The Alan Turing Institute

University of BRISTOL

Data Study Group Network Final Report: Rothamsted Research

5–9 August 2019

Tackling hidden hunger through soils

https://doi.org/10.5281/zenodo.3775489

This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, 'AI for Science and Government' programme at The Alan Turing Institute

DSG17: Tackling hidden hunger through soils

Samuel Ellick, Timo Breure, Chris U. Carmona,
Andrew Dowsey, Ben Evans, Ali Fahmi,
Stephan Haefele, Kirsty L. Hassall, Markus Loning,
Diego Perez Ruiz, Darya Shchepanovska and Cathy Thomas

November 2019

Executive summary

This report describes the work completed during a week long data study group hosted by the Alan Turing Institute. The challenge was provided by Rothamsted Research and looks at predicting soil and plant physicochemical properties from soil infrared (IR) spectra. Three datasets were explored and modelled using a combination of established and more recent data-science strategies. Due to the size, scope and variety in the datasets, multiple conclusions were drawn. Overall, our preliminary findings indicate that soil physiochemical properties were easier to model than plant physicochemical properties. Decision tree based methods were used consistently throughout the three datasets and were overall more robust than other approaches considered in our analysis. Our results are in line with the current literature; IR data can be an effective predictor of the physicochemical properties of soil and by extension, the health of the soil.

1 Introduction

Hidden hunger refers to a pandemic issue effecting large portions of the globe including sub-Saharan Africa and describes macro/micro nutrient deficiencies within humans. While the source of famine is easily identified as a lack of food or calorific intake, hidden hunger is more nuanced. A population may be getting enough nutrition in terms of calorific intake, however the source of this nutrition may be depleted in certain nutrients such as calcium, potassium and other nutrients essential for normal human function. Over a period of time the lack of these nutrients within their diet may present itself as a systemic medical condition, and indirectly affect the productivity and economic security of their families and countries.

Multiple organisations and research institutes around the globe conduct research to combat the hidden hunger problem. The African soil information service (AfSIS) is one such organisation, focusing on hidden hunger in sub-Saharan Africa. A large amount of chemical and geological data is curated by

AfSIS. With this information, AfSIS aims to provide strategies that improve the nutritional composition of the crops grown in sub-Saharan Africa and improve the quality of life for millions of people.

1.1 Challenge overview

This report presents the output of a week-long collaboration between the Alan Turing Institute and Rothamsted Research to scope automated modelling techniques for determining soil quality in East Africa from mid-range infra-red spectra (MIR) data.

The current gold standard method for determining the chemical and physical composition of soil is through wet chemical analyses; using such information farmers could better implement strategies specific to their soil to improve crop health and nutritional value. The drawback to wet chemical analyses is that it is expensive and requires specialist equipment and personnel within controlled laboratory conditions. Previous research has shown mid-range IR spectroscopy (MIR) of soil to be a promising technique to predict macroscopic soil properties such as moisture content, particle size distribution and exchangeable nutrients Na/Ca and K. MIR is an attractive method for predicting soil properties since it is much cheaper and more portable than wet-lab investigations and hence would be easier to deploy over sub-Saharan Africa. In this work, we aimed to develop similar approaches to predict micronutrient content within soils from IR spectra using the data provided by Rothamsted Research.

1.2 Data Overview

Rothamsted Research provided three data sets allowing for the exploration of various approaches to tackle this problem. All datasets contained infrared spectroscopy of the soil paired with wet chemical analyses. Differences between the datasets were: the size of the sample-set, controlled vs not controlled experimental conditions and inclusion of nutrient content of the plant biomass.

- **Dataset 1:** Controlled experiment. Six hundred samples of crops grown on 30 different soils, with 7 different fertiliser applications. Recorded data are wet chemistry and MIR spectra data on soil, data on experimental design and elemental accumulation within the crops.
- **Dataset 2:** Real word data taken from a single country, Ethiopia. Paired soil and crop samples (n=456). Recorded data are wet chemistry and MIR spectra data on soil and elemental accumulation within the crops.
- **Dataset 3:** Real word data from multiple sub-Saharan countries. Paired soil and crop samples (n=2000). Recorded data are wet chemistry and MIR spectra data on soil only.

1.3 Main objective

From our project introduction and initial discussion, three main objectives were identified:

1. Predict crop nutrient content from soil IR spectra data
2. Predict soil elemental composition from soil IR spectra data
3. Predict crop nutrient content from soil elemental composition

Being able to directly predict crop nutrient content in plant biomass from soil IR spectra is the most attractive target as it addresses the core problem (will these crops provide adequate nutrition?) using the most easily deployed technique. The second objective, predicting soil composition from infrared spectra was identified as another target and in theory should be easier as the soil composition and IR spectra will be directly correlated to each other. The third objective, predicting crop nutrition from soil composition, was identified as a third potential target and could allow for predicting crop nutrient content from the predictions of soil elemental composition from IR spectra.

1.4 Approach

Two spectral pre-processing strategies and a variety of machine learning algorithms were explored to achieve our main objectives. Normalisation and derivitisation were the pre-processing strategies applied to the spectral data. Machine learning algorithms used included regularised linear models, support vector mechanics and tree-based models. A breakdown of the approaches used and their effectiveness are given throughout this report.

1.5 Limitations

Due to time limitations and the small numbers of samples, we did not obtain confidence intervals nor assess whether differences in predictive performances of our tried out methods were statistically significant. Measurements of samples in datasets 2 and 3 are considered i.i.d. samples ignoring spatial correlations. Models built from this data may not generalise well to new samples.

Dataset 1 was recorded under controlled conditions while datasets 2 and 3 were not. For datasets 2 and 3 there was no information on farming practices, i.e. whether fertiliser was applied to the soil and how the land was cultivated. Inconsistent practices between different farms and regions may impact model predictive performance, particularly for the prediction of crop nutrient content.

