Cost-effective spectral modelling of soil data

Master's thesis

Bartosz Smoczyński
University of Amsterdam
Student ID: 13898183
b.smoczynski@student.uva.nl

Dr. William Wu
Company Supervisor
QED.ai

Prof. dr. Alfons Hoekstra
University Supervisor
University of Amsterdam

ABSTRACT

Diffusive reflectance spectroscopy (DRS) is a technique enabling rapid and inexpensive measurements of key nutrients in soil. Adapting this method in developing countries can empower local farmers and facilitate soil condition monitoring. In this study we research the possibility of reducing costs of DRS-based soil surveys by focusing on careful selection of training samples and frequency subsetting. We find that data stratification yields a significant improvement in learning rates. Results on frequency subsetting give promise for utilizing a hand-held spectrometer for carbon and nitrogen measurements. Based on our results we recommend carrying out pilot surveys.

KEYWORDS

soil, MIR spectroscopy, spectral modelling, data stratification, frequency subsetting

1 INTRODUCTION

Diffusive reflectance spectroscopy or DRS is a technique which can non-intrusively determine the presence and concentrations of chemical substances in a sample. It measures the absorption of radiation, as a function of frequency or wavelength. It is successfully used in astronomy, chemistry, physics as well as the industrial sector. One of its applications is to examine the chemical composition and properties of *soil*.

Soil is an essential natural resource. All of agriculture relies on it and its well-being. It regulates the ecosystem and is the backbone of food and fiber production. Soil surveys are a way to monitor its condition. They examine soil's capacity to support essential ecosystem services such as primary productivity, nutrient and water retention and resistance to erosion.

Traditionally, in such studies, soil samples collected from the field are transported to specialized laboratories. Samples are dried, milled and sieved to later be analyzed using wet chemistry with accordance to strictly defined protocols.

QED.ai and its partners, under whom supervision this thesis is written, are mostly interested in conducting soil surveys in developing countries. The cost of analyzing a single sample using wet chemistry can reach up to 200$ [1]. It can take weeks or even months to obtain the results, mainly due to logistics, as there are very few ISO-certified labs in the global south. Continuous power supply cannot be taken for granted, which can also greatly impede the process of analyzing samples.

On the other hand, a DRS scan takes around 30 seconds and it's much easier to perform in the field setting. Moreover, QED is developing *scanspectrum* - a handheld visible/near-infrared spectrometer which will hopefully constitute an extremely cost-effective and portable solution.

Soil's chemical properties are inferred from the DRS scan with the use of machine learning. A large set of samples is collected into what's called a calibration set. Spectral data are paired with wet chemistry measurements and machine learning models are trained to predict the latter from the former.

Using high quality data, it is possible to obtain industry acceptable accuracy of predictions using this method. This result was repeatedly confirmed by many studies [2], [3], [4]. Spectral modeling is used to assess basic nutrient concentrations as well as the pH of soil and many other properties. This research gives great promise for building and deploying spectral models in developing countries.

Africa Soil Information Service (*AfSIS*) was a large-scale collaborative project aiming to fill a major gap in soil spatial information in Africa. The quality and diversity of data associated with the project proved to be sufficient to successfully train machine learning models for the task of soil properties prediction [5]. The study was ran is two stages. In the first stage 1831 samples were collected, and in the second - 781. After incorporating samples

from the second stage into the training set models' performance dropped. This highlights the importance of sample collection and selection procedures, as it is not clear why this happened.

A similar project - *IndiaSIS* was carried out in the Bihar state of India. The goal of the study was to obtain data that could drive the precise and profitable fertility management in the region. QED was involved in building spectral models for this survey. Unfortunately, in spite of many efforts, only the predictions for pH reached industry-acceptable accuracy. After a thorough investigation by QED the data itself was identified as the main culprit of this underperformance, as the modeling techniques were revised many times and successfully applied to other datasets.

Both these incidents highlight that the research gap is still existent in the spectral modeling field, especially when working in a new locality and with limited funding. There is much need for conducting cost-effective soil surveys, but little research has been put into how exactly these should be carried out. This thesis aims to, at least partially, fill this gap.

Research questions

How to facilitate low-cost spectral-based soil surveys by adjusting sample collection and modeling procedures?

Sub-questions.

- What are the necessary sampling frames for successful development of spectral models? How does incorporation of data stratification by mineralogy or location impact models' performance? How to effectively stratify by multiple features at once?
- How many samples are needed to build a model depending on the analyte? What are the learning curves? How are they affected by data stratification?
- What is the effect of frequency subsetting on the performance and learning rates? Can the $400nm$ - $1000nm$ range be used to model elemental carbon content?

2 RELATED WORK

Research in the field of spectral modeling of soil consists primarily of the evaluation and comparison of various methods in data preprocessing, outlier detection and model construction. Traditional machine learning models dominate the field with techniques such as partial least squares regression (PLSR) and Cubist - a type of CART tree with linear models in the leaves. There is very little to none published research on low-cost studies in new localities outside US.

In [6] Dangal et al. conduct a systematic study of how various machine learning techniques perform on the task of spectral modeling. The Kellogg Soil Survey Laboratory (KSSL) library is used for calibration and evaluation. The methods include PLSR, memory-based learning (MBL), random forest (RF) and cubist. Different preprocessing, outlier detection and data transformation approaches are considered. The authors point to the importance of estimating prediction uncertainty. They found that MBL produced significantly narrower prediction intervals compared with PLSR and RF methods. They highlight that their high performance ($R^2 \geq 0.98$) in predicting OC and CO3 is likely due to the use of a very large dataset sampled from the broad soil distribution of the US.

