Supplementary information: Estimating soil fungal abundance and diversity at a macroecological scale with deep learning spectrotransfer functions

Yuanyuan Yang[1], Zefang Shen[1], Andrew Bisset[2], and Raphael A. Viscarra Rossel[1]

[1]Soil and Landscape Science, School of Molecular and Life Sciences, Curtin University, GPO Box U1987, Perth WA 6845, Australia.
[2]CSIRO Oceans and Atmosphere, GPO BOX 1538, Hobart TAS 7001, Australia.

Correspondence: Raphael A. Viscarra Rossel (r.viscarra-rossel@curtin.edu.au)

Rarefaction curves

The BASE dataset sought to produce as many sequences as resources allow with a minimum sequencing number of 10,000 per sample. Here, each sample was re-sampled at depth of 11 000 sequences to eliminate the unbalanced sequencing (Fig. S1). We chose 11 000 sequences as re-sampling depth mainly because many samples only had this sequences number but also the rate of increase in the rarefaction curves is small at this depth.

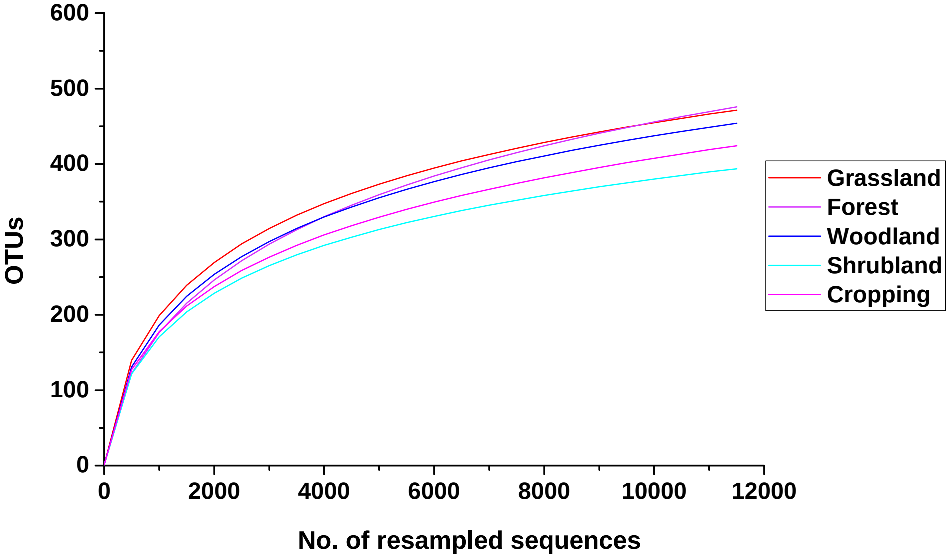

Figure 1. Rarefaction curves showing richness accumulated in terms of the observed OTUs per ecosystem types. We have showed rarefaction curves for only 11000 sequences.

Edaphic and environment covariates

Table 1. The soil, environmental, and visible–near infrared (vis–NIR) covariates used in the modelling, their resolution and the source of the data.

Set	Predictors	Resolution	Source
Digital soil maps	Total organic carbon (TOC)/ %	90 m	Viscarra Rossel et al. (2015)
	Total nitrogen (TN)/ %	90 m	
	Total phosphorus (TP)/ %	90 m	
	Bulk density (BD)/ g cm^{-3}	90 m	
	Cation exchange capacity (CEC)/ meq/100 g	90 m	
	Available water content (AWC)	90 m	
	pH	90 m	
	Sand/ %	90 m	
	Silt/ %	90 m	
	Clay/ %	90 m	
	Kaolinite/ rel. abundance	90 m	Viscarra Rossel (2011)
	Illite/ rel. abundance	90 m	
	Smectite/ rel. abundance	90 m	
Climate	Mean annual temperature (MAT)/ °C	90 m	Xu and Hutchinson (2011)
	Mean annual precipitation (MAP)/ mm	90 m	
	Potential evapotranspiration (PET)/ °C	90 m	
	Mean annual solar radiation (SolarR)/ J m^{-2} yr^{-1}	90 m	
	Prescott index (PI)	90 m	Prescott (1950)
Terrain	Elevation (DEM)/ m	90 m	Gallant et al. (2011)
	Topological Wetness Index (TWI)	90 m	
	Aspect/ °	90 m	
	Relief/ m	90 m	
	Slope/ °	90 m	
Vegetation	Fpar-raingreen (Fpar-r)	250 m	Donohue et al. (2009)
	Fpar-evergreen (Fpar-e)	250 m	
	Net primary productivity (NPP)/ g C m^{-2} yr^{-1}	1 km	Zhao et al. (2005)
Parent material	Thorium (GammaTh)/ mg kg^{-1}	100 m	Minty et al. (2009)
	Uranium (GammaU)/ mg kg^{-1}	100 m	
	Potassium (GammaK)/ mg kg^{-1}	100 m	
vis–NIR	Absorbance at 208 wavelengths	10 nm	

