GLOBAL SOIL SPECTRAL LIBRARIES & DEEP LEARNING-BASED MODELS

To improve remediation of radiocaesium contamination in agriculture

Franck Albinet
Independent Data Science & Machine Learning Consultant
Research contractor in the "Remediation of radioactive contaminated agricultural land"
Joint FAO/IAEA: CRPD15019 project

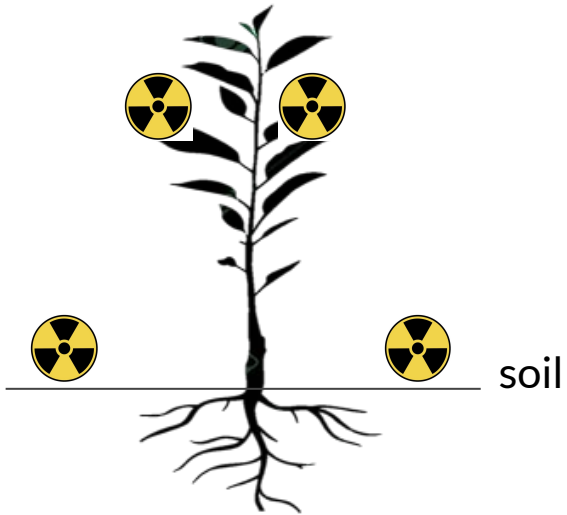
Outline

1. Our application domain & problem statement
2. Can soil spectroscopy help?
3. Model- vs. Data-centric approach
4. Our path ahead

1. Fate of radioisotopes in soil after nuclear accidents

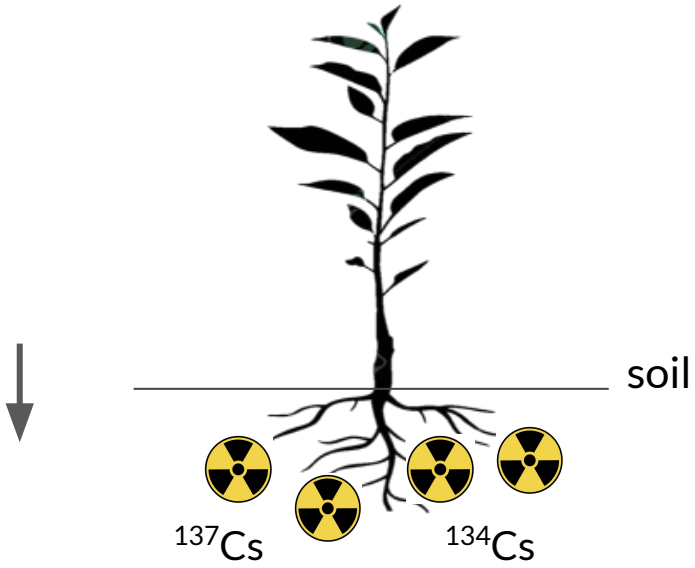
Fate of radioisotopes in soil after nuclear accidents

Radionuclides fallout



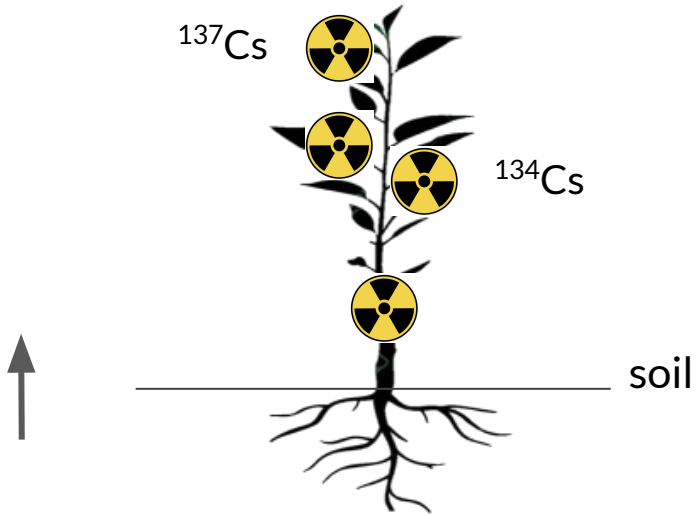
Fate of radioisotopes in soil after nuclear accidents

Migration to deeper soil layers



Fate of radioisotopes in soil after nuclear accidents

Root uptake



The role of K (Potassium)

K_{ex} competes with ^{137}Cs in root-uptake process

K fertilizer application to remediate

Where and how much to apply?

How much K_{ex}?

For a wide range of soils & ecosystems

Rapidly & at minimum cost

Up to landscape & regional scale

~~Wet chemistry~~

2. Infrared spectroscopy to the rescue?

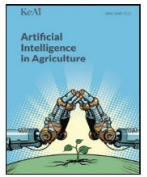
Infrared spectroscopy to the rescue?