In [7] Ng et al. examine the capability of mid-infrared spectroscopy to predict 119 different soil properties. For their experiments they use a memory-based learning (MBL) model. MBL is a locally linear model, which fits a partial least squares regression (PLSR) based on its k-nearest neighbors. They use four metrics to evaluate prediction quality for different response variables. These metrics are then used to classify the properties into four groups using k-means. The division seems arbitrary but allows for rule-of-thumb classification of properties into categories for which the expected accuracy of prediction can be indicated. The data used in the study is taken from the KSSL-MIR library. It is a large library of high quality soil data sampled from across the US. The authors find that elemental concentrations exhibiting low interquartile range are harder to successfully predict.

In [3] Seybold et al. explore the transferability of calibration. Models calibrated and validated using the KSSL library are deployed in the field to check if they retain statistical accuracy and precision. Additionally the ability to transfer calibrations from different spectrometers. The field samples were collected in the US predominantly from locations where the soil type (taxon) is classified as Mollisols. The calibration dataset was stratified based on the soil horizon (layer) and other properties e.g. clay content. A separate PLSR model is developed for each strata. The authors develop an interesting approach in which they utilize a general model to

first predict the relevant strata so that a specialized model can be used for the final prediction. The general conclusion was that the transfer is viable. Models for predicting both total and organic carbon, which is of interest to this study, performed very well, with $R^2 \geq 0.97$. It is pointed out that the sample variability for prediction was greater than in similar studies conducted thus far.

3 METHODOLOGY

USDA KSSL mid-infrared spectral library

The majority of experiments in this study is performed using the USDA KSSL mid-infrared spectral library. The database contains more than 38,000 pedons, with measured chemical and physical properties representing geographically diverse soils from across the conterminous United States, Hawaii, and Alaska, see figure 1. It is maintained by the Kellogg Soil Survey Laboratory (KSSL). The data is available on request from the facility.

The dataset contains measurements of over 200 soil properties. We limit our attention to qualities essential for agricultural utility, in particular soil fertility. Table 1 summarizes the number of samples available for each considered property. We highlight that only 2500 samples exist that have available measurements for all considered features. This in particular means that we are forced to work with a varying dataset - depending on the analyte in question a different subset of samples is used. In our experiments we consider Olsen extractable phosphorus.

Property	no. samples
Phosphorus	14200
Potassium	29000
Nitrogen	68250
pH	50300
Carbon	74450
Clay	47350
Water retention	37850
GPS coordinates	46300
MIR spectrum	333750
All features	2500

Table 1: Number of samples for each soil taxon

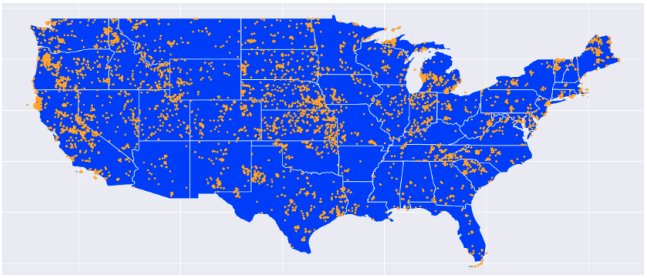

Figure 1: Locations of the soil pedons with MIR spectra available through the USDA KSSL soil database for samples from the conterminous United States

We continue by examining the joint distribution of all wet chemistry properties. Preceding the analysis we drop outliers falling outside the 99th quantile. Figure 2 depicts the distribution of each individual analyte, and figure 3 - the correlations between them. The distributions of elemental concentrations are heavily skewed.

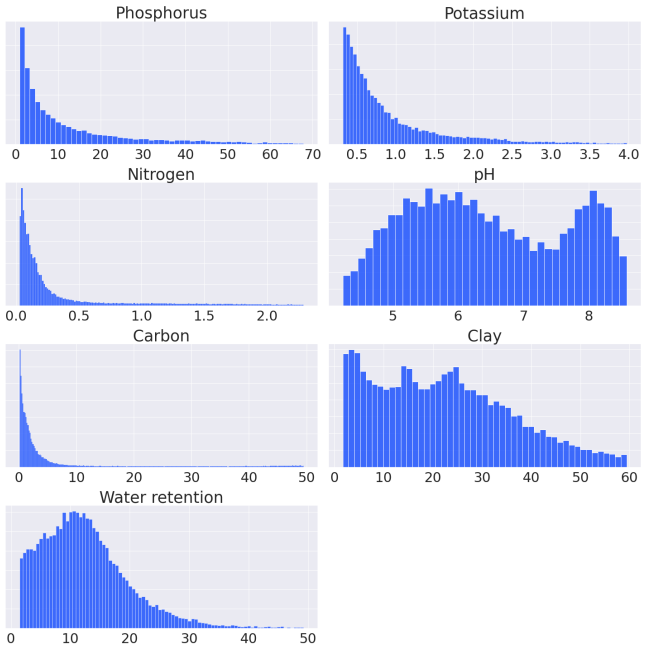

Figure 2: Distributions of wet-chemistry-measured features in the USDA KSSL database

IndiaSIS dataset

As we are most interested in remote low-cost soil surveys we explore a different spectral library - the IndiaSIS dataset. The samples were collected in Bihar state of India from 2018 to 2020. In total there are 3180 datapoints.