Algorithms of machine learning

The PLSR is a linear regression model widely used in the quantitative analysis of diffuse reflectance spectra in soil (Viscarra Rossel, 2008). This method uses a latent variable (known as component) approach to model covariance structures in two projected spaces of the predicted and observed variables (Wold et al., 2001). We performed PLSR using the `pls` library in the software R. Number of components parameter was tuned from 1 to 20 using 10-fold cross validation.

The SVM method employs classification and regression analysis to solve linear and nonlinear multivariate problems (Suykens et al., 2002). Here, a Kernel function of Gaussian radial basis function (RBF) was used. Parameters penalty (C) and gamma (γ) of the RBF were optimized during modeling. SVM was performed using the kernlab library in the software R.

The RF is a ensemble learning classification and regression algorithm consisting of many decisions trees (Breiman, 2001). It uses bagging and feature randomness when building each individual tree and merges them together to get a more accurate and stable prediction. RF prediction performance is sensitive to three user-defined parameters: the number of trees (ntree) in the forest, the minimum number of data in each node (nodesize), and the number of predictors tried at each node (mtry). RF model was performed using the RandomForest library in the software R.

The GPR model, a form of Bayesian non-linear regression (Rasmussen and Williams, 2005), was trained using using the kernlab library in the software R. A GPR model is defined primarily by the selection of a covariance function, which defines how the expected value of the target variable changes as values change across the input space. The covariance function contains several parameters, which are optimized during modeling including a length-scale for each feature (l), and a noise free signal variance (σ_f^2), the noise variance (σ_n^2).

The XGBoost was a scalable and efficient tree boosting systems (Friedman, 2001). The XGBoost algorithm is superior to the traditional gradient boosting machine method. Over-fitting was controlled with a more regular model formalization method for more reliable performance (Chen and Guestrin, 2016). The XGBoost model has been described in detail by Chen et al. (2019).The XGBoost library in the software R was used for building the XGBoost model. Several parameters including nrounds, eta, gamma,and subsample were optimized in the modeling.

The CUBIST model is a form of piece-wise linear decision tree (Quinlan, 1992),which we have used and described in some detail elsewhere (Viscarra Rossel and Webster, 2012). Briefly, CUBIST uses a recursive partitioning of the predictor variable space and partitions the data into subsets that are more similar with respect to the predictors in the data. A unique linear model is then applied to predict the response within each partition. The advantage of Cubist is that they enable different linear models to capture the linearity in different parts of the predictor variable space, leading to smaller, more interpretable. Two parameters including the committee models(C) and the number of neighbouring observations(N) were adjusted during modeling.

A convolutional neural network (CNN) consists of multiple processing layers which can extract representations of the input data at various abstract levels (Lecun et al., 2015). Its internal layers include convolutional layers, pooling layers and fully connected layers. A convolutional layer scans its input with multiple filters and generates corresponding feature maps; A pooling layer down-samples its input for dimension reduction and invariance to small shifts; Fully-connected layers follows to calculate the model outputs. The architecture of the CNNs brings about several advantages: local correlation (Lecun et al., 2015), minimal preprocessing (LeCun et al., 1990, 1995), and a high number of connections with a low number of free parameters (LeCun et al., 1990).

Convolutional neural networks have numerous applications across disciplines, such as natural language processing (Kim, 2014; Kalchbrenner et al., 2014; Collobert et al., 2011), object detection and recognition (Gonzalez, 2007; Szegedy et al., 2015), drug discovery (Wallach et al., 2015), etc. Recent studies have also exploited CNNs for soil spectroscopy (Veres et al., 2015) One-dimensional CNNs (1D-CNNs) and two-dimensional CNNS (2D-CNNs) are commonly used for soil property predictions (Liu et al., 2018; Ng et al., 2019; Padarian et al., 2019; Tsakiridis et al., 2020; Veres et al., 2015). One-dimensional CNNs take raw spectra data or preprocessed 1D array as inputs whereas 2D-CNNs process spectrograms generated from raw spectra. One-dimensional CNNs outperform other models such as Partial Least Squares regression, Cubist and Support Vector Regression, including 2D-CNNs in soil property prediction (Ng et al., 2019; Tsakiridis et al., 2020). This might because the 1D-CNNs can effectively exploit the local correlations between the adjacent spectral wavelengths (Veres et al., 2015).

