Machine Learning for Soil Attribute Prediction: An Effectiveness and Dimensionality Reduction Analysis

José Solenir L. Figuerêdo[1], **Marcos Eduardo de C. Ferreira**[1], **Rodrigo T. Calumby**[1]

[1]Department of Exact Sciences, University of Feira de Santana
Feira de Santana – BA – Brazil

{solenir.figueredo, eduarddferreira}@gmail.com, rtcalumby@uefs.br

Abstract. *Traditional soil fertility analyzes are laborious, expensive, time-consuming and produce hazardous waste. Although many works using machine learning (ML) has been done to address these issues, some algorithms and dimensionality reduction strategies require further investigation. Therefore, in this study we evaluated the potential of Support Vector Regression and Ridge regression in determining soil attributes, and compared principal components regression and partial least squares regression (PLSR). The results showed that Ridge was the most effective model. In addition, our experiments revealed that PLSR was able to achieve statistically equivalent results, and in some cases superior to the baseline, but using a much smaller average number of components.*

1. Introduction

Soil fertility is essential to achieve efficient and profitable agricultural production. Therefore, a high-density and fine-scale monitoring of soil properties[1] is a crucial step, as it allows the construction of soil maps to guide farmers in better management decisions in crop fields [Wollenhaupt et al. 1994, Wei et al. 2022]. In practice, several laboratory analyzes are necessary to diagnose soil fertility before determining which – and the amounts of – soil amendments and fertilizers to use. However, these traditional approaches are expensive and time-consuming, making it hard to increase the density of soil data analysis [Benedet et al. 2021]. Additionally, these strategies produce hazardous waste due to the use of strong acids, strong bases, and other chemical reagents. Hence, it is crucial to identify alternative approaches to assess and potentially predict the nutrient levels available in soils.

In this context, researchers have investigated different soil sensing techniques and their applicability in agriculture, such as proximal detection technologies. Proximal soil sensing technologies such as visible and near-infrared diffuse reflectance spectroscopy (Vis-NIR) and X-ray fluorescence spectroscopy (XRF) are dry chemical techniques that

[1]Soil attributes and properties are used interchangeably in this work.

allow for rapid and ecological analyzes of soil fertility [Tavares et al. 2022]. The central idea is to use these spectral data to predict agronomic attributes. To that purpose, machine learning (ML) techniques have been applied. The objective is to use these data to build specific calibrations (models), transforming spectral data into predictions of soil physical and chemical properties [Folorunso et al. 2023].

In recent years, several works have been developed, e.g., considering the task a classification problem [Suchithra and Pai 2020, M and C D 2021, Sunori et al. 2022] or predicting the physical and chemical properties of the soil as a regression problem [Benedet et al. 2021, Wei et al. 2022]. Some works have focused on preprocessing techniques, e.g., dimensionality reduction [Laili et al. 2020, Wei et al. 2022]. Indeed, spectral data often have many variables (e.g., n $>$ 300) [Wei et al. 2022], while some may not effectively contribute to the prediction. In some cases, they may even decrease the model effectiveness, due to the curse of dimensionality [Verleysen and François 2005]. Besides, it is important to consider the cost of processing these variables in a real-time setting and the trade-off with effectiveness gains.

Despite some recent advances, current methods are still not suitable for real-world practice considering the limited effectiveness and the large variations in success levels for different soil properties. In addition, many works use simple or optimistic experimental validation processes. The effectiveness of the model is evaluated using only one validation set. While this approach is common, it may lead to misinterpretations when applied to small datasets like the one utilized in [Wei et al. 2022], which comprises only 102 samples. After all, owing to the stochastic nature of data partitioning, the model's effectiveness may have been confined solely to that specific subset. Therefore, novel techniques must be proposed to allow better predictions and be validated with strict protocols for reliable conclusions. For example, alternative data preprocessing strategies could be evaluated as well as other ML algorithms.

In this work, we evaluated the partial least squares regression (PLSR) as a dimensionality reduction strategy. Unlike principal component analysis (PCA), which creates composite variables based only on the independent variables, PLSR also considers the dependent variable, so that composite variables have higher correlations with the target attribute [Liu et al. 2022]. In this sense, it is expected that the PLSR would contribute to the generation of more robust predictive models, while using fewer components than the PCA. Moreover, this work experimentally assessed the effectiveness of alternative ML algorithms, specifically Support Vector Regression (SVR) and Ridge regression, in determining soil physical and chemical attributes. Ridge regression, for example, can be used to eliminate multicollinearity, and thereby better define the relationship between independent and dependent variables.