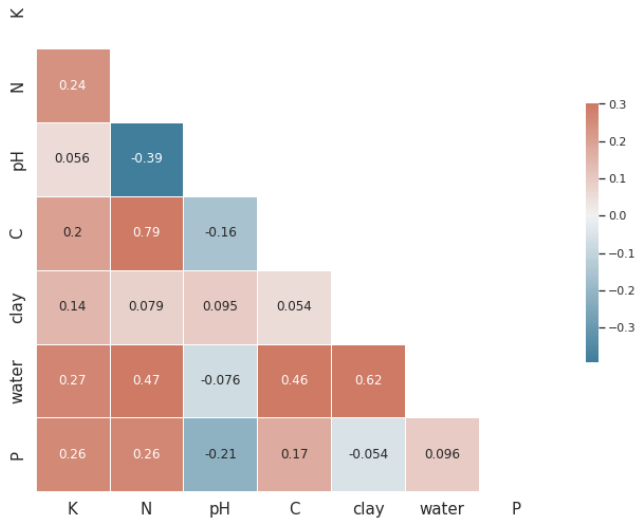

Figure 3: Correlations between wet-chemistry-measured properties in the USDA KSSL database

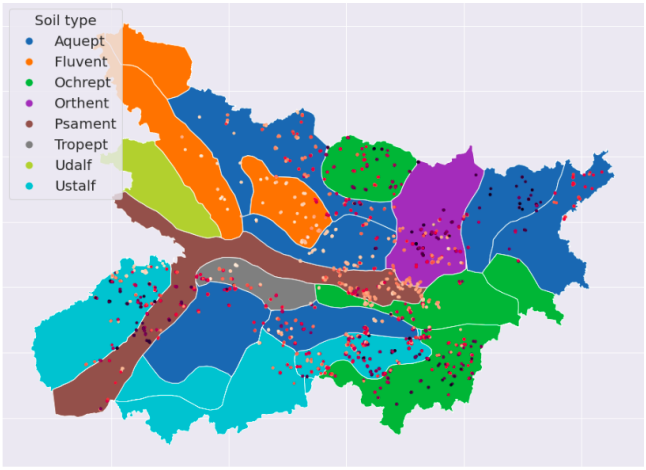

Figure 4: Locations of soil samples in the IndiaSIS dataset, along with the soil type of the region. Markers are shaded according to the pH.

We follow with the distribution analysis of wet chemistry measurements in the IndiaSIS database. Figures 11 and 12 in the appendix show the histograms and correlations of wet-chemistry-measured properties in the dataset. The histograms differ from the corresponding histograms of the KSSL data. Carbon and nitrogen distributions are far less skewed. The respective values for KSSL data are: $\mu_3(N_{KSSL}) = 2.3$, $\mu_3(C_{KSSL}) = 2.3$ whereas for IndiaSIS: $\mu_3(N_{IndiaSIS}) = 0.3$, $\mu_3(C_{IndiaSIS}) = 0.8$.

Directly from the plot we identify a data anomaly - the measurements for nitrogen only appear inside set intervals. Upon further analysis we discover that they fall into a grid with a regular spacing of 12.5 units. We are not aware what is the reason for this irregularity, but suspect it results from a particular method of measuring this element's concentration.

Spectra

We conclude this section by exploring the spectral data. Measurements come in the form of arrays of numbers - each representing the radiation absorbance for the corresponding wavenumber. We consider the mid-infrared spectra measured in the $500cm^{-1}$ - $4000cm^{-1}$ wavenumber range sampled every $2cm^{-1}$. This yields 1750 continuous features associated with each sample. Given that a typical low-cost soil surveys is able to collect around 2000 samples this means that we should be paying special attention to building models that are robust to overfitting.

To visualize the spectra, apart from two datasets introduced in the previous sections we use the AfSIS dataset. On figure 5 we provide a visual summary of the spectra from each of these databases.

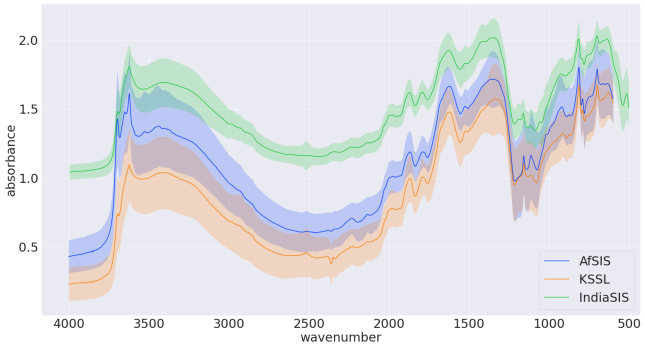

Figure 5: Median spectrograms for IndiaSIS, KSSL and AfSIS datasets, along with the standard deviation for each wavenumber.

Model development

Overall, our modelling approach is similar to that described in literature, with some differences regarding data preprocessing and chosen architecture.

Data preprocessing. The particular preprocessing method used in the final model is a result of trial and error

in choosing specific data transformations and their order.

To process the spectral data we first utilize Savitzky-Golay smoothing [8] using order three polynomials and window size of 50. This method works by fitting successive sub-sets of adjacent data points with a polynomial by the method of linear least squares. Smoothing makes the model partially invariant to local noise in the spectra and small horizontal shifts.

Subsequently, we apply the Δ operator of first order differences. First order differences can be more representative of peaks in the spectrum. At the same time they make to model completely invariant to vertical shifts of the input.

Finally, we employ baseline removal. The manner in which soil is placed into the microcup can vary the baseline in the MIR (and NIR) reflectance. Removing this baseline shift helps normalize the data and eliminate some dependence on sample placement. The algorithm iteratively performs a polynomial fitting in the data. At every iteration, the fitting weights in the regions with peaks are reduced to identify the baseline only.