Convolutional neural networks (CNNs) consist of multiple processing layers which allows CNNs to learning increasingly complex representations (Lecun et al., 2015). Recent studies showed that one dimensional neural networks (1D-CNNs) produced more accurate soil property predictions than other statistical and machine learning methods (Liu et al., 2018; Tsakiridis et al., 2020; Veres et al., 2015). Here, we developed a 1D-CNN for each target using the automated hyperparameter tuning framework for 1D-CNNs (Shen & Viscarra Rossel, 2021). We optimised hyperparameters: number of convolutional, pooling, and fully-connected layers; kernel size, number of filters, padding type (Same or Valid), strides, and activation in convolutional layers; pool type (AveragePooling or MaxPooling), pool size, padding type and strides in pooling layers; Number of units and activation in fully-connected layers; and dropout rates. In this study, the 1D-CNNs were developed using the deep learning framework TensorFlow (Abadi et al., 2016).

Architectures and optimised hyperparameters of the 1D-CNNs spectro-transfer functions

The 1D-CNN architectures are given in Table 2. The 1D-CNNs consists of a number of Convolutional layers and Fully-connected layers, joined by a Flatten layer. In the case of Glomeromycoata, pooling layers were also used. A Dropout layer was also used after each Convolutional and Fully-connected layer to prevent overfitting.

Table 2. Architectures of the 1D-CNN spectro-transfer functions

Phyla and diversity	Layer type	Kernel size	Filters/Units	Padding	Strides	Activation
Ascomycota	Convolutional	(5,1)	48	Same	2	Swish
	Convolutional	(3,1)	126	Same	1	Swish
	Flatten	-	-	-	-	-
	Fully-connected	-	509	-	-	ELU
	Fully-connected	-	239	-	-	ELU
	Fully-connected	-	141	-	-	SELU
	Fully-connected	-	1	-	-	Linear
Basidiomycota	Convolutional	(6,1)	102	Valid	4	SELU
	Convolutional	(3,1)	99	Valid	2	ReLU
	Flatten	-	-	-	-	-
	Fully-connected	-	283	-	-	ELU
	Fully-connected	-	184	-	-	ELU
	Fully-connected	-	98	-	-	SELU
	Fully-connected	-	1	-	-	Linear
Mortierellomycota	Convolutional	(3,1)	4	Same	2	SELU
	Flatten	-	-	-	-	-
	Fully-connected	-	404	-	-	ReLU
	Fully-connected	-	399	-	-	SELU
	Fully-connected	-	83	-	-	ELU
	Fully-connected	-	1	-	-	Linear
Glomeromycota	Convolutional	(7,1)	71	Valid	2	ReLU
	Convolutional	(3,1)	84	Same	1	LeakyReLU
	AveragePooling	(7,1)	-	Same	1	-
	Convolutional	(6,1)	48	Same	1	LeakyReLU
	Convolutional	(4,1)	124	Valid	3	ReLU
	MaxPooling	(4,1)	-	Same	3	-
	Flatten	-	-	-	-	-
	Fully-connected	-	213	-	-	SELU
	Fully-connected	-	80	-	-	ELU
	Fully-connected	-	33	-	-	ELU
	Fully-connected	-	1	-	-	Linear
Mucoromycota	Convolutional	(5,1)	121	Same	4	SELU
	Convolutional	(8,1)	17	Valid	1	Swish
	Convolutional	(3,1)	24	Same	1	SELU
	Flatten	-	-	-	-	-
	Fully-connected	-	183	-	-	SELU
	Fully-connected	-	137	-	-	ELU
	Fully-connected	-	115	-	-	Swish
	Fully-connected	-	1	-	-	Linear
Diversity	Convolutional	(2,1)	66	Same	2	LeakyReLU
	Flatten	-	-	-	-	-
	Fully-connected	-	469	-	-	SELU
	Fully-connected	-	321	-	-	ReLU
	Fully-connected	-	255	-	-	Swish
	Fully-connected	-	1	-	-	Linear

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning, in: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp. 265–283, 2016.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Chen, S., Liang, Z., Webster, R., Zhang, G., Zhou, Y., Teng, H., Hu, B., Arrouays, D., and Shi, Z.: A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution., Science of The Total Environment, 655, 273–283, 2019.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, pp. 785–794, 2016.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P.: Natural language processing (almost) from scratch, Journal of Machine Learning Research, 12, 2493–2537, 2011.