1.6 Report

This report goes into detail about the work done during the data study group. A breakdown of the three datasets, as well as common pre-processing strategies, are discussed in section 2. A breakdown of the approaches applied to datasets

1, 2 and 3 are given in sections 3, 4 and 5 respectively; these sections may be further broken down where appropriate. Finally, our closing remarks and ideas for future work are given in section 6

2 Experimental

2.1 Dataset

The data provided by Rothamsted Research can be divided into three categories: the spectral data, wet chemistry data and meta-data. The standard operating procedure for data generation and collection is not referenced in this report; however, it was provided by the data challenge providers and is available in the literature on the wider AfSIS project. [Heng, T 2014]

The wet chemistry data was the most diverse data type within this challenge. Both physical and chemical measurements were provided, however, the work in this report focuses on predicting the chemical measurements such as the elemental composition. The elemental composition was available for both the soil and the crop biomass dependant on the dataset. The soil elemental composition is split into two categories: total elemental composition and available elemental composition. The total elemental composition describes the raw elemental composition within the soil *i.e.* independent of whether that element is organic or inorganic, while, the available elemental composition aims to simulate the elemental nutrients available to the plant after rainfall. For the available elemental composition, the soil undergoes an extraction process to wash any organic elemental components off the soil, the elemental concentration of this eluent wash is then measured. Both the total and available elemental concentration was measured by inductively coupled plasma emission spectroscopy (ICP-OES) while the crop biomass elemental composition was determined by x-ray fluorescence (XRF).

The spectral data comprises of the infrared absorbance of the soil; for datasets 1 and 2 the mid-infrared region was collected (500 - 4000 cm^{-1}) and for dataset 3 both the mid and near-infrared region was collected (500 - 12000 cm^{-1}). The instruments used to collect this data are not known nor whether this was kept consistent between the datasets. The metadata comprises of geographical locations of the samples for data set 3.

2.1.1 Dataset description

Dataset 1: Dataset 1 was a controlled experiment looking at the relationship between soil fertilisation and the development of a single crop, wheat. The overall experiment was a randomised split-plot design and done under greenhouse conditions. 30 Kenyan soils and one standard control compost were used. The soil had one of 7 possible fertiliser treatments: Full fertiliser, control (no fertiliser) and 5 full fertilisers each missing an elemental component (K, N, P, S, Zn). For each soil 20 repeats (pots) were produced; 5 repeats were used of the full fertiliser and the control (no fertiliser), 2 repeats were made for each of

the fertilisers missing an elemental component. The sample-set size was 620 (31 soils, 20 pots).

The analytical data provided for dataset 1 is:

- Infrared spectrum of each of the soils
- Elemental and physicochemical properties of the soils
- Elemental concentration of the plants' biomass, sampled at 3 different stages (grain, harvest 1 and harvest 2)

Dataset 2: Dataset 2 was not a controlled experiment and represented real-world analytical data. 456 samples site across Ethiopia were used; for each a soil sample was taken for elemental composition and spectral analysis, additionally some of the crop grown was taken to measure its nutritional value. A variety of crops were analysed in this dataset, most commonly wheat and teff.

The analytical data provided for dataset 2 is:

- Mid-infrared spectra of the soil
- Elemental and physicochemical properties of the soil
- Elemental concentration of the crop biomass at harvest

Dataset 3: Dataset 3 is similar to dataset 2 as while dataset 2 was collected from Ethiopia only, dataset 3 was collected from 20 different countries in sub-Saharan Africa. The sample size of dataset 3 is larger than dataset 2, n = 2002 versus n=456 respectively. Additionally, dataset 3 does not contain any data on crop biomass nutritional content, only paired soil wet chemical and spectral analyses were collected.

The analytical data provided for dataset 2 is:

- Mid and near-infrared spectra of the soil
- Elemental and physicochemical properties of the soil
- Location of the soil sample including country and geographical co-ordinates.

2.1.2 Data quality issues

The dataset was generally well-curated and clearly documented with a few minor issues: Although not a data quality the resolution of the infrared spectra was inconsistent across the datasets. While all datasets contained the mid-infrared range of numbers (500 - 4000 cm^{-1}) the length and distance between measurements were different; one data set may have had 1000 evenly spaced values between 500 and 4000 cm^{-1} while another data set may have had 1800 evenly spaced values. This made transferring models from one dataset to another difficult without limiting models to wave numbers used in all three datasets or to implement some form of interpolation. There was also some inconsistent variable names for the same measurement across the three datasets.

2.2 Pre-processing

The pre-processing strategies applied to spectroscopic data effect the correlations observed in the data. Pre-processing is required for spectroscopic data as samples may have slightly different base-lines between runs, particularly for dark samples such as soil. The two major pre-processing strategies outlined in this work are per-sample z-score standardisation and first-order derivitisation. Methods such as baseline subtraction and smoothing are used in literature, but not implemented directly in this work. The data provider may have performed some corrections to the spectral data before the analyses in this report.

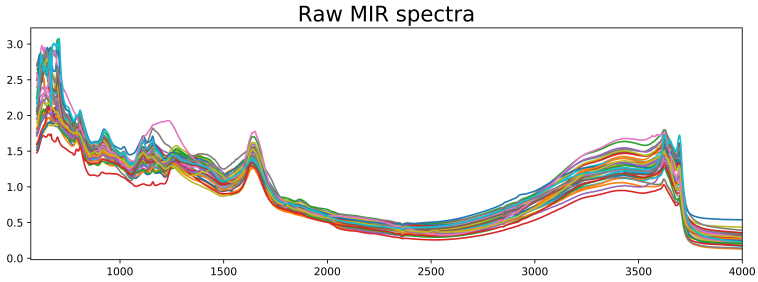
Traditionally, z-score standardisation (subtract the mean, divide by standard deviation, AKA standard normal variate) is a common technique for feature scaling. In this work, however, it is not applied in the traditional sense and is done on a per-sample basis. This normalisation does help to correct some of the baseline inconsistencies between samples. Other, more effective spectral correction techniques are available, however, per-sample z-score standardisation is quick and easy to implement. The first-order derivative is another common pre-processing strategies for spectral data. Its significant benefits are that it removes all baseline noise and is quick to implement; its negative is that the top of an absorbance peak will have a derivative of 0 making it hard to extract chemical importance from derivitised IR data.

Examples of raw, per-sample z-score standardisation and first-order derivatives are given below in figure 1. Both techniques were used throughout the report with the technique used for a given approach is clearly noted.