Wet chemistry data doesn't require as much preprocessing as it is much less complex. We remove outliers falling outside the 1st and 99th percentiles. We normalize it to zero mean and unit variance. For some analytes the distributions are skewed, in which case we iteratively apply the logarithm function, until the skewness drops below 1. This is done since many regression models rely on the normality of the distribution and fit to the the data better in such setting.

Machine learning model. Machine learning architecture themselves are not a focus of this study, as most of them yield similar results. We thus, after experimenting with several of them, confine to using only one throughout the rest of this study.

Preliminary analysis showed that Extremely Randomized Trees regressor [9] gives the best predictions for most analytes. This model utilizes an ensemble of weak predictors to generate a single strong learner. Decision trees are used as the weak learners. Just like random forests, models of this class are resilient to overfitting owing to bootstrap aggregating. They are particularly useful for tasks with large numbers of covariates which justifies their use for spectral modeling. In addition they generally perform better than vanilla random forests. The Extremely Randomized Trees regressor differs from other gradient boosting techniques by strongly randomizing both attribute and cut-point choice when splitting a tree node. In the extreme case, it builds totally randomized trees whose structures are independent of the output values of the learning sample.

Evaluation. The primary metric of model's performance is the coefficient of determination R^2. It represents the ratio of the variance explained by the model to the total variance of the target values. The exact formula is given below.

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2}$$

In most settings models reaching $R^2 = 0.7$ or above are considered accurate enough for industry applications.

4 EXPERIMENTAL SETUP

To establish a baseline and validate our modelling approach, we start by replicating state-of-the-art results for the task of wet chemistry predictions on the KSSL database. Once this is accomplished, we move on to experimenting with smaller sample sizes, hoping to imitate the low-cost field setting.

Experiments on data stratification

Stratified sampling is a method of sampling that ensures a balanced distribution of a chosen feature among the samples. In our experiments we aim to determine the impact of stratification on model's performance. We selected features that can be estimated in the field without the use of any expensive equipment, namely clay content, water retention, spacial coordinates and pH. Based on this estimate it can then be decided whether to proceed with further analysis of a sample. This can help reduce the cost and prevent unnecessary lab work.

The impact of stratification is most pronounced when size of the calibration set is relatively small. Thus, for the experiments on stratification we consider sample sizes of 100, 200, 400 and 1000. To reduce the variance of the measurement, in each case 10 or more subsets drawn from the entire database were used with a 5-fold cross validation for each of them. We define three ways of stratifying the data. For a simplified visualization see figure 6.

Random. A control trial where we don't employ any kind of stratification. This constitutes a baseline to compare against.

Balanced. We construct even sized strata based on the quantiles of the feature. Then we form the splits in the cross validation that have equal proportions of samples from each strata.

Adverserial. This way of sampling is meant to simulate the worst possible stratification. Splits in the cross validation each come from a different strata.

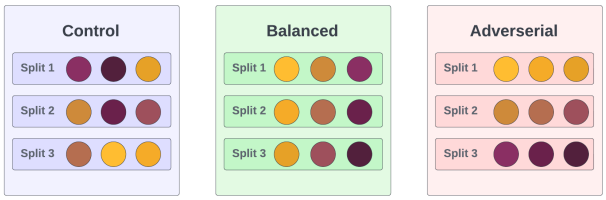

Figure 6: Illustration of different stratification strategies. Color intensity represents the magnitude of a feature by which stratification happens

Multi-feature stratification

In some cases it is desirable to stratify by several features at once. An obvious example is that of stratifying by longitude and latitude of the sampling site. We can also imagine taking independent measurements of clay content and pH and stratifying by both of these properties. We develop several techniques that aim to address this problem. Figure 3 illustrates an overview of these methods.

Simple Cartesian stratification. Quantiles are independently calculated for each feature and the sample space is cut into Cartesian products of intervals between subsequent quantiles. Samples falling into each of so formed multidimensional boxes are taken to be the strata.

Selective Cartesian stratification. This method works similarly to the previous one, except an equal number of samples is selected from each box and the rest are discarded. This flattens out the distribution and yields equal sized strata.

Recursive stratification. This method is defined recursively as follows. In the case of a single feature it reduces to one dimensional balanced stratification described previously. In case of multiple features we compute the quantiles of the first feature. We divide the data points into groups based on these quantiles and proceed recursively with each group. The groups obtained in the last step form the strata. Like in the previous method each strata has the same number of samples.

K-means stratification. We cluster the data points in the space of the features that we are stratifying by using the k-means algorithm. Each cluster forms a stratum. The drawback of this method is that it is not clear how to apply it in practice when collecting samples.

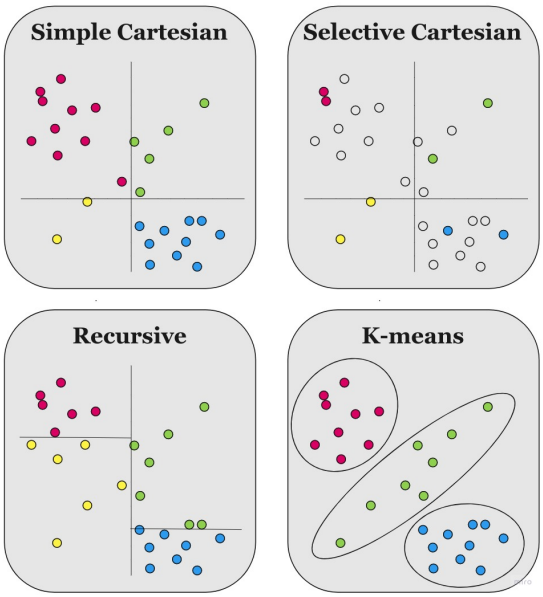

Figure 7: Illustration of multidimensional stratification methods. Each color designates a different stratum

We run experiments to compare the effectiveness of each method.