Artificial Intelligence in Agriculture 6 (2022) 230–241

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Prediction of exchangeable potassium in soil through mid-infrared spectroscopy and deep learning: From prediction to explainability

Franck Albinet [a,b,*], Yi Peng [c], Tetsuya Eguchi [d,b], Erik Smolders [e], Gerd Dercon [b]

[a] Independent Researcher & Consultant, Guéthary, France
[b] Soil and Water Management & Crop Nutrition Laboratory, Joint FAO/IAEA Centre of Nuclear Techniques in Food and Agriculture, Seibersdorf, Austria
[c] Global Soil Partnership, Food and Agriculture Organization of the United Nations, Viale delle Terme di Caracalla, 00153 Rome, Italy
[d] Agricultural Radiation Research Center, Tohoku Agricultural Research Center, National Agriculture and Food Research Organization, Fukushima, Japan
[e] Soil and Water Management Unit, KU Leuven, Belgium

https://www.sciencedirect.com/science/article/pii/S2589721722000186

Predicting K_{ex} used to be challenging

No strong spectral features to support calibration model

Small & local SSLs impose low model capacity (overfitting)

“Blind” to complex spectral patterns

Our hypothesis, success conditioned:

Availability of large SSLs

Use of model class with highly scalable model capacity

Investigate Deep Learning (CNN) properties

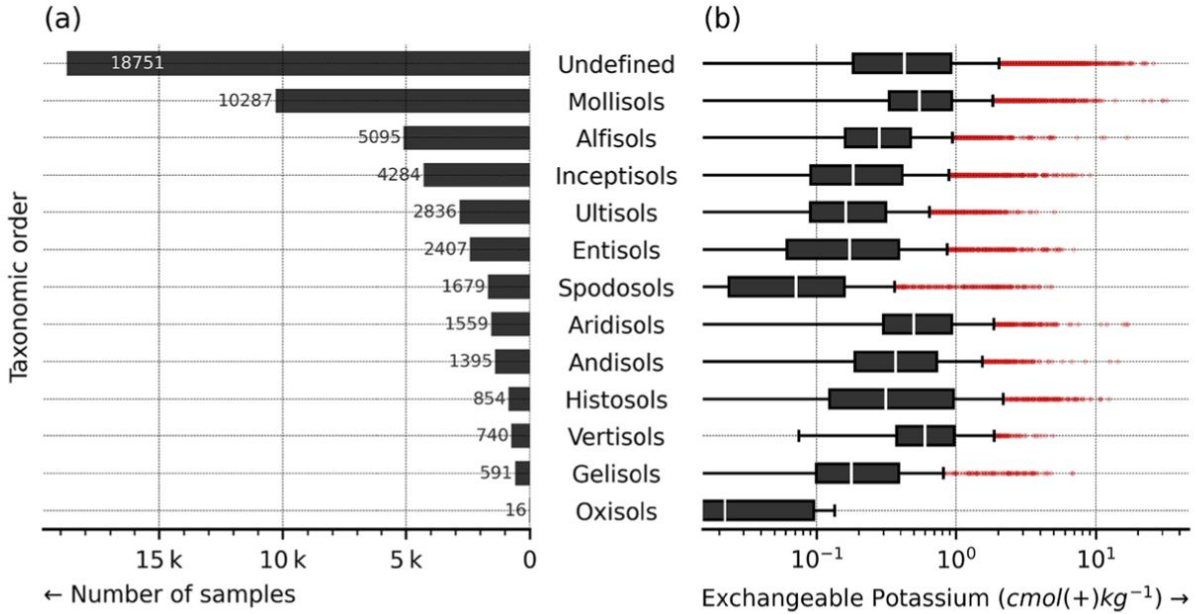
Does including all Soil Taxonomy orders (Global) work?

Interpretability

Prediction uncertainty

Viable candidate for global soil prediction service?

The USDA/KSSL Soil Spectral Library (~45K)



A “simple” (though deep) CNN architecture

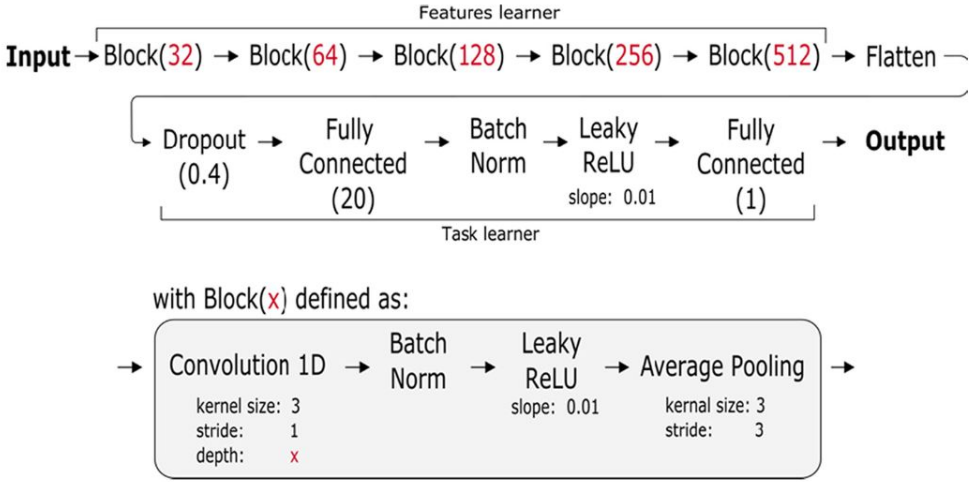


Fig. 2. Convolutional Neural Network architecture used in this study composed of (i) 5 stacked Convolutional-Pooling blocks with constant and small kernel size and increasing depth (feature learner) and (ii) a task learner.

