. Proposed method

Overall, SSL trained ViT on RGB data were able to predict both pH values with satisfying accuracy given the range of ground truth values ([7.32–10.40], [7.02–9.65], [41–99] for pH H_2O, pH KCl and Si respectively). As shown in Figs. 5 and 6, the models demonstrated the capacity to accurately map soil properties while adhering to overarching patterns, even in the presence of occasional outliers (most notably for Si). This handling of outliers was further discussed in section 4.4.

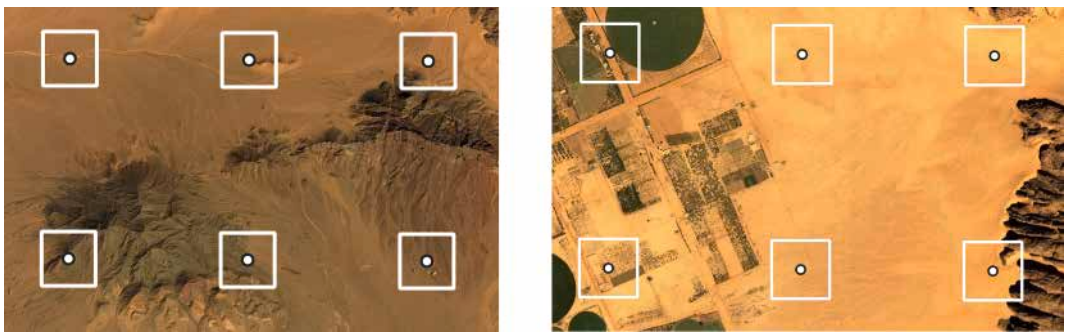


Fig. 3. Examples of 224 × 224 pixel tiles around ground sampling points.

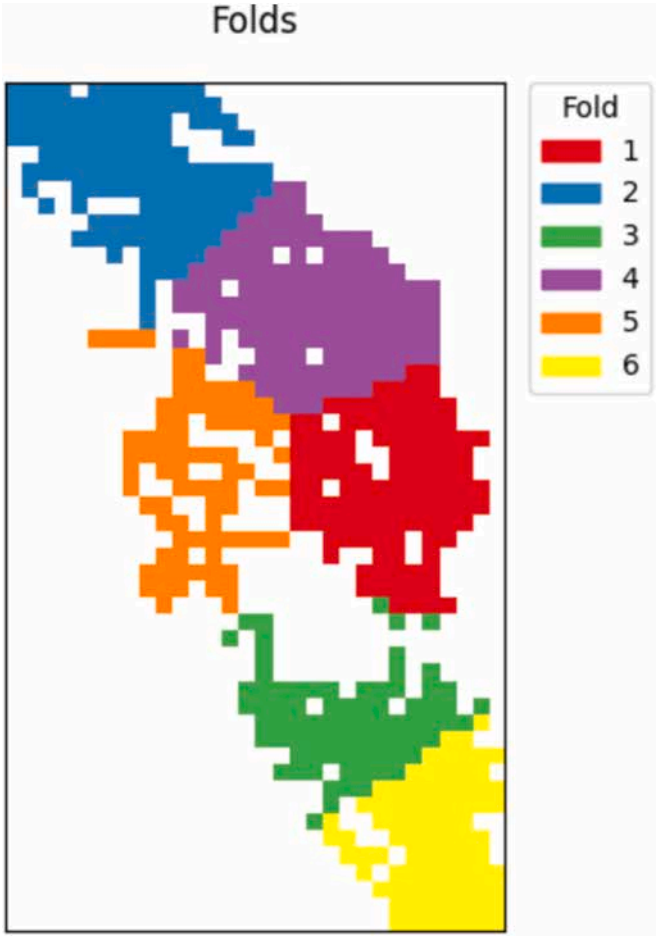


Fig. 4. Geographical separation of folds.

Table 2
Mean RMSE values (± std. deviation) by fold for the studied target variables for different input covariates and prediction methods.