Learning curves

We explore how many samples are needed to calibrate a model for particular nutrients. We again use 5-fold cross validation on at least 10 different subsets of the database in order to narrow the confidence intervals. To quantify the impact of decreasing the number of samples in the calibration set, whilst not affecting the validation set, we subset only the splits selected for training in the cross validation. We evaluate the impact of data stratification on learning curves by applying the methods described earlier and comparing against random stratification.

Frequency subsetting

As mentioned in the introduction QED is developing scanspectrum - a low-cost hand-held spectrometer operating in the visual-infrared range. Using VNIR spectra for nutrient prediction has been researched in other studies [10], [11]. However these studies use a much wider frequency range of 350-2500nm obtained using expensive equipment. Scanspectrum provides readings in the range of 400-1000nm. We therefore truncate the spectra in the database to this range and examine their ability to model elemental carbon content.

Additionally QED recommended that another cost effective solution can rely on using specially designed crystals that react only to specific frequencies in the MIR range. Therefore we evaluate the effect of subsetting the frequencies used for prediction. In order to select which features to retain, we first calibrate the model on a 5000 sample training set. Then, we use the impurity-based feature importance criterion to determine a small subset of origin features. Finally we include wavenumbers within $2cm^{-1}$ from those selected in the previous step to compensate for small horizontal misalignments.

5 RESULTS AND ANALYSIS

Baseline results

We report performance metrics for three spectral models:

- The spectral model described by Ng et al. in [7] calibrated and evaluated on the KSSL spectral library.
- Our own model also evaluated on the KSSL spectral library.
- A model developed using the same approach but trained and evaluated on the IndiaSIS dataset.

Table 2: Baseline R^2 scores on KSSL and IndiaSIS datasets compared to Ng et al.

analyte	Ng et al.	KSSL	IndiaSIS
P	.37	.52	.16
K	.48	.59	.22
N	.85	.95	.15
pH	.85	.87	.69
C	.95	.99	.26

Our model achieved a higher performance then the Ng et al. model for all analytes. This proves our modeling approach effective. At the same time, it performs poorly on the IndiaSIS dataset, which confirms previous findings by QED.

Data Stratification

Remark. In this and following parts we use a varying number of samples depending on the experiment. This was done since learning rates are different for different analytes. In general, we try to use as few samples as possible in order to get the best possible translation into a low-cost survey.

Figures 8 and 9 show the impact of stratification on model's performance on predicting carbon and potassium. Refer to the appendix for similar plots on nitrogen and phosphorus. We abbreviate water retention at 15 bar to 'water', latitude and longitude of the soil pedon to 'LAT' and 'LONG', and finally clay content to 'clay'. For simplicity we only consider several possible criteria to stratify by. In each case for every individual feature only two strata are created. Preliminary analysis showed that increasing the number of strata does not yield any further significant improvements to the performance. As an example we noted the following scores for carbon prediction when increasing the number of clay strata: no stratification: .86, 2 strata: .892, 3 strata: .894, 5 strata: .895, 10 strata: .897. From this point onward, unless otherwise stated we use two-fold stratification per feature. In case of multi-feature stratification we use the simple Cartesian method described above. We emphasize again that if we were using a fixed set of samples, all scores obtained using random stratification would be equal for a given analyte. However this is not the case as we are using a varying dataset as mentioned before.

As expected, in all cases adverserial stratification yields the worst evaluation, whereas balanced stratification - the best. Using a twofold stratification by clay we were able to outperform Ng et al. phosphorus model using only 1000 samples (R^2 = .39 vs R^2 = .37), compared to over 14000 samples that are available. In some cases the score obtained using adverserial stratification drops to 0, which could equivalently be achieved by model which outputs a constant value. We discuss the implications of this in the next section.

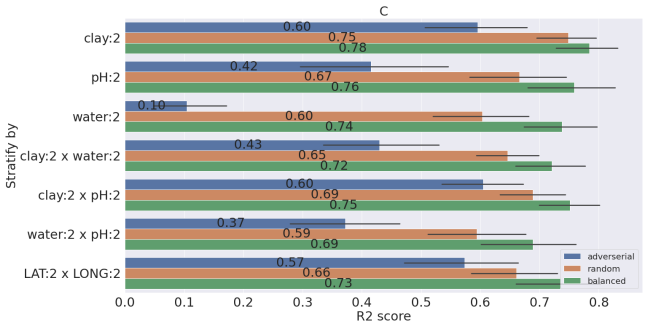

Figure 8: Impact of data stratifications on R^2 scores for predicting carbon using 100 samples. .95 confidence intervals are indicated by black horizontal lines.

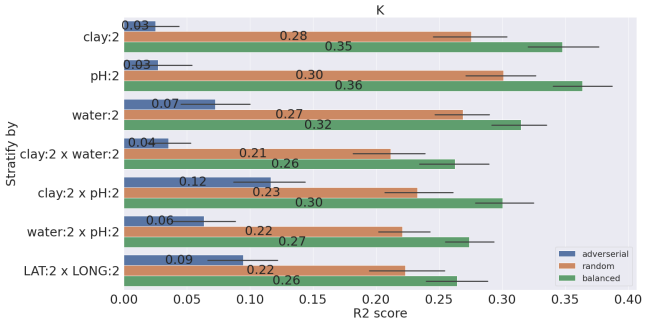

Figure 9: Impact of data stratifications on predicting potassium using 1000 samples.