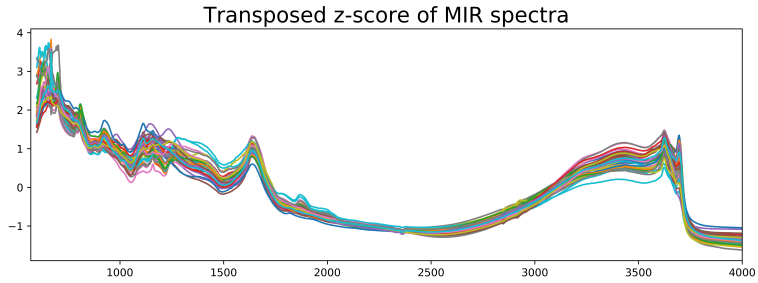
Modelling of the elemental composition of the soil and plant biomass were the primary focus of the approaches discussed in this report. It was observed that the distribution of elemental concentration was positively skewed for some datasets. To account for this skew, some of the approaches took the logarithm of the elemental concentration prior to modelling. If this pre-processing step was carried out, as such it would be stated in the approach.

3 Dataset 1

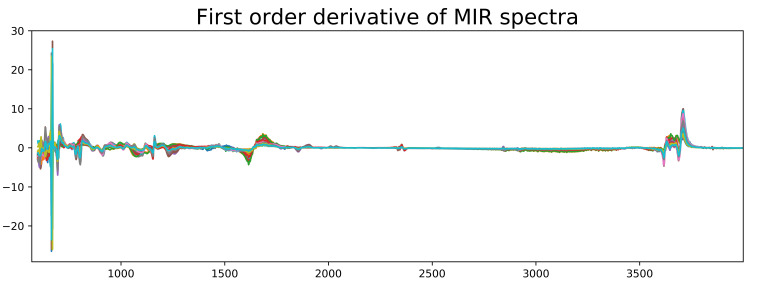
As discussed in section 2, dataset 1 was generated from a controlled experiment. Crops were grown under greenhouse conditions using different fertilisers and soils. Elemental data was provided for both the soil and the crop, moreover, the crop elemental composition was recorded at different stages of the crop's development. The variety of data in dataset 1 allowed more flexibility with explored strategies. Two approaches are discussed, the first looking at predicting crop nutrition using random forests and regularised regression and the other looking at Bayesian networks.



(a) Raw IR spectra



(b) z-score corrected IR spectra



(c) first order derivitive IR spectra

Figure 1: Example IR spectra (a) Raw, (b) transpose z-score corrected and (c) first order derivative IR spectra

3.1 Predicting crop elemental composition from spectral, pH and fertilisation treatments

This section describes our attempts to build a predictive model for crop nutrition by using data from the spectrometry of the soil, pH levels, and various fertilization treatments. One of the interesting features of this data set is that it is possible to combine information about two main sources of nutrients for the plant: 1) Those from untreated soil and 2) those from soils given various different treatments. The target variables for this exercise are the level of concentration (in ppm) of various chemical elements in the plant, at three stages of development (H1, H2 and GRAIN). We aimed to predict continuous concentrations using regression methods.

3.1.1 Method

The predictor variables comprised of the MIR data, the pH level of the soil and the fertiliser used. The fertiliser was included as a one-hot encoded categorical feature. For the MIR data each sample z-score standardised (section 2.2); the MIR data was included as a set of 100 continuous co-variates, which are the result of averaging the observed raw spectra in intervals of length 1/100th of the observed wavenumbers. 9 elements of interest (Ca, Cl, K, Mg, Mn-L, Mn-T, P, S, Zn) were used as target variables, at each development phase (H1, H2, grain). Two algorithms were compared in this approach, LASSO regression and boosted random forests. These were chosen due to the wide dimensionality of this dataset as well as the ease of their implementation.

The models were trained twice. In the first instance the target variables were kept in their raw state; in the repeat the target variables were log normalised to correct the skew in the distribution. For the log-normalised experiment, the results were inverse transformed before calculation of evaluation metric i.e. Pearson-correlation coefficient of the true and predicted values (r^2).

The following figure (figure 2) is an example of this predictive comparison for calcium, we display the scatter-plot of predicted vs observed values for three lasso models each corresponding to a different stage of the plant development.

It was observed that the prediction is sensitive to the continuous values of covariates, such as the soil spectra and pH, and also to the categorical treatment covariate, forming clusters in some cases.

LASSO Regression and boosted random forests are compared in the heatmaps below (figure 3). The r^2 value of the predicted values *v.s.* the true values are given within the heatmaps, the target variables were not log-scaled for these models. Overall predictive performance is mixed: predictions of Ca, S and P are generally high with r^2 values around 0.8-0.9. The predictive power of the models are notably weaker when applied to the grain, this makes rational sense as the grain would not have had the opportunity to absorb nutrients from the soil when being sown.

The logarithmic transformation of the response before fitting the model provided mixed results, slightly improving some of the models, but worsening oth-