2. Related Works

In [Benedet et al. 2021], Generalized Linear Model (GLM) and Random Forest (RF) were used to generate and fit predictive models able to explain some soil properties related to fertility (available potassium (av.K), available phosphorus (av.P), exchangeable Ca^{2+} (ex.Ca), exchangeable Mg^{2+} (ex.Mg), exchangeable Al^{3+} (ex. Al) and remaining phosphorus (P-rem)), which were extracted from different types of soil. The models were evaluated from multiple measures, such as determination coefficient (R^2), root mean square

error (RMSE), and mean absolute error (MAE). Among the generated models, the RF achieved the best effectiveness. Considering R^2 (0.49, 0.56, 0.68, 0.70, 0.71, 0.83 for ex. Al, av. K, av. P, ex. Mg, P-rem, ex. Ca, respectively).

[Yang et al. 2020] were able to predict several soil properties (P, N, OC, K and pH) using ML methods such as artificial neural network (ANN), PLSR, SVR, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). The best architecture was CCNVR, which combines CNN and RNN, with the lowest RMSE value (6.40, 0.45, 3.30 and 0.35 for OC, N, CEC and pH, respectively) and the highest R^2 (0.73, 0.70, 0.73 and 0.86 for OC, N, CEC and pH, respectively).

In [M and C D 2021], soil fertility was classified as low, medium and high, based on soil attributes: pH, EC, OC, P, K, S, Zn, B, Fe, Cu and Mn. A set of classifiers such as Naive Bayes, Logistic Regression, Support Vector Machines, Decision Trees, Boosted Regression Tree (BRT) and RF were used to estimate fertility levels. In general, the RF classifier outperformed the other classifiers considering multiple measures.

In [Wei et al. 2022], prediction models were adjusted from regression methods with the proposal to predict soil properties such as Clay, OM, CEC, pH, V(%), P, K, Ca and Mg, which are related to soil fertility. The best values of R^2 were obtained by the principal components regression (PCR) and Lasso regression, ranging from 0.33 to 0.96 and 0.03 to 0.84 for XRF and Vis-NIR sensors, respectively.

Despite the recent results, there are still some limitations to be tackled. Although some of these studies, for instance [Yang et al. 2020], have applied a cross-validation process for defining hyperparameters, the effectiveness of the model is evaluated using only one validation set. Moreover, it is worthwhile to explore algorithms that incorporate regularization strategies such as Ridge regression or those capable of leveraging nonlinearity in the data like SVR. Additionally, although some works explore dimensionality reduction, there is still a lack of analysis that compares different reduction strategies. Thus, in this study, we explored SVR and Ridge regression in the task of determining soil physical and chemical attributes and compared them to PLSR and PCR, which have demonstrated promising effectiveness in previous works [Wei et al. 2022].

3. Experimental Workflow and Settings

The experimental process proposed in this work is illustrated in Figure 1. There are four stages: Data Collection (Section 3.1), preprocessing and model training (including optimization and validation) (Section 3.2), and, finally, the evaluation and analysis of the model (Section 3.3).

3.1. Data Collection

The dataset used comprises 102 soil samples from the soil database of the Laboratory of Precision Agriculture at the Luiz de Queiroz College of Agriculture at the University of São Paulo [Tavares et al. 2022]. Two fields used for agricultural production had samples collected from 0 to 20 cm deep; 58 samples were collected in the municipality of Piracicaba, State of São Paulo (Field 1), and the remainder n = 44 in the municipality of Campo Novo do Parecis, Mato Grosso (Field 2).