Capacity to leverage growing data regime

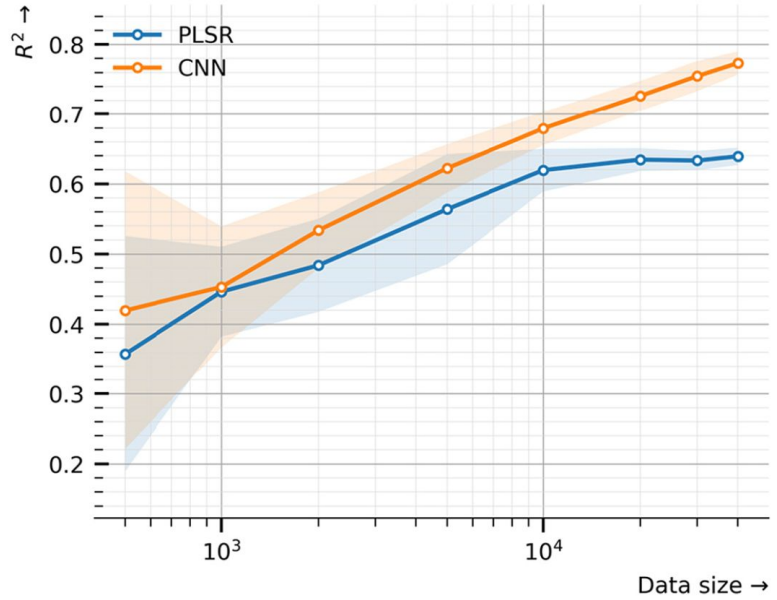


Fig. 3. Learning curves of PLSR and CNN models trained to predict K_{ex} in a growing data regime using all Soil Taxonomy Orders. The mean and the standard deviation of the coefficient of determination R^2 calculated on 20 different test splits as the size of the dataset increases are reported.

Some metrics

Table 3
Comparison of PLSR and CNN models performance[a] in predicting K_{ex} when trained on all Soil Taxonomy Orders and tested on individual Soil Taxonomy Orders separately. Figures reported are averages and standard deviations of metrics computed on test sets over 20 different train/validation/test splits. N is the number of samples. Performances on Oxisols are not reported as the size of its test set contains often a single sample.

			Metrics			
Orders	N	Models	R^2	LCCC	RMSE	MAPE
All	4032	PLSR	0.64 ± 0.01	0.780 ± 0.005	1.06 ± 0.45	135.0 ± 3.3
		CNN	0.79 ± 0.08	0.884 ± 0.005	0.60 ± 0.16	30.9 ± 0.9

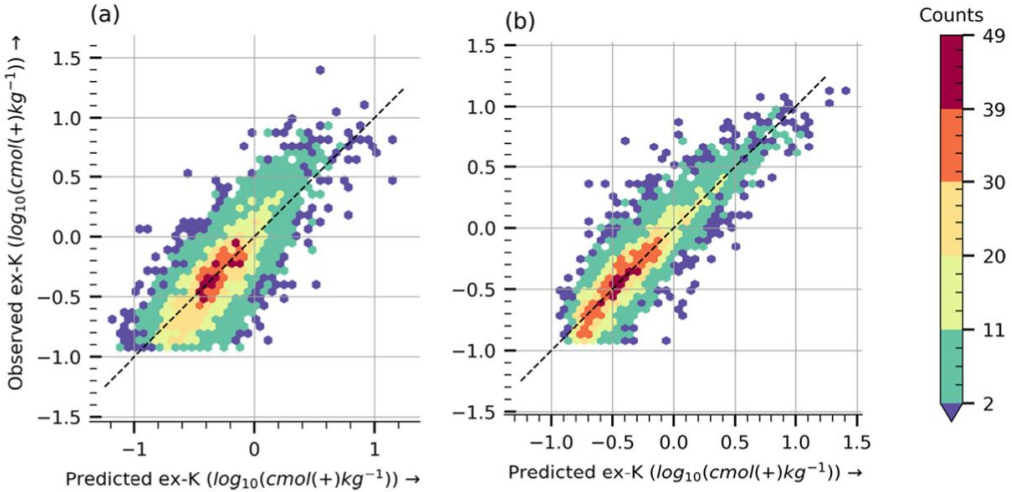


Fig. 4. Observed vs. Predicted K_{ex} with (a) PLSR and (b) CNN using all Soil Taxonomy Orders.