Method	Covariates	pH H_2O	pH KCl	Si
RF	RGB	0.56 ± 0.03	0.51 ± 0.06	7.89 ± 1.66
	RGB + DEM	0.55 ± 0.07	0.47 ± 0.03	8.13 ± 1.95
ViT	RGB	0.91 ± 0.13	1.52 ± 0.41	46.83 ± 5.00
	RGB + DEM	1.05 ± 0.19	1.05 ± 0.19	27.17 ± 3.93
ViT (SSL)	RGB	**0.51 ± 0.03**	**0.40 ± 0.03**	6.93 ± 1.03
	RGB + DEM	0.59 ± 0.12	0.50 ± 0.10	22.11 ± 3.26
ViT (SSL) + RF	RGB	0.57 ± 0.06	0.43 ± 0.03	7.92 ± 1.78
	RGB + DEM	0.52 ± 0.02	0.59 ± 0.07	**6.73 ± 1.39**

Table 3
R^2 values for the studied target variables for different input covariates and prediction methods.

Method	Covariates	pH H_2O	pH KCl	Si
RF	RGB	0.01	−0.12	0.08
	RGB + DEM	0.03	0.03	0.03
ViT	RGB	−1.76	−10.43	−31.18
	RGB + DEM	−2.52	−2.91	−9.74
ViT (SSL)	RGB	**0.07**	**0.23**	−0.36
	RGB + DEM	−0.10	−0.10	−5.93
ViT (SSL) + RF	RGB	−0.05	0.18	0.05
	RGB + DEM	−0.17	−0.14	**0.33**

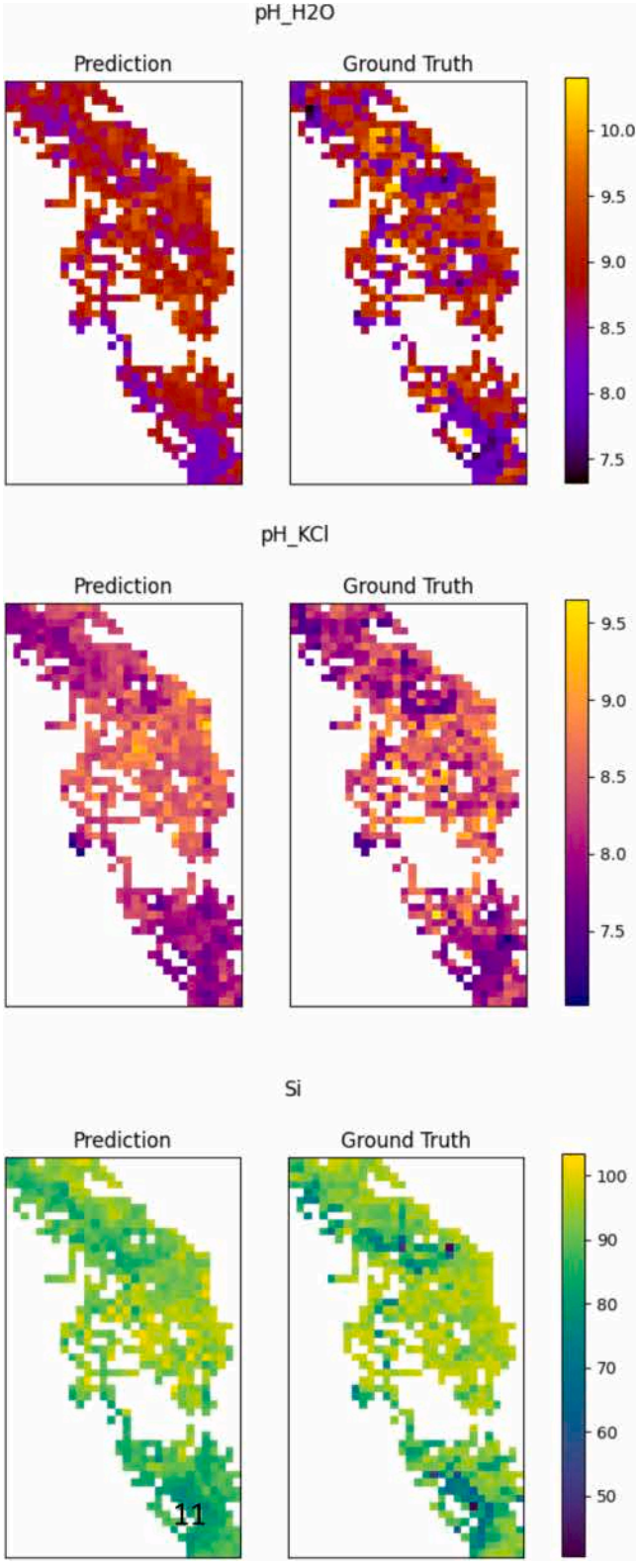


Fig. 5. Maps of model predictions for the three target variables.