The samples were analyzed following Vis-NIR and XRF analysis techniques. Soil fertility analyzes were carried out in a commercial laboratory, where the following vari-

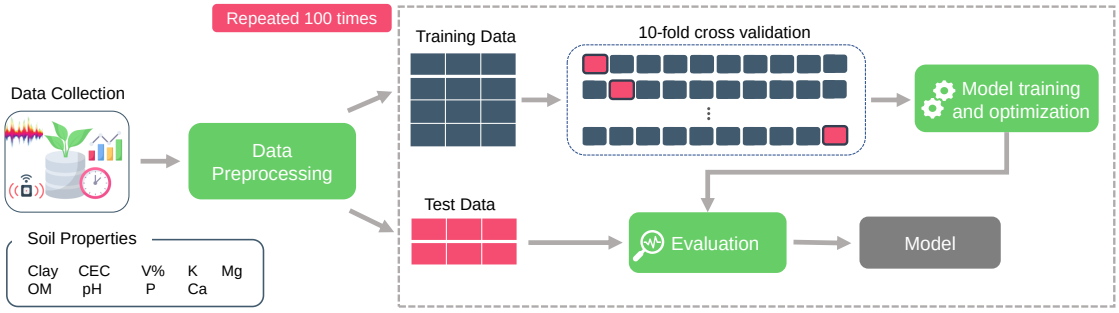

Figure 1. Experimental workflow conducted in this work.

ables intrinsically related to soil fertility were determined: Clay (g kg^{-1}); OM: organic matter content (g kg^{-1}); CEC: cation exchange capacity ($mmol_c kg^{-1}$); pH: potential of hydrogen; V(%): base saturation ($mmol_c kg^{-1}$); P: phosphorus ($mmol_c kg^{-1}$); K: potassium ($mmol_c kg^{-1}$); Ca: calcium ($mmol_c kg^{-1}$); Mg: magnesium ($mmol_c kg^{-1}$).

3.2. Data Preprocessing and Experimental Setup

This study used two main algorithms, i.e., SVR and Ridge regression. In addition, experiments were also carried out using PLSR and Ridge regression with PCA (named from now on as Ridge-PCA). Basically, PLSR is a technique that reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components instead of the original data [Liu et al. 2022]. PLSR is especially useful when the predictors are highly collinear. In our context, this algorithm is applied to generate a model capable of identifying the relation between the spectral measurements and the physical and chemical properties of soil. With this, it is expected to adjust the model using a highly reduced number of components, and still keep competitive effectiveness.

Table 1. Hyperparameters tested for each of the algorithms used in this work

Algorithm/Classifier	Hyperparameters	Tested values	Standardization
Ridge Regression	Alpha	[0,100] ÷ 100	No
Ridge-PCA	Alpha Components	[0,100] ÷ 100 [1,30]	Yes
Support Vector Regression	C Gamma Kernel	(0.1, 1, 10) (1,0.1,0.01) (rbf, linear, sigmoid)	No
PLSR	Components	[1,30]	Yes

The dataset was partitioned into training and test sets. It was randomly partitioned using the $75/25$ ratio, with 75% for training and the remaining 25% for the test set. This process was performed 100 times considering each of the soil properties separately. The training set was used to build the models that were optimized via hyperparameter optimization (Table 1) through grid search using cross-validation based on k-folds, with $k = 10$ and considering the R^2 as maximization criteria. The test set is used to evaluate the overall effectiveness of the models. Before carrying out the training, for some of the algorithms (PLSR and Ridge-PCA), the data were standardized using the z-score normalization ($z = (x-\mu)/s$), where x is the sample, μ is the mean of the training samples, and s is the standard deviation of the training samples. Preliminary experiments using standard-

ization for SVR and Ridge-PCA were also conducted but this decreased the effectiveness of the generated models, therefore z-score was not used with these algorithms.

The experimental process presented in Figure 1 was executed 100 times for each algorithm, considering each of the soil properties individually. Thus, the result reported in this work corresponds to the average of these executions. To keep comparative compatibility, the same procedure was applied to the baseline, described in Section 3.3.

3.3. Evaluation

The effectiveness of the models was assessed based on the Root Mean Squared Error (RMSE) and R^2. The models generated were compared with the ones developed in [Wei et al. 2022]. Specifically, we used the PCR method as a baseline, which achieved promising results by reducing the number of components. Specifically, PCR is a regression method that combines principal components analysis with least squares regression. For the strict comparison of the effectiveness results, the developed models were compared to the baseline using Wilcoxon's Signed Rank Test in order to assess the statistical significance of the results.

4. Results and Discussions

A general analysis of the models generated in this work in comparison to the baseline is presented in Section 4.1. Given the baseline mainly regards the process of dimensionality reduction, in Section 4.2 we provide a comparative analysis between the PLSR, Ridge-PCA and PCR.