Donohue, R. J., McVicar, T., and Roderick, M. L.: Climate-related trends in Australian vegetation cover as inferred from satellite observations, 1981–2006, Global Change Biology, 15, 1025–1039, 2009.

Friedman, J. H.: Greedy function approximation: A gradient boosting machine., Annals of Statistics, 29, 1189–1232, 2001.

Gallant, J., Wilson, N., Dowling, T., Read, A., and Inskeep, C.: SRTM-derived 1 second digital elevation models version 1.0, Geoscience Australia: Canberra, ACT, 2011.

Gonzalez, T. F.: ImageNet Classification with Deep Convolutional Neural Networks Alex, Handbook of Approximation Algorithms and Metaheuristics, pp. 1–1432, https://doi.org/10.1201/9781420010749, 2007.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P.: A convolutional neural network for modelling sentences, 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference, 1, 655–665, https://doi.org/10.3115/v1/p14-1062, 2014.

Kim, Y.: Convolutional neural networks for sentence classification, EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 1746–1751, https://doi.org/10.3115/v1/d14-1181, 2014.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D.: Handwritten digit recognition with a back-propagation network, in: Advances in neural information processing systems, pp. 396–404, 1990.

LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks, 3361, 1995, 1995.

Lecun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, https://doi.org/10.1038/nature14539, 2015.

Liu, L., Ji, M., and Buchroithner, M.: Transfer learning for soil spectroscopy based on convolutional neural networks and its application in soil clay content mapping using hyperspectral imagery, Sensors (Switzerland), 18, https://doi.org/10.3390/s18093169, 2018.

Minty, B., Franklin, R., Milligan, P., Richardson, M., and Wilford, J.: The radiometric map of Australia, Exploration Geophysics, 40, 325–333, 2009.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., and McBratney, A. B.: Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra, Geoderma, 352, 251–267, https://doi.org/10.1016/j.geoderma.2019.06.016, 2019.

Padarian, J., Minasny, B., and McBratney, A. B.: Using deep learning to predict soil properties from regional spectral data, Geoderma Regional, 16, e00 198, https://doi.org/10.1016/j.geodrs.2018.e00198, 2019.

Prescott, J. A.: A climatic index for the leaching factor in soil formation, Journal of Soil Science, 1, 9–19, 1950.

Quinlan, J. R.: Learning with continuous classes, in: 5th Australian joint conference on artificial intelligence, vol. 92, pp. 343–348, Singapore, 1992.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2005.

Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D., and Vandewalle, J.: Least Squares Support Vector Machines, 2002.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: Going deeper with convolutions, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015, 1–9, https://doi.org/10.1109/CVPR.2015.7298594, 2015.

Tsakiridis, N. L., Keramaris, K. D., Theocharis, J. B., and Zalidis, G. C.: Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network, Geoderma, 367, 114 208, https://doi.org/10.1016/j.geoderma.2020.114208, 2020.

Veres, M., Lacey, G., and Taylor, G. W.: Deep Learning Architectures for Soil Property Prediction, Proceedings -2015 12th Conference on Computer and Robot Vision, CRV 2015, pp. 8–15, https://doi.org/10.1109/CRV.2015.15, 2015.

Viscarra Rossel, R. A.: ParLeS : Software for chemometric analysis of spectroscopic data, Chemometrics and Intelligent Laboratory Systems, 90, 72–83, 2008.

Viscarra Rossel, R. A.: Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra, Journal of Geophysical Research: Earth Surface, 116, F04 023, 2011.

Viscarra Rossel, R. A. and Webster, R.: Predicting soil properties from the Australian soil visible–near infrared spectroscopic database, European Journal of Soil Science, 63, 848–860, 2012.

Viscarra Rossel, R. A., Chen, C., Grundy, M., Searle, R., Clifford, D., and Campbell, P.: The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project, Soil Research, 53, 845–864, 2015.

Wallach, I., Dzamba, M., and Heifets, A.: AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery, pp. 1–11, http://arxiv.org/abs/1510.02855, 2015.

Wold, S., Sjostrom, M., and Eriksson, L.: PLS-REGRESSION: A BASIC TOOL OF CHEMOMETRICS, Chemometrics and Intelligent Laboratory Systems, 58, 109–130, 2001.

Xu, T. and Hutchinson, M.: ANUCLIM version 6.1 user guide, The Australian National University, Fenner School of Environment and Society, Canberra, 2011.

Zhao, M., Heinsch, F. A., Nemani, R. R., and Running, S. W.: Improvements of the MODIS terrestrial gross and net primary production global data set, Remote Sensing of Environment, 95, 164–176, 2005.