3.2. Usage of deep learning over random forests

Deep learning with SSL on RGB data outperformed random forests fitted using RGB or RGB and DEM. However, using random forests after extracting deep learning features seem to be able to better fit the data in some cases than deep learning alone or random forests alone. However,

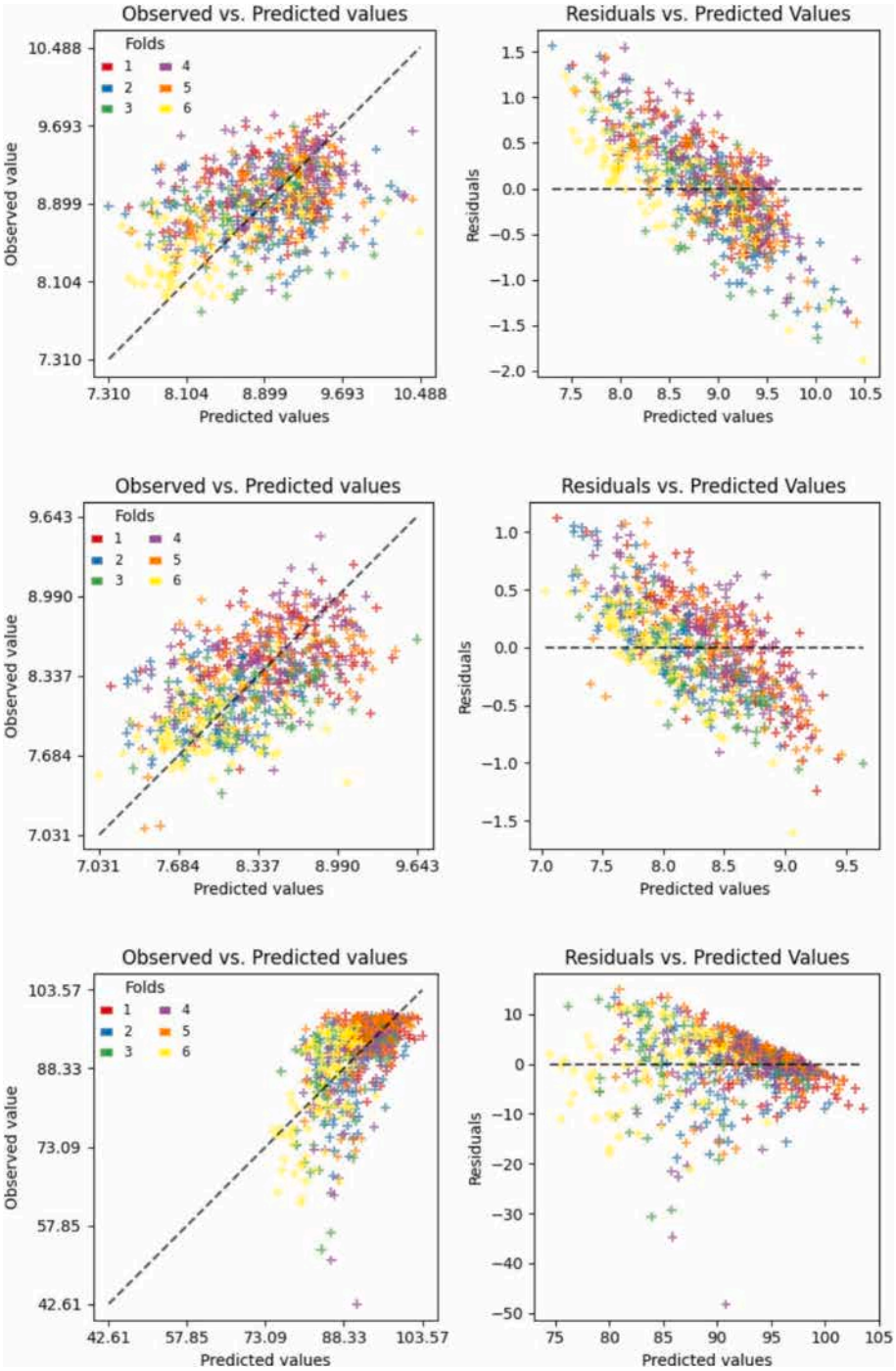


Fig. 6. Observed vs. Predicted and Residuals vs. Observed plots for the three target variables from top to bottom pH H_2O, pH KCl and Si.