Model-centric: iterate on model ↗ R^2 ~ 0.84

```
dblock = DataBlock(blocks=(SpectraBlock, AnalyteBlock(analytes=[725])),
                   splitter=RandomSplitter(),
                   item_tfms=[RandWAvgTfm(), NormalizeTfm(mean=0.8813, std=0.4324),
                              LogTfm()])

dls = dblock.dataloaders(path.ls(),
                         bs=16)

model = xresnet18(ndim=1, c_in=1, ks=3, n_out=1).to(device)
learn = Learner(dls, model, loss_func=MSELossFlat(), metrics=R2Score())
learn.lr_find()
learn.fit_one_cycle(40, 1e-3)
```

PyTorch fast.ai ...

XResNet
xresnet50_deeper
xresnet34_deeper
xresnet18_deeper
xresnet50_deep
xresnet34_deep
xresnet18_deep
xresnet152
xresnet101
xresnet50
xresnet34
xresnet18
xse_resnext50_deeper
xse_resnext34_deeper
xse_resnext18_deeper
xse_resnext50_deep
xse_resnext34_deep
xse_resnext18_deep
xsenet154
xse_resnet152
xresnext101
xse_resnext101
xse_resnet101
xresnext50
xse_resnext50
xse_resnet50
xresnext34
xse_resnext34
xse_resnet34
xresnext18

https://docs.fast.ai

All good but …

Large-scale deployment would tell a different story

Iterate around data rather than model mainly

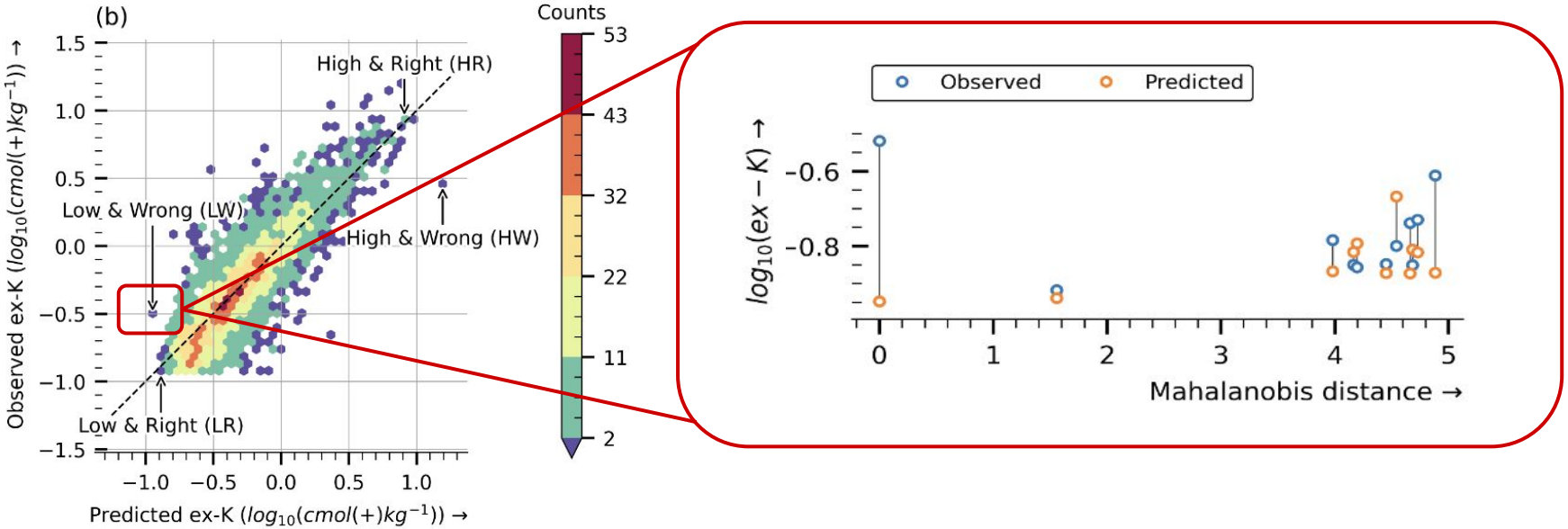
Error analysis is crucial lever to activate

3. Model- vs. Data-centric approach

Data-centric approach: iterate on data

1. to identify lack of consistency of annotation (wet chemistry)
2. To debug algorithm & model
3. To gain further insight on data (distribution, typology)
4. To inform further data collection

Annotation error (wet chemistry)?



What are the performance by Soil Taxonomy orders?

Table 3
Comparison of PLSR and CNN models performance[a] in predicting K_{ex} when trained on all Soil Taxonomy Orders and tested on individual Soil Taxonomy Orders separately. Figures reported are averages and standard deviations of metrics computed on test sets over 20 different train/validation/test splits. N is the number of samples. Performances on Oxisols are not reported as the size of its test set contains often a single sample.