4.1. General effectiveness analysis

Table 2 presents the effectiveness of the models developed in this study, as well as of the baseline. The statistically significant superiority of the evaluated algorithms in relation to the baseline is highlighted in bold. In turn, statistically significant inferiority is marked with "*". The remainder indicates statistical equivalence.

Considering the models generated in this study, Ridge regression achieved the best results, considering both sensors and both measures. It was even statistically higher than the baseline for all target properties, except for the *P* in Vis-NIR, and *OM* and *CEC* in XRF. Thus, in a scenario where the effectiveness of the model is the most important factor, the model obtained by Ridge would be the most suitable for practical use. It is noteworthy that, specifically for the *P*, none of the evaluated algorithms was able to generate a model that effectively captures the relationship between the input data and that property. Thus, future experiments must be conducted to deal specifically with this target.

As observed, in general Ridge regression achieved the best effectiveness considering individual properties, which may be the result of the underlying regularization process present in that method. When applying a penalty, the slope is reduced and, therefore, the model becomes less sensitive to changes in the independent variable [Saleh et al. 2019]. Therefore, the regression fits better to the data of interest, generating a more robust model. From these findings, some experiments were conducted to evaluate the use of PCA with Ridge regression. Considering the Vis-NIR sensor, the Ridge-PCA outperform the baseline for *CEC* and *Ca* attributes. In turn, taking XRF sensor, the Ridge-PCA was statistically superior to baseline for the *CEC* attribute. For the rest of the proprieties was

statistically equivalent. Therefore, when accurate predictions of *CEC* and *Ca* properties are crucial, Ridge-PCA is the recommended approach over the baseline.

Table 2. Experimental results of the algorithms evaluated in this work

		Baseline			PLSR			Ridge-PCA			Ridge		SVR	
Sensor	Target	#C	RMSE	R^2	#C	RMSE	R^2	#C	RMSE	R^2	RMSE	R^2	RMSE	R^2
Vis-NIR	Clay	19.79	30.808	0.882	10.8	30.937	0.880	23.23	*32.221	*0.872	**29.446**	**0.892**	*31.973	*0.874
	OM	19.87	3.185	0.696	9.96	3.184	0.695	24.14	*3.293	*0.677	**3.037**	**0.725**	**2.978**	**0.736**
	CEC	13.65	17.623	0.490	7.65	17.340	0.504	22.44	**16.758**	**0.540**	**16.645**	**0.543**	**16.431**	**0.561**
	pH	20.03	0.330	0.394	11	0.330	0.392	25.81	*0.356	*0.307	**0.270**	**0.591**	**0.296**	**0.511**
	V%	24.46	9.431	0.805	13.08	**9.095**	**0.819**	24.99	*10.350	*0.764	**8.019**	**0.858**	9.481	0.803
	P	3.34	13.890	-0.096	3.04	13.916	-0.107	3.77	**13.804**	**-0.082**	13.813	-0.127	*14.402	-0.162
	K	12.07	1.407	0.618	7.04	*1.439	*0.604	12.12	1.401	0.624	**1.358**	**0.646**	1.421	0.610
	Ca	16.58	10.982	0.646	8.83	11.059	0.641	24.85	**10.596**	**0.672**	**10.359**	**0.686**	**10.485**	**0.680**
	Mg	15.97	8.046	0.553	8.73	*8.281	*0.526	22.17	*8.239	0.536	**7.832**	**0.573**	8.056	0.557
	Average	17.80	10.227	0.636	9.64	10.208	0.632	22.47	10.402	0.624	9.621	0.689	10.140	0.655
XRF	Clay	23.82	37.458	0.827	5.97	**36.105**	**0.840**	24.27	37.347	0.829	**32.953**	**0.865**	**32.597**	**0.868**
	OM	13.3	4.619	0.371	3.63	*4.688	*0.352	12.23	4.613	0.373	*5.008	*0.258	*4.927	*0.282
	CEC	19.55	14.951	0.630	3.03	**14.780**	**0.638**	19.57	**14.901**	**0.633**	*15.504	*0.599	14.833	0.633
	pH	11.83	0.390	0.169	3.28	0.392	0.159	12.78	0.390	0.170	**0.364**	**0.252**	**0.369**	**0.238**
	V%	24.99	8.264	0.849	6.05	**7.751**	**0.868**	24.21	8.261	0.849	**6.576**	**0.906**	**6.297**	**0.913**
	P	11.54	13.955	-0.114	3.27	13.938	-0.119	11.45	13.900	-0.104	**12.984**	**-0.024**	13.996	-0.101
	K	16.41	1.353	0.650	4.2	**1.337**	**0.657**	16.55	1.352	0.649	**0.926**	**0.837**	**0.938**	**0.833**
	Ca	25.94	7.630	0.830	3.65	**7.495**	**0.835**	25.98	7.646	0.829	**6.888**	**0.861**	**6.907**	**0.860**
	Mg	25.44	7.088	0.652	3.76	**6.850**	**0.673**	25.31	7.091	0.651	**5.772**	**0.766**	**5.766**	**0.767**
	Average	20.16	10.219	0.622	4.20	9.925	0.628	20.11	8.914	0.623	9.249	0.668	9.079	0.674