random forests seem to be more able to fit more consistently than deep learning models, most notably for Si, our most unbalanced target variable (with a majority of points within the same values and only some outliers). These differences are further explored in section 4.2 in the discussion.

3.3. Impact of SSL

Without SSL pre-training, deep learning models were unable to fit correctly given the low number of samples. For instance, the R^2 scores achieved when fitting for Si clearly indicated overfitting of the model. Even using more balanced target variables such as pH H_2O or pH KCl, deep learning regression head were not able to fit and predict reliably. This kind of behaviour is to be expected when performing supervised learning only given our number of samples that is very low regarding deep learning standards (Safonova et al., 2023).

3.4. Addition of covariates

While the addition of covariates improved the performance of random forests, the performance of deep learning degraded by the addition of the DEM as a new channel in input. This can be explained by the large domain change required when fitting a new format of data (RGB + DEM input, *i.e.* four channels, rather than three with RGB) (see section 4.1 in discussion).

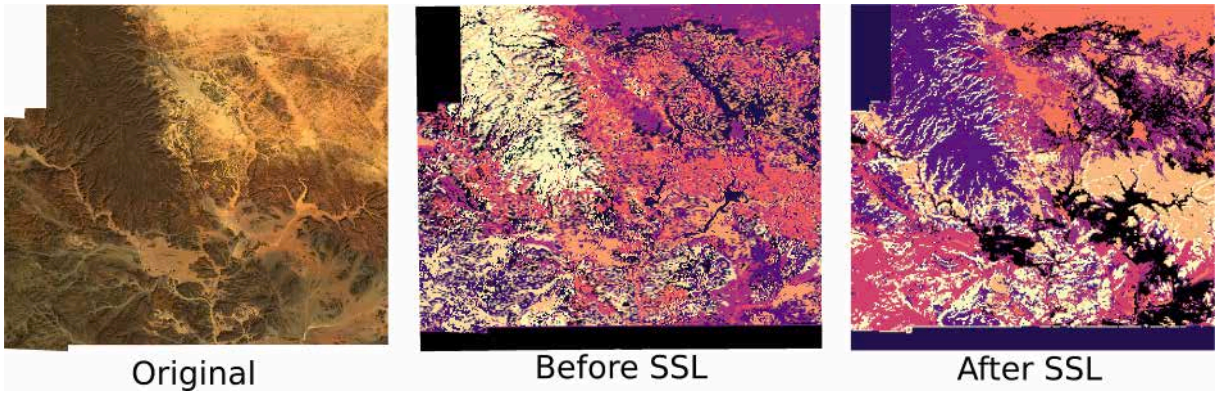


Fig. 7. Clustering of the model's feature space before and after SSL.

4. Discussion

4.1. Versatility of the method and advantages of self-supervised learning

Self-supervised pre-training has been shown to ease domain changes (Bucci et al., 2021). In our case, during SSL pre-training, the model feature space should change from one describing accurately natural images, such as the ones that can be found in the ImageNet dataset, to a feature space describing more adequately desert scenes seen from above. To check this hypothesis, we performed a simple K-means clustering of the feature space of the model when fed the entire satellite image (K = 8). A map of this clustering shows that the ViT backbone seems to better group and identify landscapes after SSL pre-training (see Fig. 7).

An other experiment was to reduce the dimension of the feature space of a model and plot the result to have an intuition of which points that were clustered or spread out. UMAP (Uniform Manifold Approximation and Projection) algorithm (McInnes et al., 2018) is often used because it is able to conserve the topology of the feature space. In Fig. 8, we reduced the 768 dimension feature space into 2 dimensions to be able to plot it, both for the backbone before and after SSL. SSL seems to spread out the points in the feature space, possibly making it easier for the model to discriminate and learn features during the supervised phase.