			Metrics			
Orders	N	Models	R^2	LCCC	RMSE	MAPE
All	4032	PLSR	0.64 ± 0.01	0.780 ± 0.005	1.06 ± 0.45	135.0 ± 3.3
		CNN	0.79 ± 0.08	0.884 ± 0.005	0.60 ± 0.16	30.9 ± 0.9
Undefined	1553	PLSR	0.65 ± 0.01	0.788 ± 0.007	1.15 ± 0.52	153.0 ± 7.2
		CNN	0.81 ± 0.01	0.897 ± 0.006	0.76 ± 0.28	31.6 ± 1.3
Mollisols	977	PLSR	0.60 ± 0.02	0.742 ± 0.012	0.77 ± 0.20	94.3 ± 3.1
		CNN	0.78 ± 0.02	0.868 ± 0.014	0.44 ± 0.07	27.4 ± 1.2
Alfisols	422	PLSR	0.54 ± 0.05	0.728 ± 0.029	0.41 ± 0.18	82.6 ± 4.4
		CNN	0.69 ± 0.03	0.822 ± 0.018	0.38 ± 0.17	27.3 ± 1.3
Inceptisols	289	PLSR	0.54 ± 0.04	0.715 ± 0.021	0.56 ± 0.12	117.6 ± 6.8
		CNN	0.72 ± 0.03	0.838 ± 0.023	0.40 ± 0.07	35.0 ± 3.1
Ultisols	192	PLSR	0.30 ± 0.08	0.581 ± 0.056	0.32 ± 0.05	80.9 ± 9.0
		CNN	0.62 ± 0.05	0.763 ± 0.038	0.26 ± 0.06	32.5 ± 2.4
Entisols	165	PLSR	0.54 ± 0.10	0.734 ± 0.051	0.44 ± 0.09	124.3 ± 15.2
		CNN	0.77 ± 0.05	0.875 ± 0.033	0.32 ± 0.05	30.8 ± 3.4
Aridisols	163	PLSR	0.45 ± 0.16	0.686 ± 0.067	2.13 ± 2.74	145.9 ± 56.2
		CNN	0.69 ± 0.04	0.822 ± 0.022	0.66 ± 0.32	35.1 ± 3.1
Andisols	133	PLSR	0.59 ± 0.03	0.729 ± 0.028	0.59 ± 0.29	106.4 ± 11.3
		CNN	0.74 ± 0.04	0.856 ± 0.026	0.48 ± 0.15	32.6 ± 3.3
Vertisols	95	PLSR	0.55 ± 0.10	0.732 ± 0.062	0.39 ± 0.10	98.7 ± 12.3
		CNN	0.75 ± 0.06	0.857 ± 0.037	0.27 ± 0.06	26.9 ± 3.3
Histosols	80	PLSR	0.64 ± 0.09	0.778 ± 0.055	1.16 ± 0.38	243.1 ± 46.1
		CNN	0.76 ± 0.06	0.870 ± 0.029	0.87 ± 0.27	45.4 ± 7.4
Spodosols	64	PLSR	0.70 ± 0.08	0.824 ± 0.043	0.48 ± 0.13	157.1 ± 19.7
		CNN	0.78 ± 0.05	0.880 ± 0.025	0.41 ± 0.11	37.2 ± 4.7
Gelisols	61	PLSR	0.63 ± 0.10	0.767 ± 0.060	0.65 ± 0.14	193.8 ± 31.4
		CNN	0.75 ± 0.08	0.860 ± 0.047	0.59 ± 0.14	47.3 ± 9.9

[a] N is the number of samples in the test set. The coefficient of determination R^2 and Lin's concordance correlation coefficient (LCCC) are unitless; the Root Mean Square Error (RMSE) is expressed in cmol(+).kg^{-1}. The Mean Absolute Percentage Error (MAPE) is expressed in %. Both the mean and standard deviation of the metrics calculated over the 20 different random splits are reported.