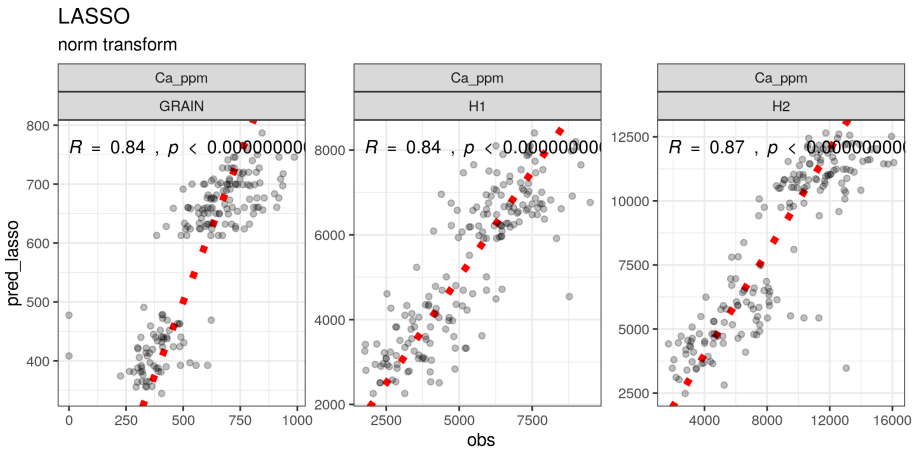


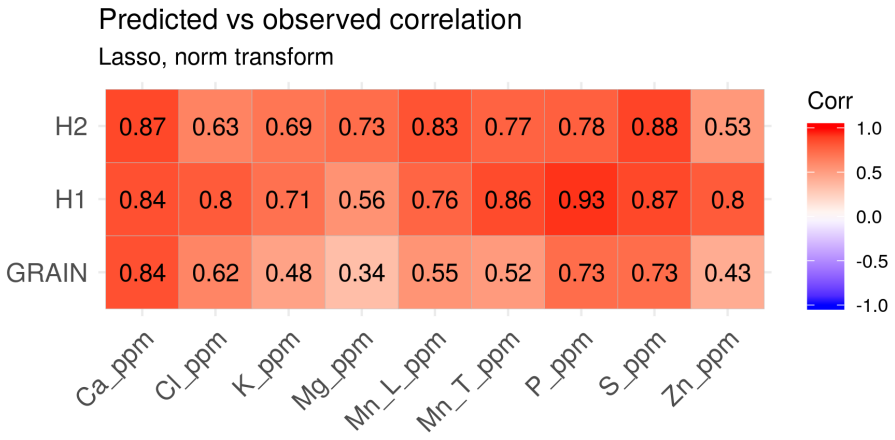
Figure 2: Observed vs predicted values for Ca at three stages of growth (Grain, H1, H2)

ers. Figure 4 shows the values for lasso with the log transformation (GRAIN stage omitted due to time constraints).

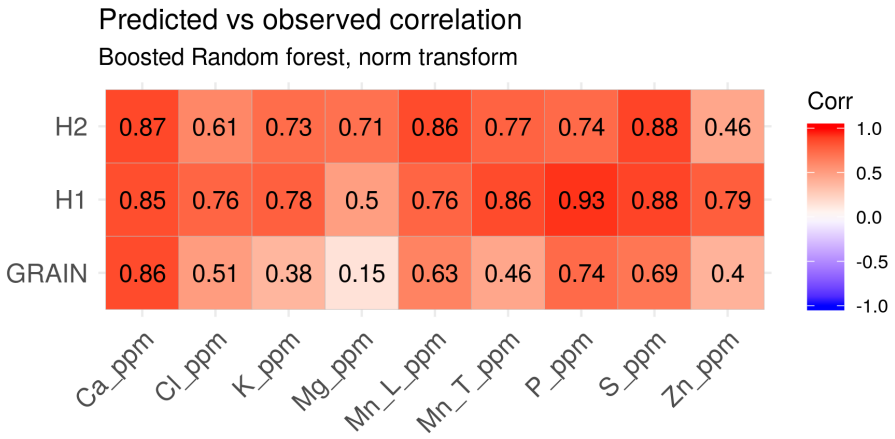
The coefficient for the LASSO models could be used to identify specific wavenumbers important for modelling spectral data. The supplementary section (section 6.2) shows examples of such information for Ca and P.

3.2 Bayesian networks

Bayesian networks (BNs) are probabilistic graphical models with a set of variables connected by conditional dependencies represented by directed acyclic graphs (Fenton & Neil, 2012). BN models can express risk assessment situations and provide visual as well as probabilistic platforms to analyse the interaction between different variables. In this study, we developed a BN model based on experts' knowledge and considering the available data in dataset 1, which includes various soil characteristics, their elements and spectra, and different treatments through fertilizing. The aim was to measure the effect of soil characteristics on each other as well as measuring the effectiveness of treatment intervention on the plant elements. As shown in Figure 5, soil variables are placed on the left side of the model, and treatment-related ones are on the right side. Soil variables consist of soil texture, soil organic carbon, 7 main elements (including Fe, Se, Ca, Zn, N, P, and K), the maximum spectrum of MIR scan, and minimum spectrum of MIR scan. We also added the pH variable which is associated with Ca and K elements, according to our domain experts. Treatment variables are categorised into environmental factors, available or uptake elements, straw, and external factor or fertilizer. Environmental variables are soil texture, soil organic carbon, room, and bench. These variables form different environments that experts had performed the experiments. Available elements are the amount



(a) Lasso regression results



(b) Boosted random forest results

Figure 3: Heatmap of correlation coefficients for (a) Lasso and (b) random forest algorithms using non-pre-processed target variables

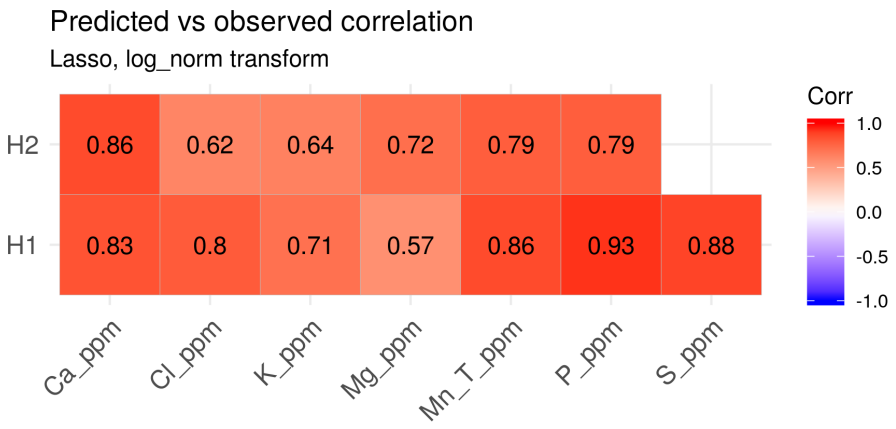


Figure 4: Results of Lasso regression models where target variables were log-normalised

of 6 elements (except N) out of above-mentioned elements of soil which are absorbed by the plant. These 6 elements are associated with straw or dried parts of the plant. Treatment includes 7 different fertilizers which are associated with all available elements.