These experiments as well as our results show that the features produced by the SSL-pre-trained backbone are potentially more discriminant when describing the scenes we are focusing on and could potentially be relevant when interpreting model outputs (see Wadoux et al., 2020).

Furthermore, we achieved meaningful performances while avoiding overfitting despite using a small dataset compared to common deep learning practices (Safonova et al., 2023). Indeed, our dataset only has 662 points associated with ground truth data, which is very small compared to classicaldeep learning datasets (for comparison, the classic MNIST dataset contains 70 K images and ImageNet 14 M images Deng, 2012; Russakovsky et al., 2015). This can be achieved because the model has already learned relevant features during pretraining.

While in our case, the SSL backbone trained on RGB + DEM data was not able to overcome the backbone trained on RGB alone, recent studies in remote sensing show that SSL pre-trained backbone are a reliable way to produce quality features on a variety of domains other than RGB, such as multispectral or hyperspectral data (Jakubik et al., 2023; Cong et al., 2023; Braham et al., 2022)

Such explorations of the feature space of the SSL pre-trained backbone could be a way to easily cluster and map data with limited human input, thus helping expert planning.

4.2. Usage of random forests with deep learning features

In this study, we have fitted random forests using deep learning features produced by the SSL backbone. While providing less accurate results than fitting a deep learning head, the random forests were able to fit more consistently than deep learning heads. Indeed, while neural networks can be applied to a wide range of tasks and datasets, using them on constrained, unbalanced and small datasets increase the risk of overfitting. Our results then suggest that using a random forest as regression head rather than a deep learning head after the backbone could be a more reliable way to fit and predict data when faced with a limited number of samples.

During the last decade, deep learning and neural networks have mostly been treated as monolithic, with an input and an output, without considering the descriptors and features extracted by the model – except some notable exceptions like Mask RCNN (He et al., 2018) or U-Net (Ronneberger et al., 2015) that have been adapted with a variety of backbones. With SSL back bones now being able to provide quality and spatially aware descriptors (see for instance experiments conducted by Oquab et al. (2023)), the state of the art evolves again towards two step methods, separating feature extraction and target task. Indeed, it is common for SSL studies to assess the performances of their backbones by relying on simple classification methods such as linear or KNN classifers (see for instance Oquab et al. (2023) and Caron et al. (2021).

Then, in contexts with limited labelled data, a two step method com bining deep learning features and random forests (or other machine learning methods requiring high quality descriptors as input) appears to be viable.

This could then overcome the need for handcrafted descriptors of classical machine learning and the need for numerous laboratory measured data for supervised deep learning.

However, the features produced by a deep learning backbone can be noisy and this can for instance explain why we were not able to overcome deep learning heads when fitting random forests after a deep learning encoder. Simple manipulations of the features could be a way to overcome these difficulties,such as Principal Component Analysis (PCA) or UMAP to reduce the dimensionality of the features. Fig. 9 shows the result of a PCA on the features produced by an SSL pretrained ViT when fed RGB + DEM data. The model seems to be able to distinguish different zones, thus providing an unsupervised map of the area, integrating both RGB and DEM variables.

4.3. Use of remote sensing and visual data

An inherent limitation in employing remote sensing for digital soil mapping lies in its capacity to capture surface-level information exclusively (aside from radar data for instance (Huisman et al., 2003)). In the context of our present study, which focuses on arid landscapes characterised by predominantly exposed or sparsely covered soils, this limitation poses minimal concern. However, when considering soil mapping in regions with substantial vegetation cover or human-made structures, numerous factors can impact soil characteristics, many of which remain