A thorough analysis of the learning process

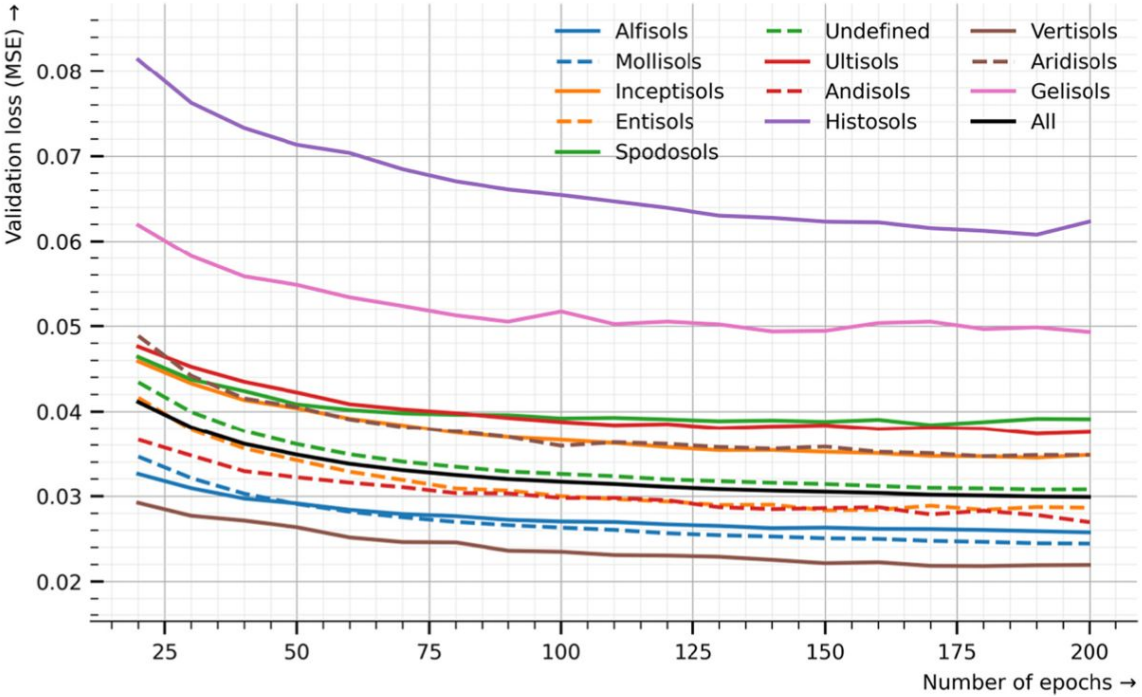