Using the bnlearn package of R, a BN model was built and training was initialised. We expected promising outcomes based on the structure of BN model and the availability of data, so that we could investigate the effect of soil characteristics on each other and the effect of fertilizers on the available elements considering the soil characteristics. However, the results of the BN model were not given in time for the data study group deadline.

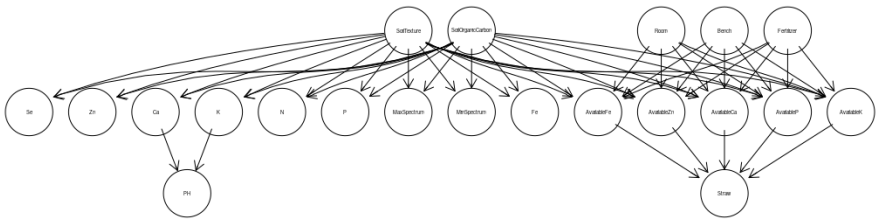


Figure 5: Bayesian network architecture

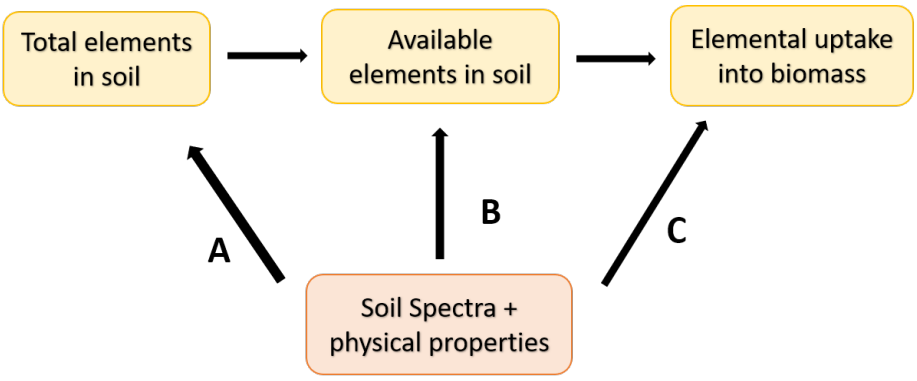


Figure 6: DS2 strategies scheme

3.3 Dataset 1 summary

Two key approaches were explored using dataset 1. The first demonstrated Lasso and boosted random forest regression to model key nutrients within the crop at various stages of growth and soil conditions. Results indicated that the accuracy of the model is dependant on the target nutrient and specific stage of growth, *e.g* Ca obtained good results at each stage of growth, while Mg was more accurately predicted at the latest stage of growth. The second approach utilised BN; using the 'BNlearn' R package a BN was constructed, however, due to time constraints could not be trained on this dataset.

4 Dataset 2

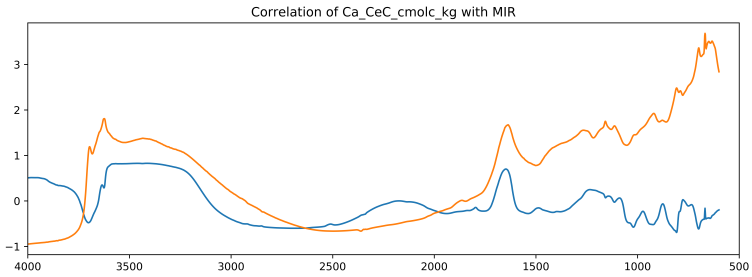
The majority of the work performed during the data study group focused on dataset 2. This was due to is relatively large sample size (456) as well as having chemical information on both the soil and the crop biomass, allowing for multiple approaches.

Unlike dataset 1, dataset 2 was not taken under controlled conditions and represented real-word data. After an initial literature review and group discussion 3 strategies were formed (figure 6): **A** - prediction of total element concentration from MIR and physical data, **B** -prediction of bio-available elements in the soil from MIR and physical data and **C** - prediction of crop biomass elemental concentration from MIR and physical data. Strategy **C** represents the most rewarding goal, as aa good prediction of element absorption allows direct information on the nutritional value of the crop and could be used to detect hidden hunger directly. Strategies **A** and **B** also have benefits as the prediction of the elemental fingerprint in the soil would enable tailored fertilisation strategies maximising the effect of fertilisation.

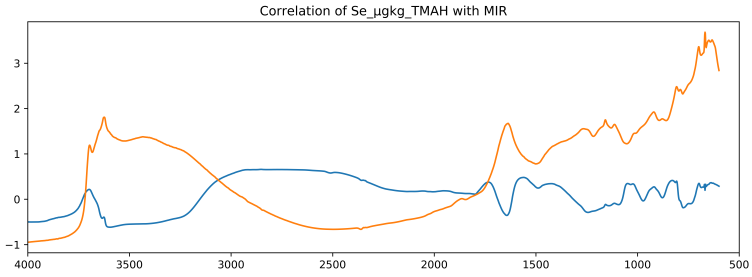
4.1 Exploratory data analysis

Prior to modelling, exploratory data analysis (EDA) was applied to the data to determine the existence of any correlations with elemental chemistry. The two approaches employed were calculating the Pearson correlation coefficients of wavenumbers with target variables and principal component analysis (PCA).

By calculating the Pearson correlation coefficient of each wave-number with elemental concentration, it was possible to determine which MIR peaks had a positive, negative or negligible correlation with that element. This is exampled using available Ca, and Se; the correlation of elemental concentration for these target variables are plotted against wave-number in figure 7. The correlation coefficient is shown in blue with a reference MIR spectra shown in yellow. A peak at 1600 cm^{-1} has a positive correlation with Ca and negative correlation with Se, suggesting that this could be a viable technique for targeted analysis. Another note on these plots is that the correlation coefficient is relatively flat despite the high spectral baseline level at 1500 cm^{-1} and below. It was suspected that the poor baseline at this region is a consistent trait for the IR spectra collected in this work and should have minimal impact on downstream modelling.