Multi-feature stratification

We follow with the comparison of different multi-feature stratification methods proposed in the previous section. In table 3 we report metrics for carbon using 200 samples, nitrogen using 400 samples, phosphorus and potassium using 800 samples. For every property we considered the same 4 combinations of features to stratify by.

In most cases selective Cartesian stratification yielded the best results, whereas random stratification - the worst. Overall, regardless of the specific method used, stratification improves performance.

Learning curves

Figure 10 plots the R^2 score as a function of the number of samples using standard, and clay-stratified sampling. Table 4 gives further evaluations for other properties.

Table 4 reveals a clear pattern: stratification has the most impact on performance when using a small training set. In general, it improves the learning rates of the models.

Table 3: Multi-feature stratification methods R^2 scores comparison.

prop.	strat.	Random	Cartesian	Selective	Recursive	Cluster
N	clay, water	.62	.7	**.74**	.67	.69
	clay, pH	.76	.79	**.84**	.8	.8
	pH, water	.78	.82	**.85**	.84	.84
	lat, long	.84	.87	**.88**	.86	.86
C	clay, water	.76	.83	**.86**	.83	.84
	clay, pH	.79	.82	**.87**	.8	.83
	pH, water	.66	.78	**.84**	.6	.78
	lat, long	.75	.86	**.94**	.85	.88
P	clay, water	.21	.25	**.26**	.25	**.26**
	clay, pH	.32	**.37**	.35	.36	**.37**
	pH, water	.18	**.24**	**.24**	**.24**	**.24**
	lat, long	.11	.15	**.17**	.14	.16
K	clay, water	.17	.23	**.29**	.23	.24
	clay, pH	.21	.29	**.34**	.29	.29
	pH, water	.12	.19	**.27**	.2	.2
	lat, long	.19	.25	**.28**	.25	.25

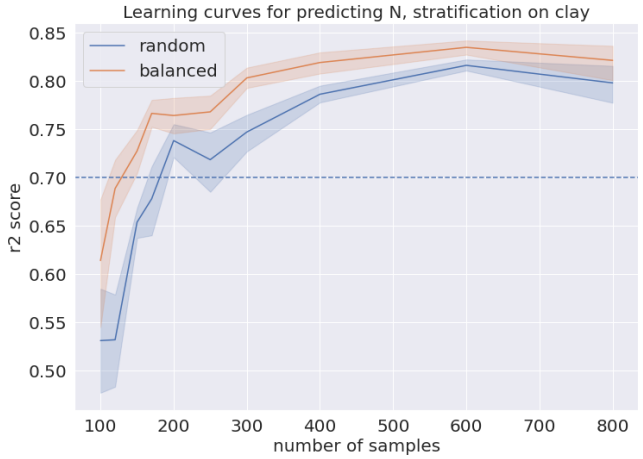

Figure 10: Learning curves comparison. Shaded regions denote the .95 confidence intervals

VNIR spectra

In table 5 we report evaluation metrics for predicting carbon from the VNIR spectrum. Both the full and truncated spectra were used for comparison.

Table 4: R^2 scores for different calibration set sizes. For each analyte we use random sampling (denoted as rnd.) and clay-stratified sampling (denoted as st.).

	Analyte							
	N		P		K		C	
n	st.	rnd.	st.	rnd.	st.	rd.	st.	rd.
100	.7	.62	.25	.13	.1	.09	.76	.72
200	.8	.71	.24	.15	.13	.11	.86	.82
300	.83	.79	.26	.21	.18	.14	.88	.85
500	.84	.82	.32	.25	.24	.17	.91	.89
1000	.87	.85	.4	.34	.35	.27	.93	.93
2000	.89	.88	.45	.41	.43	.4	.95	.94

Table 5: R^2 scores obtained by carbon models using VNIR spectrum

n	truncated 400-1000nm	full spectrum 350-2500nm
100	.12	.34
200	.44	.63
300	.6	.75
500	.68	.8
1000	.73	.86
2000	.75	.88

Using 300 samples the full spectrum model was able to cross the .7 R^2 threshold, whereas a 1000 samples were needed to achieve this using the truncated spectrum. Nonetheless, our results confirm that using the 400-1000nm range is enough to obtain an acceptable accuracy for elemental carbon predictions.

Frequency subsetting

Tables 6 and 7 show the effect of using a subset of frequencies for prediction. Be mindful that the actual number of features used for prediction is higher by about 20% to 30% than the number of origin features due to the incorporation of adjacent wavenumbers.

As expected, learning occurs faster when using a small subset of features. Incorporation of additional wavenumbers becomes more beneficial with growing number of samples used for training.

Table 6: N

	No. origin features			
	30	50	100	1700 (all)
n				
50	.67	**.74**	.7	.24
100	.71	.78	**.79**	.66
400	.86	.88	**.89**	.86
1000	.89	.9	**.91**	.9
3000	.9	.92	**.93**	**.93**

Table 7: C

	No. origin features			
	30	50	100	1700 (all)
n				
50	**.45**	0	0	0
100	.87	**.9**	.88	.84
400	.95	**.96**	**.96**	**.96**
1000	.96	.97	**.98**	**.98**
3000	.97	.97	**.98**	**.98**

6 DISCUSSION

Although we managed to outperform the Ng et al. model, building another spectral model for the KSSL library was not the main objective of this study. There exist spectral models optimized to predict only particular elements that often report higher accuracy metrics [12]. Evaluation metrics obtained for the IndiaSIS dataset hint that there exists an underlying problem with the data itself.