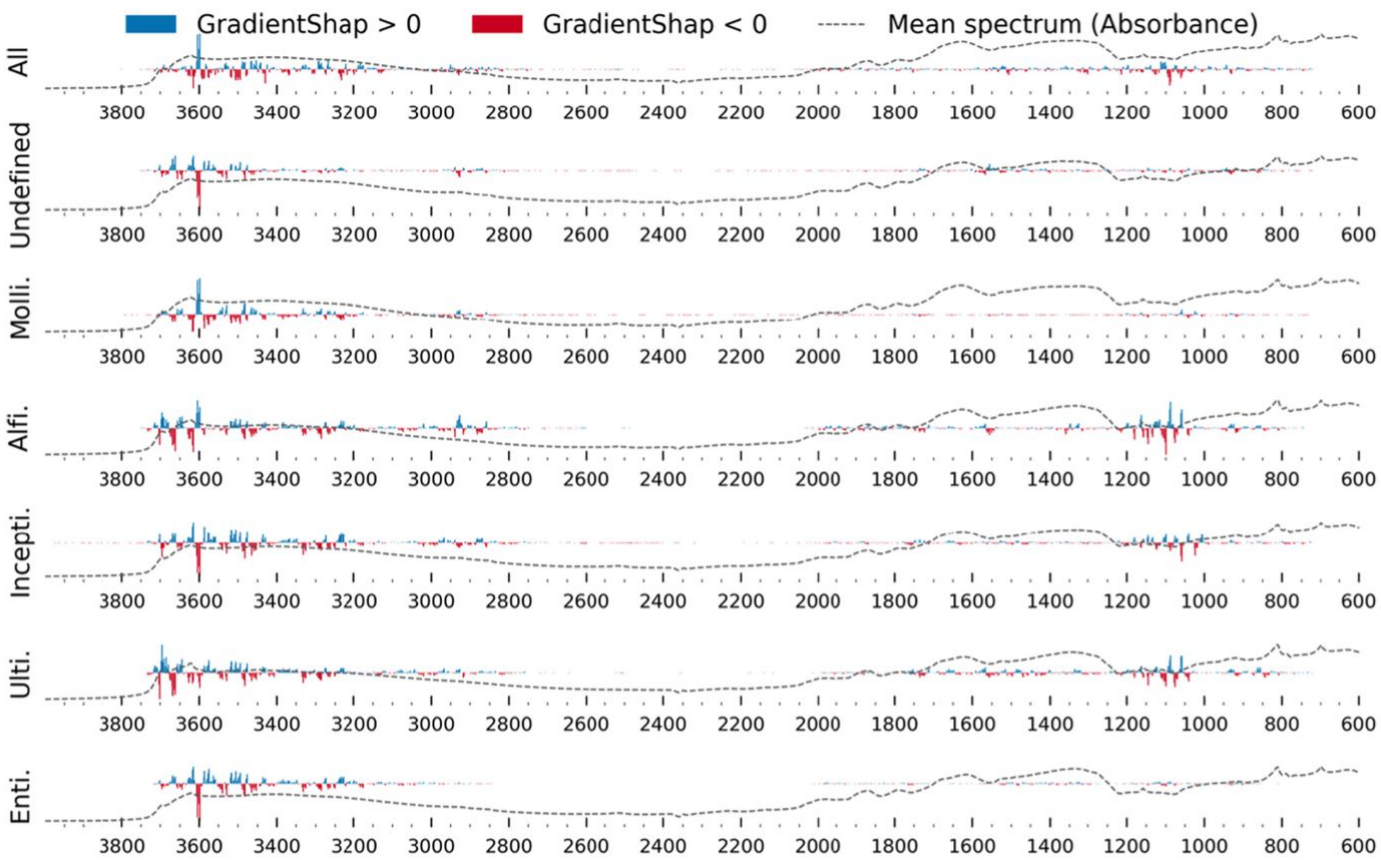
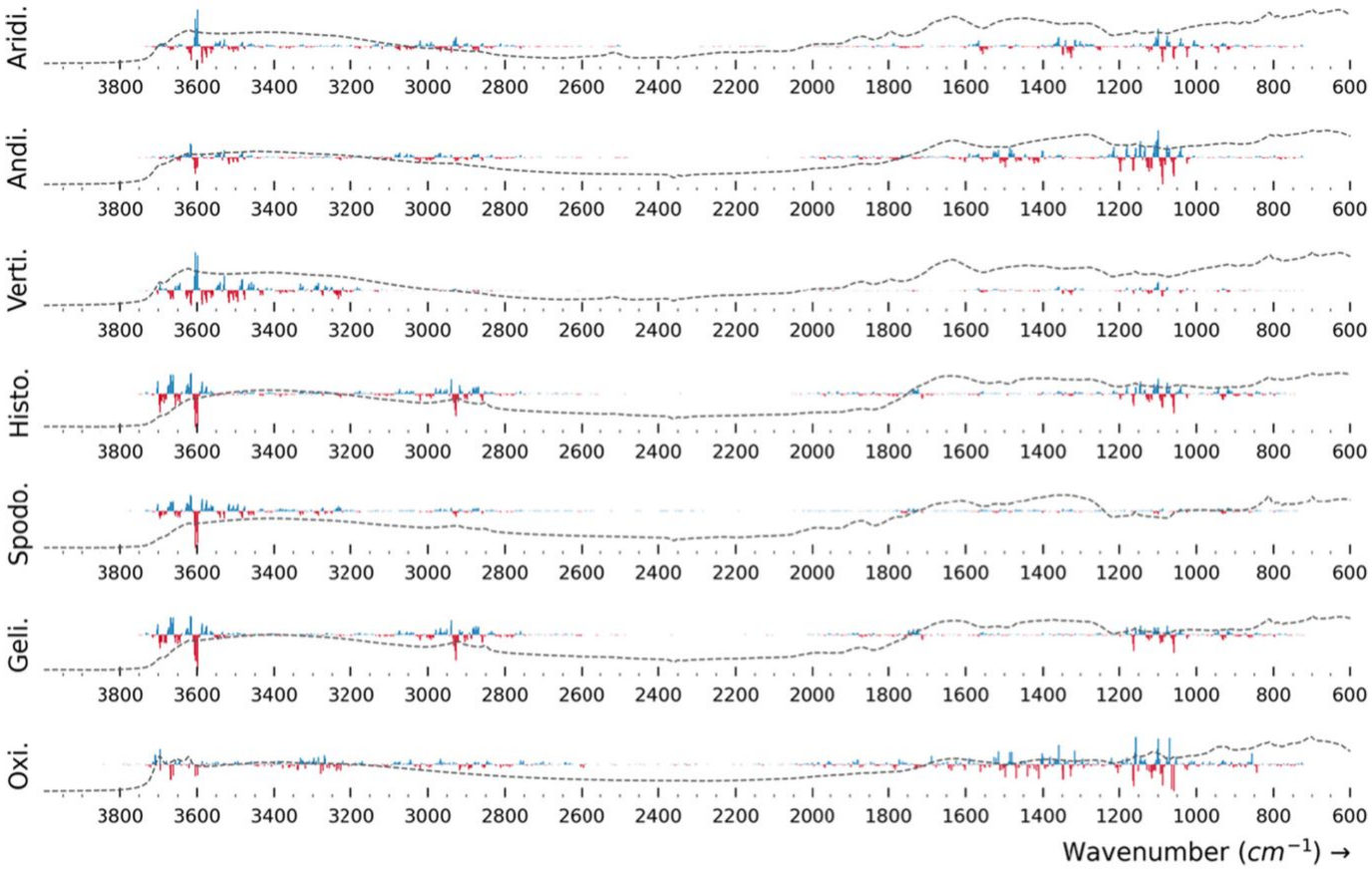