(a) Correlation of Ca with IR data



(b) Correlation of Se with IR data

Figure 7: Pearson correlation coefficients of (a) Ca and (b) Se with IR wavenumbers. The correlation coefficient is given in blue; for reference the mean of all IR spectra is given in yellow

PCA was also applied to the data set. The first-order derivative of the spectral data was taken prior to PCA. The first two principal components of the spectral data are plotted below using target variables as a colour gradient (available Calcium and Selenium) (figure 8). Correlations between these target variables and the PCA data can be observed, this is particularly apparent for Ca (figure 8 (a)) where the first principal component pulls apart samples with high and low Calcium.

The correlation between an available soil element and its content in the plant biomass was explored. The exploration was limited to the micro and macro nutrients of importance. Surprisingly, the only correlation between an available soil element and it's content in the plant biomass is for the micro-nutrient Selenium. The correlogram is given in figure 9. The results may seem poor, however, one factor that wasn't considered at the time was the crop's natural bio-availability. The correlogram was produced using data from multiple crops, however, different crops may naturally absorb more of a particular element. In the future, this would need to be taken account of when calculating correlations between soil content and plant uptake.

4.2 Strategy A - predicting total elemental concentration in the soil

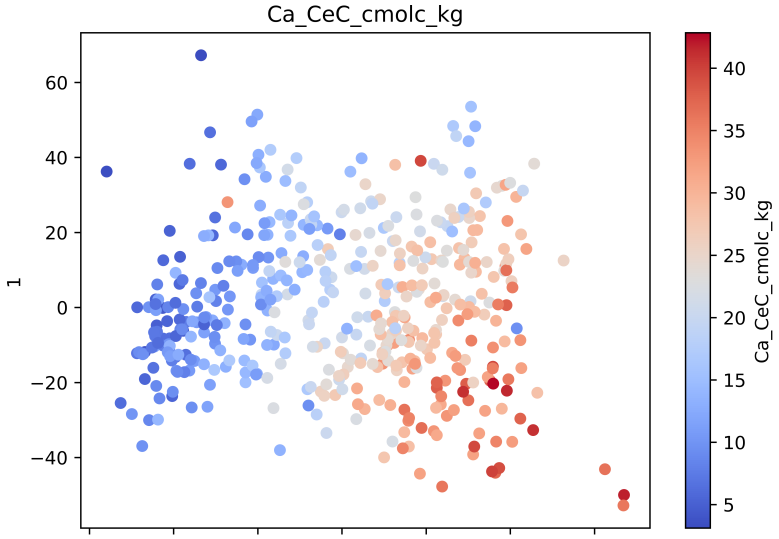
This approach focuses on the prediction of the total Ca and Se in the soil directly from the MIR spectra. Ca was selected as the literature indicated it was easily modelled, while Se was selected as the data challenge owners stated Se was an element of particular interest. Two algorithms are compared: Linear support vector regression (SVR) and random forests (RF)

The IR data was transpose z-score standardised prior to modelling (see section 2.2), the target variables were left as default *i.e.* not log normalised. The entire of dataset two was used (n=456) with training and test sets drawn randomly at 70% and 30% respectively and kept using the same split for models throughout this approach. The Pearson correlation coefficient between observed and predicted values (r^2), as well as the root mean squared error are given below in table 1.

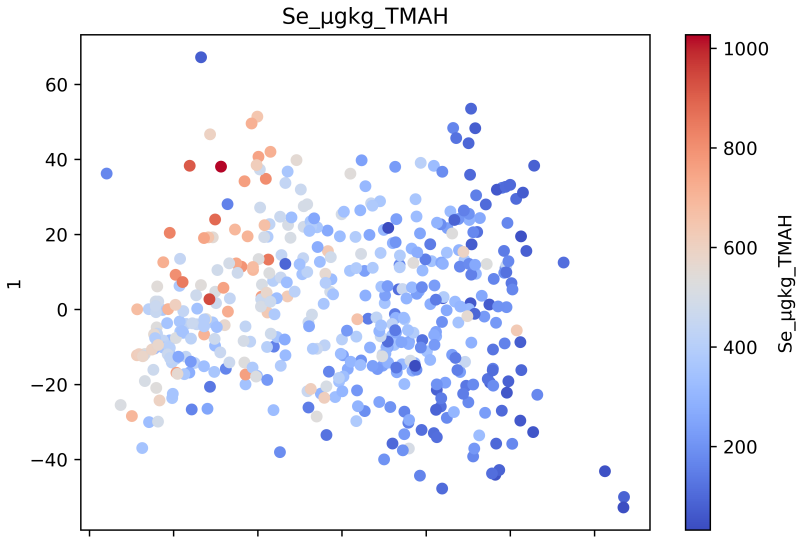
Table 1: Table comparing the prediction of total Ca and Se in soil from support vector regression and random forest

Element	Model	r^2	$RMSE$
Ca	SVR	0.98	1.89
Ca	RF	0.87	4.62
Se	SVR	0.89	82.54
Se	RF	0.79	112,88

Overall SVR outperformed RF and obtained a r^2 of 0.98 for Ca and 0.89 for Se. At the time the Ca result obtained by SVR was typical of literature values, whilst Se was above what was seen in the literature. Other test set splits were



(a) PCA of IR data coloured by Ca



(b) PCA of IR data coloured by Se

Figure 8: PCA analysis IR data coloured by elemental concentration: (a) Ca and (b) Se with IR wavenumbers. The first and second principal components are plotted.

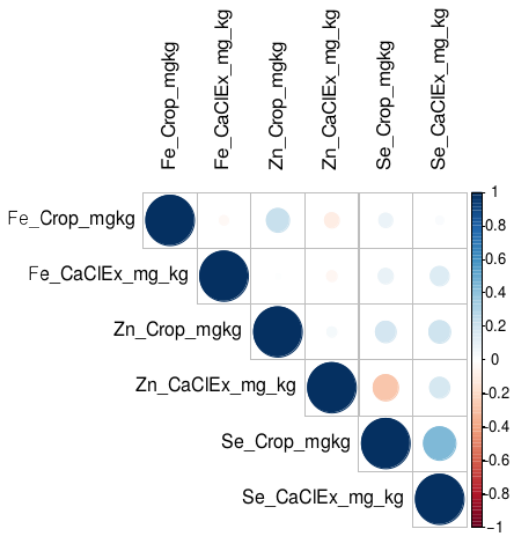


Figure 9: Correlation between an element's concentration in soil and crop biomass

note explored, in future work, a more robust validation strategy, such as k-fold validation, should be applied. Tuning of the algorithms hyperparameters was not explored for this approach.