We demonstrated that using basic stratification techniques can significantly improve the performance and learning rates. It is important to assess if these techniques can realistically be applied in the field. QED's soil scientist has confirmed that clay content can be effectively estimated using hand measurements. Since our method distinguishes only two strata of clay content, this way of its estimation is sufficient. Similarly, pH meters are widely accessible and portable which makes them easy to use in the field. The only property that poses a limitation is that of water retention which requires the use of more specialized equipment.

Very poor scores obtained using adverserial stratification are an indication that using a spectral

model in a location with mineralogy deviating from that of the area where training samples were collected is doomed to failure.

We carried out an extensive analysis of multi-feature stratification techniques in order to demonstrate the subtle differences between them. We weren't able to find any other sources on how to approach this problem. Out results suggest tthat the Selective Cartesian stratification is the most effective one. This is fortunate since it can also be easily incorporated into the sample collection process, unlike for instance the Cluster stratification method. We failed to perform experiments using even more stratification features. Moreover, due to the varying dataset used throughout the experiments our results don't give a clear recommendation as to what features are optimal for stratification for a given analyte. Nonetheless, based on the results we obtained we suspect that using a combination of spacial coordinates, clay content and pH is a promising combination.

Experiments on learning curves revealed that using as little as 100 samples can be enough to calibrate a spectral model for carbon and nitrogen prediction on a industry acceptable level. This motivates the idea of running pilot studies in new localities. Such survey could collect a small number of samples and assess the chances of success of a full scale survey. Such a procedure could have prevented futile spectral measurements performed on IndiaSIS samples.

We were able to successfully calibrate a model for carbon prediction using the truncated 400-1000nm VNIR range. A considerable limitation of this result is that the spectral measurements come from a high quality specialized spectrometer. A hand-held device does not yield such accurate and noise-free measurements. More research is needed before we can confirm the viability of ScanSpectrum for carbon prediction.

We demonstrated that it is possible to achieve a high performance of the models using only a small subset of the frequencies. This however is only a preliminary result, since we based the choice of the frequencies to retain on the models. The question of the possibility of using specialized crystals as a substitute for a traditional spectrometer remains open. Nonetheless, we discovered yet another way of accelerating learning. As our results show, subsetting the frequencies enables us to obtain a reasonable accuracy using as few as 50 samples. This can further aid with the execution of pilot surveys. This idea is however limited by the transferability of characteristic frequencies for an analyte between different spectrometers, as we used a 5000 sample subset to determine them.

Finally, further research is required that would explore combining the presented methods of stratification and frequency subsetting. In our analysis on the impact of stratification we used only the simple Cartesian method. However, as subsequent results show, the Selective scheme usually yields better performance. Additionally, as mentioned before, incorporating more features to stratify by could also boost the efficiency. Bringing together all these techniques has a potential for reducing sample size required for calibration even further.

7 CONCLUSION

Careful sample selection procedure can significantly improve the performance and learning rates. We recommend stratifying collected samples by clay, pH and spacial coordinates. For multiple feature stratification the Selective Cartesian scheme should be employed. In order to build a model reaching acceptable performance for carbon and nitrogen prediction as few as 100 samples are needed, given that proper data stratification is used, since it accelerates learning. Subsetting the frequencies can further decrease the minimal calibration set size. The 400-1000nm VNIR range was found to be sufficient for elemental carbon modelling using 1000 samples.

In order to facilitate spectral-based soil surveys we recommend executing pilot studies in new localities. Such studies could help validate laboratory procedures, spectral equipment and soil's susceptibility to spectral modelling in an early stage. We prescribe using appropriate data stratification and frequency subsetting procedures in pilot studies as they have a major impact on the performance with such small sample sizes.

REFERENCES

[1] D. J. Brown, K. D. Shepherd, M. G. Walsh, M. Dewayne Mays, and T. G. Reinsch, "Global soil characterization with vnir diffuse reflectance spectroscopy," *Geoderma*, vol. 132, no. 3, pp. 273–290, 2006.

[2] J. M. Soriano-Disla, L. J. Janik, R. A. V. Rossel, L. M. Macdonald, and M. J. McLaughlin, "The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties," *Applied Spectroscopy Reviews*, vol. 49, no. 2, pp. 139–186, 2014.

[3] C. A. Seybold, R. Ferguson, D. Wysocki, S. Bailey, J. Anderson, B. Nester, P. Schoeneberger, S. Wills, Z. Libohova,

D. Hoover, and P. Thomas, "Application of mid-infrared spectroscopy in soil survey," *Soil Science Society of America Journal*, vol. 83, no. 6, pp. 1746–1759, 2019.

[4] W. Ng, B. Minasny, S. H. Jeon, and A. McBratney, "Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions," *Soil Security*, vol. 6, p. 100043, 2022.

[5] W. Wang, L. Hu, B. K. Wilson, and M. D. Keller, "Potential for predicting soil properties using low cost near-infrared spectroscopy and machine learning," in *Optical Sensors and Sensing Congress*, p. AW4I.2, Optica Publishing Group, 2020.

[6] S. R. S. Dangal, J. Sanderman, S. Wills, and L. Ramirez-Lopez, "Accurate and precise prediction of soil properties from a large mid-infrared spectral library," *Soil Systems*, vol. 3, no. 1, 2019.