Fig. 5. Validation loss (MSE: Mean Squared Error) of the CNN during the training process (200 epochs) calculated on the entire validation set (All) and by Soil Taxonomy Orders. The different curves represent the mean MSE over 20 different random splits.

What does the model consider important?



What does the model consider important?



Similarity between Soil Taxonomy Orders?

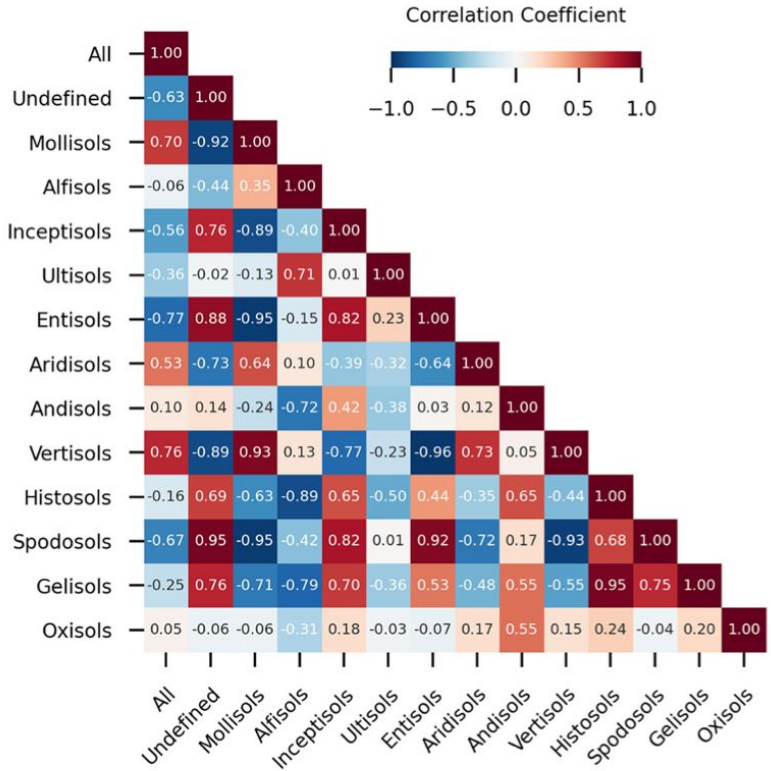


Fig. 8. Correlation matrix of GradientShap values computed based on the CNN model prediction of K_{ex} (log_{10}-transformed) by Soil Taxonomy Orders.

Data-centric approach: iterate on data (illustrated)

1. to identify lack of consistency of annotation (wet chemistry)
2. To debug algorithm & model
3. To gain further insight on data (distribution, typology)
4. To inform further data collection

The key reasons we use DL & CNN

1. Reuse learned features across Soil Taxonomy Orders
2. Tap into vast amount of model zoology & best practices
3. Introspection allowed
4. Transfer Learning (instrument cross-calibration also?)

A very different mindset

Domain expertise downstream rather than upstream

Don’t inject a priori domain expertise

Let the Neural Net learn it >> Data greedy (variability needed)

Reuse, reuse, reuse, …

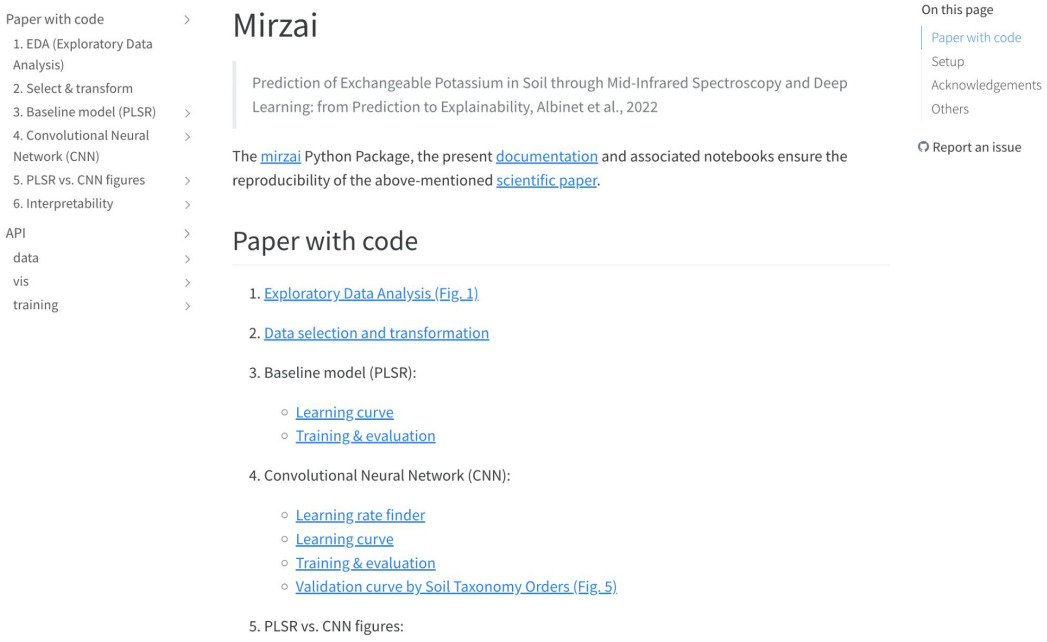
Our path ahead

Multiscale Characterization of Exchangeable Potassium Content in Soil to Remediate Agricultural Land Affected by Radioactive Contamination using Machine Learning , Soil Spectroscopy and Remote Sensing.

Promotor: Erik Smolders (KU-Leuven)
Co-Promotor(s): Raphael Viscarra Rossel (Curtin University, Australia)
Adviser(s): Gerd Dercon (Joint FAO/IAEA Center)

- MIRS + NIRS + Remote Sensing (Multi+hyperspectral) +soil forming predictors
- Taking up the challenge of transfer learning, instrument cross-calibration along the way…
- Leveraging FAO/IAEA SSLs on underrepresented soils (e.g Andisols)
- …

Thank you for the attention!



https://fr.anckalbi.net/mirzai