4.3 Strategy B - predicting bio-available elemental concentration in the soil

Earlier exploratory analysis had shown that the availability of (some) nutrients is predictive of their eventual crop uptake. We therefore investigated the potential to build predictive models of available nutrients in the soil from the spectra, with available nutrients acting as a proxy for eventual crop nutrition.

Each of the target variables were attractive targets, with the target variable log transformed prior to modelling to correct skew. The test dataset was drawn randomly and consisted of 30% of the entire data set. Support vector regression (SVR) models were trained using grid search cross validation to predict optimal hyperparameters. The correlation coefficient between the true and predicted values was used as a metric (r^2), results shown in table 2, (hyper-parameters used for SVR are given in section 6.3, table 13).

Table 2: Prediction results of the available element concentration in soil from support vector regression

Element	Process	Spectra	r^2
Fe	Fe_OxEx_mg_kg	Raw	0.656
P	pbi	Raw	0.869
K	K_CaClEx_mg_kg	Raw	0.16
P	P_CaClEx_mg_kg	Raw	0.2
Fe	Fe_CaClEx_mg_kg	Raw	0.0566
Zn	Zn_CaClEx_mg_kg	Raw	0.455
Se	Se_ugkg_KH2PO4	Raw	0.271
Se	Se_ugkg_KNO3	Raw	0.314
Se	Se_ugkg_TMAH	Raw	0.717
Zn	ZnDTPA_mgperkg	Raw	0.293
Fe	Fe_OxEx_mg_kg	1der	0.717
P	Olsen_P_mg_kg	1der	0.521
P	pbi	1der	0.9
K	K_CaClEx_mg_kg	1der	0.316
P	P_CaClEx_mg_kg	1der	0.372
Zn	Zn_CaClEx_mg_kg	1der	0.436
Se	Se_ugkg_KH2PO4	1der	0.33
Se	Se_ugkg_KNO3	1der	0.396
Se	Se_ugkg_TMAH	1der	0.586
Zn	ZnDTPA_mgperkg	1der	0.292

Good models can be easily trained to predict the available soil elements from the MIR, especially Phosphorus, Iron and Selenium. As a future direction, supplementing the spectra with easily measured additional covariates, *e.g.* pH could help to bridge the "predictive gap" between soil and crop nutrients (for example if pH or clay composition modulates the ability of the plants to up-

take Iron). Predicting crop nutrition directly from the MIR spectra is a hard problem, however from the limited time available to train the models, there is a reason to be optimistic. Models trained on the spectra alone are already able to capture portion of the variance between predicted and observed crop nutrient concentrations (25.9 % for Zinc) and the same methods are able to build highly predictive models of the soil available nutrients (90 % for Phosphorus) suggesting a small number of additional measures may enable models to uncover the relationship between the spectra and eventual crop nutrition.

4.4 Strategy C - predicting crop biomass concentration from soil spectra

Strategy C was the most explored area of data set 2. This is because of the high potential impact of being able to predict the nutritional value of a crop directly. Accurate predictions of crop nutritional content could be used to examine hidden hunger directly and would mostly mitigate the need for strategies A and B.

A variety of crops were used in dataset 2; this adds a limitation as each crop type will have a different bio-availability. To bypass this limitation, the crops were modelled independently, with the effort focused on two crops, wheat and teff, with samples sizes of 134 and 110 respectively. These crops were focused on as they were the largest in terms of sample size, they were also assumed to be more prevalent in sub-Saharan Africa, and therefore more important in terms of crops to target. For both the wheat and teff dataset, training and test datasets were created randomly with the test dataset consisting of 30 % of the total dataset; the split was kept consistent between the models to allow direct comparison. The metrics we use are r^2 for time series methods and the predicted mean square error for regression models.

4.4.1 Regression based approaches to predicting crop nutrition

This section describes the regression models implemented in order to predict plant nutrients from a large number of mid-infrared (MIR) spectra. We focus on predicting a set of micro nutrients: Ca, Fe, Zn, Se and macro nutrients phosphorus (P) and potassium (K). The goal is to predict continuous measurement of crop nutrient value from MIR spectra data using functional data and time series forest regression models.

The models used in these approaches are as follows:

- **tsf_mir:** time series forest regressor using raw MIR spectra data, ensemble of 200 time series trees, where each time series tree (i) randomly segments the spectra into 69 (sqrt of length of spectra) random intervals, (ii) then extracts three features on each interval (mean, std and slope), and (iii) finally fits a regression tree on the extracted features (Deng 2013)

- **tsf_min_max_mir:** like tsf_mir, but with additional features (min and max)

- **fPCA-Regression:** functional regression base on principal component analysis
- **Kernel-Regression:** non-parametric regression based on a function kernel estimator
- **LMDC-Regression:** Regression approach base on local maxima distance correlation; identifying impact points selection of functional predictors
- **SVR_mir:** Support vector regression applied to the MIR data - hyperparameters given in supplementary

We present the results of applying a functional regression based on principal component analysis (fPCA-Regression), a nonparametric regression on functional kernel estimator, and regression using local maxima distance correlation (LMDC). We summarise the results in table 3 for teff and table 4 for wheat. Showing the different methods and different set of nutrients.

Table 3: Model results for predicting crop nutrition in teff

methods	Zn	Fe	Se	Ca	K
tsf_min_max_mir	-8.84	-0.34	-9.09	-34.08	-25.37
tsf_mir	-8.65	-0.36	-9.01	-34.21	-25.44
fPCA-Regression	1.25	1.08	0.46	1.26	1.20
Kernel-Regression	1.04	1.05	0.41	1.04	0.99
LMDC-Regression	0.91	0.99	0.50	0.91	1.17

Notes: The table shows the MSE of true versus predicted values.