#C: Number of components **Average**: Average of results without considering property *P*

Similar to the Ridge regression, the SVR also showed results that were statistically superior to the baseline for most of the analyzed attributes. This is even more highlighted for the SVR generated for the XRF data, especially for the *V%*, *K*, *Ca*, *Mg* properties. Still considering XRF, it is also observed that the SVR presents numerical superiority in relation to the Ridge for some of the target variables, such as the *CEC*, *OM* and *V%*. Therefore, in a scenario where these attributes are the most important, the SVR would be the more appropriate.

In Table 2 we also present the average of the measures used. The average of components was also calculated for the algorithms that relied on a reduction strategy. Considering the Vis-NIR, Ridge regression obtained the best overall result ($R^2 = 0.689$). On the other hand, for XRF, SVR achieved the highest effectiveness ($R^2 = 0.674$). Hence, in a real-world situation, combining these two models would yield the optimal outcome, considering the assessed models. For the algorithms relied on a reduction strategy, the baseline obtained the highest average ($R^2 = 0.636$), followed by PLSR ($R^2 = 0.632$) on Vis-NIR. In turn, for XRF, PLSR demonstrated the highest effectiveness ($R^2 = 0.628$), while the baseline achieved ($R^2 = 0.622$). For both cases, it is noted that the numerical difference between them is practically insignificant. However, PLSR achieved this effectiveness using a much smaller number of components than the baseline, especially for the XRF sensor, where the average number of components used was approximately 5 times lower than the baseline. Regarding that, further discussion is provided in Section 2.

4.2. Dimensionality reduction analysis

Besides the effectiveness of the models, Table 2 also indicates the number of components (#C) used by algorithms that relied on a reduction strategy (PLSR, Ridge PCA and Baseline). According to the results, the PLSR produced the smallest number of components

for both sensors. In addition, the PLSR was as effective as the baseline, with the advantage of using fewer components (less than half in many cases). Considering the XRF, it was statistically superior for *Clay*, *CEC*, *V%*, *K*, *Ca*, and *Mg* properties. Thus, in a scenario where effectiveness and efficiency are both important factors, PLSR would be the most recommended to be used, considering that it achieved competitive results with a highly reduced number of components. These outcomes suggest that the soil attributes are strongly correlated with some directions[2] in the data, especially for XRF.

Considering PLSR, its application on XRF raw data used fewer components when compared to Vis-NIR raw data, except for *P*. This behavior was different for Ridge-PCA and the baseline. Such findings differ from the ones in [Wei et al. 2022], where the application of PCR on raw XRF data used fewer components when compared to raw Vis-NIR data, except for *OM*, *pH* and *P*. The observed result could have been influenced by the validation sample used in the baseline work. In that study, the authors assessed the effectiveness of the approach using only one hold-out validation set. In contrast, our study employs validation across multiple (100 times) random test sets, providing a more strict evaluation of the model's effectiveness. This approach enhances the reliability and robustness of our findings.

Figure 2 illustrates the relationship between the number of components and the effectiveness of the baseline, PLSR, and the Ridge-PCA. Considering the dispersion, PLSR was able to obtain competitive results (y-axis), and using fewer components. This outcome is quite relevant since PCA is the common choice in many applications. In contrast, for determining soil properties, the experiments indicate that PLSR is more appropriate, considering that this model would need to process less input data to make the decision.