[7] W. Ng, B. Minasny, S. H. Jeon, and A. McBratney, "Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions," *Soil Security*, vol. 6, p. 100043, 2022.

[8] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[9] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, p. 3–42, apr 2006.

[10] N. Knox, S. Grunwald, M. McDowell, G. Bruland, D. Myers, and W. Harris, "Modelling soil carbon fractions with visible near-infrared (vnir) and mid-infrared (mir) spectroscopy," *Geoderma*, vol. 239-240, pp. 229–239, 2015.

[11] A. Gholizadeh, C. Neumann, S. Chabrillat, B. van Wesemael, F. Castaldi, L. Borůvka, J. Sanderman, A. Klement, and C. Hohmann, "Soil organic carbon estimation using vnir–swir spectroscopy: The effect of multiple sensors and scanning conditions," *Soil and Tillage Research*, vol. 211, p. 105017, 2021.

[12] J. Sanderman, K. Savage, and S. R. Dangal, "Mid-infrared spectroscopy for prediction of soil health indicators in the united states," *Soil Science Society of America Journal*, vol. 84, no. 1, pp. 251–261, 2020.

A ADDITIONAL FIGURES

This part includes additional figures that couldn't fit in the main text due to space limitations.

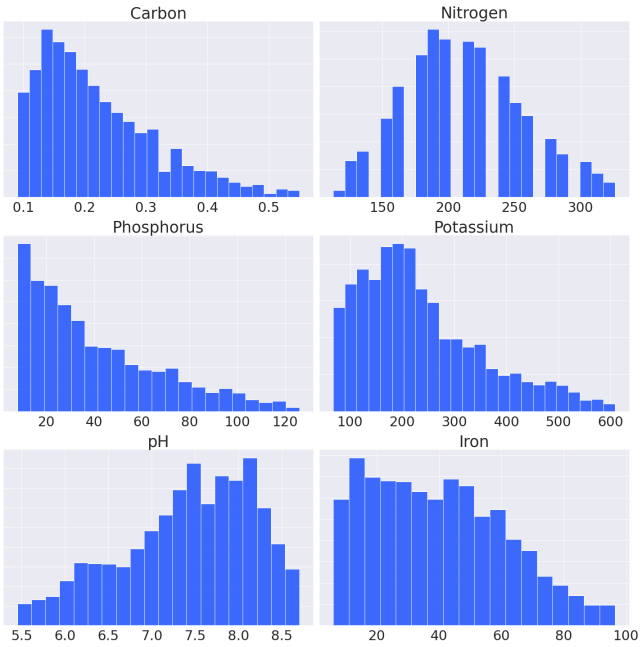

Figure 11: Distributions of wet-chemistry-measured features in the IndiaSIS dataset

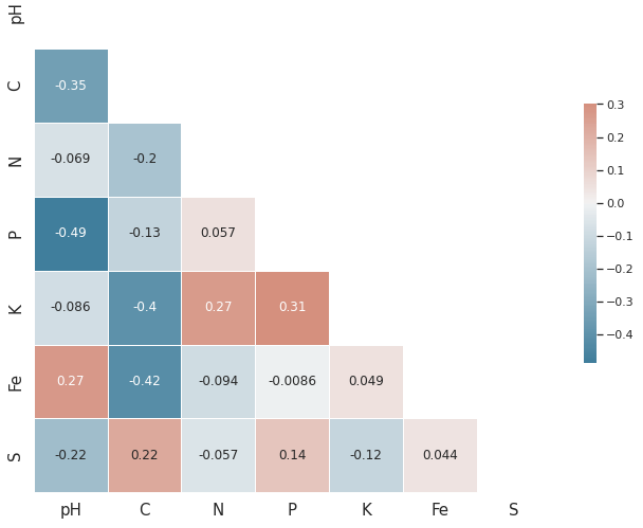

Figure 12: Correlations between wet-chemistry-measured properties in the IndiaSIS dataset

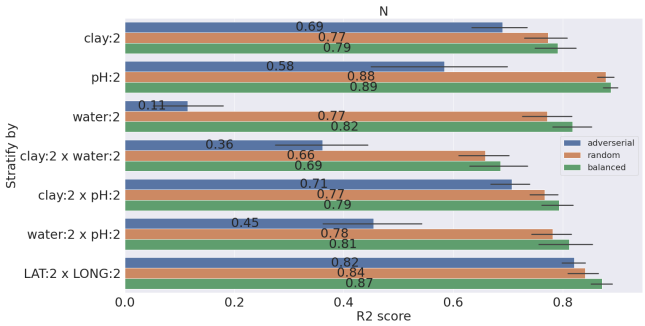

Figure 13: Impact of data stratifications on predicting nitrogen using 400 samples.

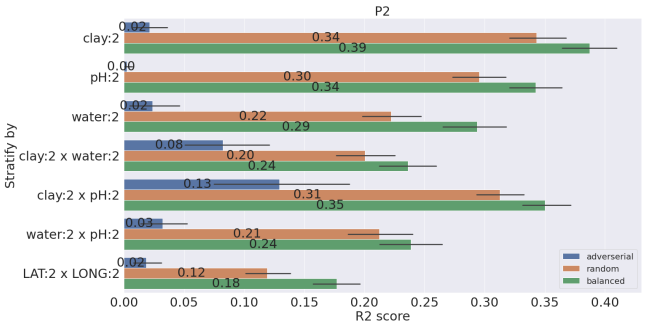

Figure 14: Impact of data stratifications on predicting phosphorus using 1000 samples.