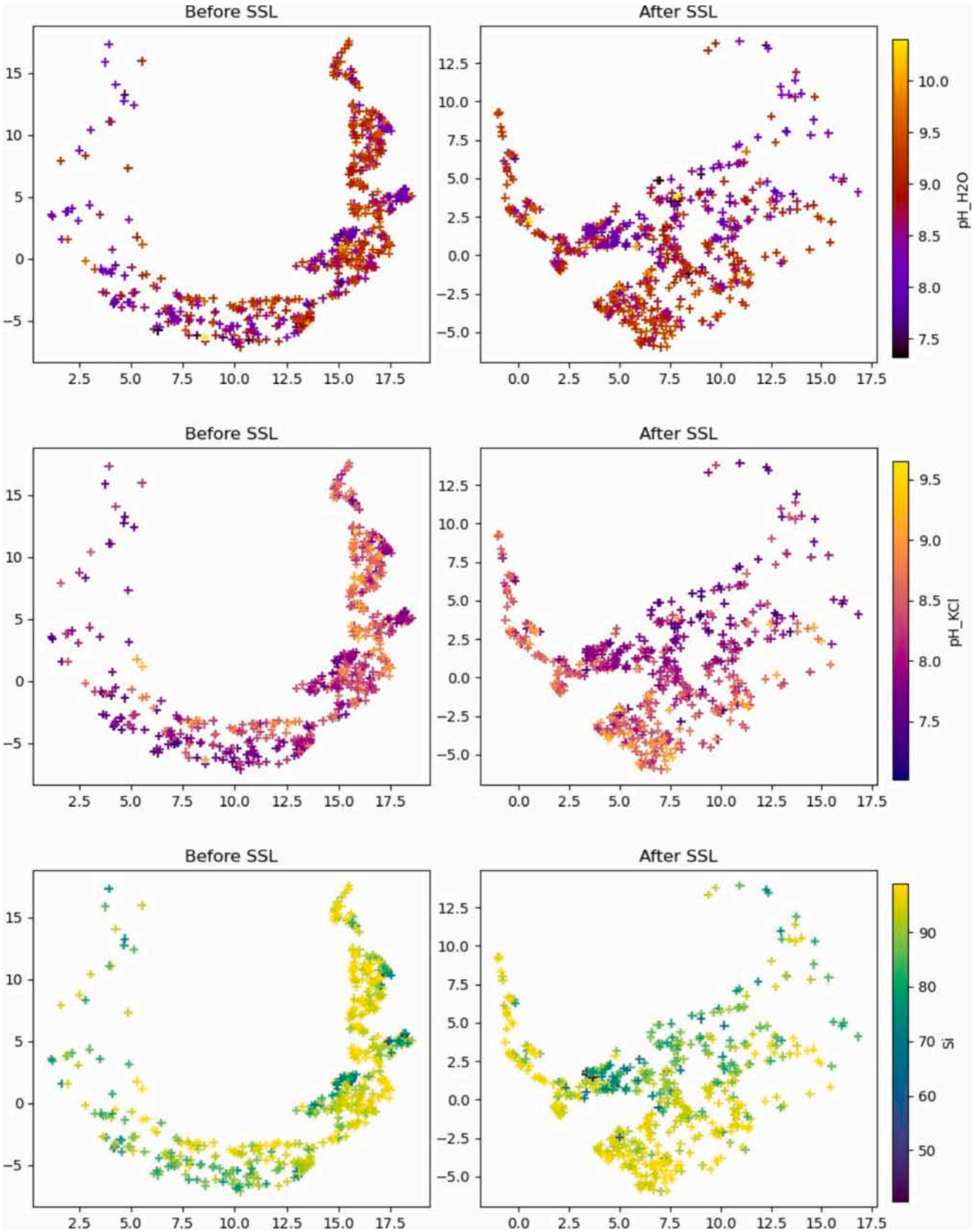


Fig. 8. UMAP projection of the ViT backbone feature space before and after SSL with color values corresponding to our three target variables.