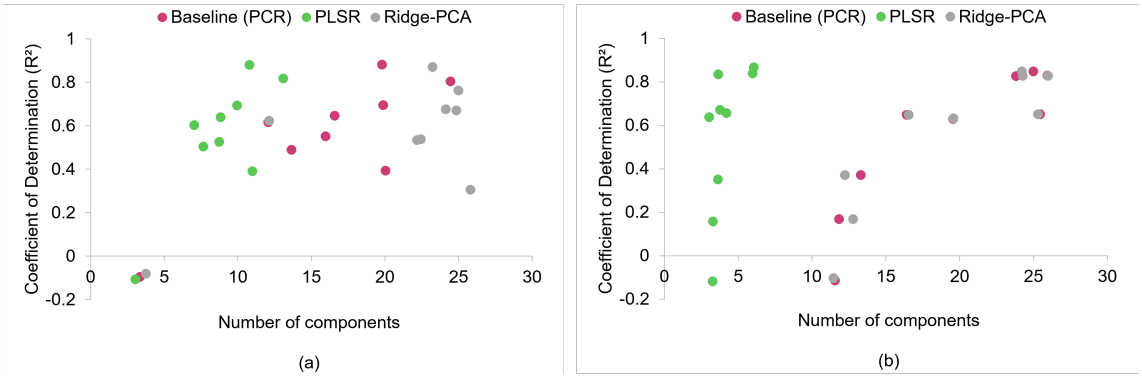

Figure 2. Scatter between the number of components and the effectiveness. In (a) for Vis-NIR sensor; in (b) for XRF

5. Conclusion

In this study, the potential of SVR and Ridge regression to determine physical and chemical soil attributes was evaluated. Furthermore, we compare the PCR, PLSR and Ridge-PCA dimensionality reduction approaches. The experimental results pointed to the Ridge regression as the most effective model, being statistically superior to the baseline for most of the analyzed properties. This indicates that this model was able to better capture the

[2]Directions represent the coefficients orientation of the linear combination. The direction is determined by a sign, which in turn, indicates the direction of association of the predictor with the composite variable.

existing relationships between the spectral data and the properties of the soil. It was also verified that the PLSR was mostly superior to the PCR in the dimensionality reduction process, considering that there was a greater reduction of the components, still maintaining a statistically superior or equal effectiveness to the baseline. Despite these findings, it is still necessary to carry out further investigations. Given the small number of samples from the dataset, future works should evaluate the strategies in this study on larger sets. Furthermore, new experiments should direct efforts to deal specifically with the prediction of property *P*, given that none of the evaluated algorithms was able to generate an appropriate model for this soil attribute.

Acknowledgement

This work was partially supported by UEFS AUXPPG and CAPES PROAP 2023 grants.

References

Benedet, L. et al. (2021). Rapid soil fertility prediction using x-ray fluorescence data and machine learning algorithms. *Catena*, 197:105003.

Folorunso, O. et al. (2023). Exploring machine learning models for soil nutrient properties prediction: A systematic review. *Big Data and Cognitive Computing*, 7(2).

Laili, A. R. et al. (2020). Prediction of soil macronutrient (nitrate and phosphorus) using near-infrared (NIR) spectroscopy and machine learning. *AIP Conference Proceedings*, 2203(1):020061.

Liu, C. et al. (2022). Partial least squares regression and principal component analysis: similarity and differences between two popular variable reduction approaches. *General Psychiatry*, 35(1).

M, S. and C D, J. (2021). Classification of soil fertility using machine learning-based classifier. In *2021 2nd ICSCCC*, pages 138–143.

Saleh, A. et al. (2019). *Theory of Ridge Regression Estimation with Applications*. Wiley Series in Probability and Statistics. Wiley.

Suchithra, M. and Pai, M. L. (2020). Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters. *IPA*, 7(1):72–82.

Sunori, S. K. et al. (2022). Design of ann based classifiers for soil fertility of uttarakhand. In *3rd INCET*, pages 1–5.

Tavares, T. R. et al. (2022). Spectral data of tropical soils using dry-chemistry techniques (vnir, xrf, and libs): A dataset for soil fertility prediction. *Data in Brief*, 41:108004.

Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *IWANN*, pages 758–770. Springer.

Wei, M. C. F. et al. (2022). Dimensionality reduction statistical models for soil attribute prediction based on raw spectral data. *AI*, 3(4):809–819.

Wollenhaupt, N. C. et al. (1994). Mapping soil test phosphorus and potassium for variable-rate fertilizer application. *JPA*, 7(4):441–448.

Yang, J. et al. (2020). Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using vis–nir spectroscopy. *Geoderma*, 380:114616.