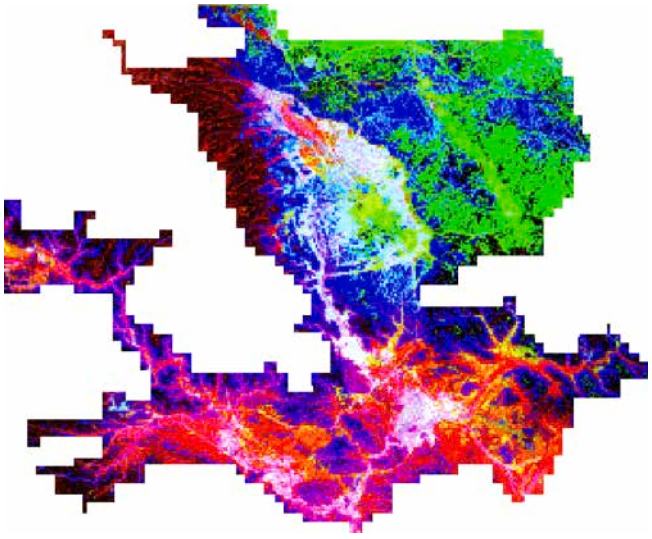


Fig. 9. PCA projection of the features produced by a SSL pretrained ViT model when feeded RGB + DEM data (first component is mapped on the R channel, second on B and third on G channel).

concealed from remote sensing observations (*e.g.* vegetation composition or inter-species interactions which are not directly discernible through aerial or satellite imagery).

4.4. Handling of outliers

For each target variable, our model seems to correctly identify the global patterns in soil characteristics (see Fig. 3), meaning that overall acid or basic regions will be correctly identified as such. However, points with local characteristics differing from the surrounding region are often predicted with a larger error. This is for instance visible for points in the north-eastern part of the grid that are covered with crops which influences pH and Si content at a local scale (see Figs. 3 and 3). In our case, sampling points around agricultural lands represent only 48 of the 662 points, which was probably too few for the model to accurately learn their properties. In the case where it is impossible to provide more points, a perspective would be to change the balance of the dataset and give more weights to points with crops during supervised training. While these techniques are common in deep learning, it is important to ponder the use of class weighting as not to alter the relevance and generalizability of the dataset for further usage (Johnson and Khoshgoftaar, 2019). An other perspective would be to change the MSE loss used here during training for a loss with a different handling of outlier, such as a L1 loss or Huber loss, for instance (Wang et al., 2020).

5. Conclusion

Our study shows that the use of SSL allows to leverage large amount of remote sensing data to predict soil features using few ground truth data compared to common deep learning practices. Then, with the use of SSL, the predictive abilities of the state of the art models such as ViTs and the availability of remote sensing data can be harnessed to map soil properties. Moreover, the study of the feature space of a SSL trained model opens new perspectives to efficiently create maps that summarize complex input remote sensing data. The automated spatial analysis can assist experts in designing optimal sampling strategies and streamlining fieldwork in challenging or large areas.

CRediT authorship contribution statement

Paul Tresson: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Maxime Dumont:** Writing – review & editing, Validation, Methodology, Data curation. **Marc Jaeger:** Writing – review & editing, Project administration, Funding acquisition. **Fŕederic Borne:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Śtephane Boivin:** Validation, Resources, Conceptualization. **Loïc Marie-Louise:** Writing – review & editing, Validation, Software. **Jeŕemie Fran¸cois:** Software, Project administration, Conceptualization. **Hassan Boukcim:** Project administration, Funding acquisition. **Herv́e Goëau:** Writing – review & editing, Supervision, Software, Project administration, Funding acquisition, 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.

Data availability

The data that has been used is confidential.

Acknowledgments

This work is supported by the French National Research Agency as part of ANR-21-PRRD-0034-01 support from Plan France Relance program, Valorhiz SA and CIRAD AMAP joint unit.

Use of Generative AI statement

Generative AI was used during the redaction of this article for syntax and language correctness purposes only.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2024.117056.

References

Bao, H., Dong, L., Piao, S., Wei, F., 2021. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254.

Boettinger, J., Ramsey, R., Bodily, J., Cole, N., Kienast-Brown, S., Nield, S., Saunders, A., Stum, A., 2008. Landsat spectral data for digital soil mapping, in: Digital soil mapping with limited data. Springer, pp. 193–202.

Braham, N.A.A., Mou, L., Chanussot, J., Mairal, J., Zhu, X.X., 2022. Self-supervised learning for few shot hyperspectral image classification, in:IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium. IEEE 267–270.

Bucci, S., D'Innocente, A., Liao, Y., Carlucci, F.M., Caputo, B., Tommasi, T., 2021. Self-supervised learning across domains. IEEE Trans. Pattern Anal. Mach. Intell. 44, 5516–5528.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 9650–9660.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D.B., Ermon, S., 2023. arXiv:2207.08051 Satmae: Pre-Training Transformers for Temporal and Multi-Spectral Satellite Imagery.

Deng, L., 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process Mag. 29, 141–142.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929.

Dumont, M., Brunel, G., Tresson, P., Nespoulous, J., Boukcim, H., Ducousso, M., Boivin, S., Taugourdeau, O., Tisseyre, B., 2024. Operational sampling designs for poorly accessible areas based on a multi-objective optimization method. Geoderma 445, 116888.

Ericsson, L., Gouk, H., Hospedales, T.M., 2021. How well do self-supervised models transfer?. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5414–5423.

Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on tabular data? arXiv:2207.08815.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009.
He, K., Gkioxari, G., Dollar, P., Girshick, R., 2018. Mask r-cnn. arXiv:1703.06870.
Huisman, J.A., Hubbard, S.S., Redman, J.D., Annan, A.P., 2003. Measuring soil water content with ground penetrating radar: a review. Vadose Zone J. 2, 476–491.
Jakubik, J., Chu, L., Fraccaro, P., Bangalore, R., Lambhate, D., Das, K., Oliveira Borges, D., Kimura, D., Simumba, N., Szwarcman, D., Muszynski, M., Weldemariam, K., Zadrozny, B., Ganti, R., Costa, C., Watson, C., Mukkavilli, K., Roy, S., Phillips, C., Ankur, K., Ramasubramanian, M., Gurung, I., Leong, W.J., Avery, R., Ramachandran, R., Maskey, M., Olofossen, P., Fancher, E., Lee, T., Murphy, K., Duffy, D., Little, M., Alemohammad, H., Cecil, M., Li, S., Khallaghi, S., Godwin, D., Ahmadi, M., Kordi, F., Saux, B., Pastick, N., Doucette, P., Fleckenstein, R., Luanga, D., Corvin, A., Granger, E., 2023. HLS Foundation. https://doi.org/10.57967/hf/0952.
Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6, 1–54.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. Commun. ACM 60, 84–90.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Li, Q., Zhu, Y., Shangguan, W., Wang, X., Li, L., Yu, F., 2022. An attention aware LSTM model for soil moisture and soil temperature prediction. Geoderma 409, 115651.
Maurice, K., Bourceret, A., Youssef, S., Boivin, S., Laurent-Webb, L., Damasio, C., Boukcim, H., Selosse, M.A., Ducousso, M., 2023. Anthropic disturbances impact the soil microbial network structure and stability to a greater extent than natural disturbances in an arid ecosystem. In: Science of the Total Environment, p. 167969.
Maynard, J.J., Levi, M.R., 2017. Hyper-temporal remote sensing for digital soil mapping: characterizing soil-vegetation response to climatic variability. Geoderma 285, 94–109.
McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.
Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning for digital soil mapping. Soil 5 (1), 79–89.
Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., Greve, M.H., 2015. Modeling soil organic carbon at regional scale by combining multispectral images with laboratory spectra. PLoS One 10, e0142295.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. arXiv:1505.04597.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.
Safonova, A., Ghazaryan, G., Stiller, S., Main-Knorn, M., Nendel, C., Ryo, M., 2023. Ten deep learning techniques to address small data problems with remote sensing. Int. J. Appl. Earth Obs. Geoinf. 125, 103569.
Wadoux, A.M.C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59–70.
Wadoux, A.M.C., Samuel-Rosa, A., Poggio, L., Mulder, V.L., 2020. A note on knowledge discovery and machine learning in digital soil mapping. Eur. J. Soil Sci. 71, 133–136.
Wadoux, A.M.C., Heuvelink, G.B., De Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. Ecol. Model. 457, 109692.
Wang, Q., Ma, Y., Zhao, K., Tian, Y., 2020. A comprehensive survey of loss functions in machine learning. Ann. Data Sci. 1–26.
Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., et al., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716.
Zhang, L., Cai, Y., Huang, H., Li, A., Yang, L., Zhou, C., 2022. A cnn-lstm model for soil organic carbon content prediction with long time series of modis-based phenological variables. Remote Sens. (Basel) 14, 4441.