Table 4: Model results for comparing crop nutrition in wheat

methods	Zn	Fe	Se	Ca	K
tsf_min_max_mir	-8.37	-1.05	-32.02	-2.65	-31.09
tsf_mir	-8.22	-1.04	-32.90	-2.67	-30.92
fPCA-Regression	1.13	1.02	0.81	0.86	1.22
Kernel-Regression	0.86	0.98	0.79	0.94	1.0
LMDC-Regression	0.83	0.96	0.90	0.77	1.13

Notes: The table shows the MSE of true versus predicted values

Regression methods show the mean square error of the predictions in the test set. We observe from tables 4 and 3 that the kernel regression and LMDC methods perform well with respect to the time series methods for some ingredients. We can observe that there are some ingredients for which the regression methods perform better, for example, the Se. However, we observe that for the Zn, Fe, Ca and K, the functional principal component regression method does not perform as well as we expected. We observe that for two different nutrients, Se and Ca, the predicted mean square error is lower than 1. For this particular dataset, the number of spectra curves in the test set is greater than the test set

for the Teff dataset. For the macronutrient (K), the predicted MSE is greater than 1 for all the regression models. Additionally, it was decided that the mean squared error was not the best metric for comparing models on different target elements; each elemental target has a different range and typical abundance this is not taken into the MSE metric.

Predicting crop nutrition using SVR

Support vector regression was also applied to the prediction of crop elemental concentration. This section is reported separately from section 4.4.1 primarily as the metric for model accuracy is the r^2 value between the prediction and results not MSE. Additionally, the SVR models had their hyperparameters tuned (details in supplementary section 6.3). The benefit of using r^2 as a metric is that the prediction of multiple target variables with different ranges can be more directly compared.

For this work the elemental concentration in the teff crop was predicted and the same train-test split as in section 4.4.1 was used. The best models Ca, Zn, Se are shown in table 5

Table 5: Predicting crop nutrition in wheat using support vector regression

Element	spectral pre-processing	Transformed target variable	R^2
Ca	z-score	No	0.122
Zn	z-score	No	0.108
Se	derivatised	Yes	0.24

Overall the r^2 values are low, below 0.25, indicating that either crop elemental concentration is hard to model with the available data, or different approaches are required. The results also indicate that a variety of pre-processing strategies as well as model hyper-parameters may be needed to fully exploit the data. Ca and Zn were best modelled when the IR spectra were not baseline corrected and the target variables not log-normalised, however, Se was best modelled when the derivatives of the IR spectra were taken and the target variable log-normalised. The optimal SVR kernel is dependant on the modelled element as well with Zn and Se being more effectively modelled with a radial basis function kernel (RBF) and Ca with a linear kernel.

Overall the prediction of crop elemental concentration using MIR spectra is challenging. A variety of algorithms including time-series forests, SVR and various functional regression algorithms were used to model data set 2. The low results between all of the approaches suggest that more information is needed (*e.g.* on farming practices or rainfall amounts) or that different modelling strategies are required.

4.4.2 Binned classification approaches to predicting crop nutrition

For this approach, the elemental concentration of Zn and Fe within the plant biomass was modelled directly. A classification type approach was chosen over regression; the two continuous target variables were binned into three classes

using the 25th and 75th quartile ranges. There are two justifications for reducing the prediction problem to a classification problem: (i) predicting classes may be easier than predicting exact continuous values, particularly if signal does not vary linearly with elemental concentration (ii) being able to predict whether crop nutrient contents will be low, medium or high may already be enough to inform decisions about farming practices. Whilst binning the target variable simplifies the problem, it may create boundaries between these pseudo-classes. Classification of samples near these quartile boundaries may be difficult.

Two crops were chosen and modelled independently, wheat and teff, these were chosen due to their higher sample size. It is worth noting that the binning of the target variables was done prior to crop selection, therefore, classes would be biased against a crops normal bio-accumulation of Fe and Zn. In future work, variable binning should be performed after crop selection. Data was randomly split into two disjoint sets, a training set used for fitting methods and a hold-out test set used for evaluating the fitted method on unseen data, using 75% of the samples for training and 25% for evaluation. The models explored were:

- **rf_wetchem:** random forest classifier on wet chemistry data was used as a baseline, for comparison with other methods that used raw MIR spectra data
- **rf_mir:** random forest classifier using raw MIR spectra data, ignoring any ordering or serial correlation of the wavenumbers in the spectra
- **tsf_mir:** time series forest classifier using raw MIR spectra data, ensemble of 200 time series trees, where each time series tree (i) randomly segments the spectra into 69 (sqrt of length of spectra) random intervals, (ii) then extracts three features on each interval (mean, std and slope), and (iii) finally fits a classification tree on the extracted features (Deng 2013)
- **tsf_mir_min_max:** like tsf_mir, but with additional features (min and max)

Table 6: Predicting crop nutrition of Wheat and Teff using binned classification method

methods	Crop	Zn	Fe
rf_wetchem:	Wheat	0.54	0.86
rf_mir:	Wheat	0.43	0.75
tsf_mir:	Wheat	0.61	0.79
rf_wetchem:	Wheat	0.65	0.47
rf_mir:	Teff	0.71	0.56
tsf_mir:	Teff	0.79	0.44
tsf_mir_min_max:	Teff	0.43	0.68

Notes: The table shows the r2 scores between true and predicted values.

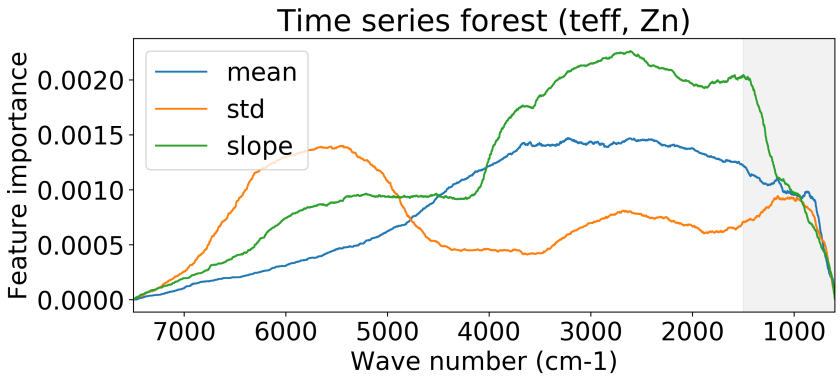


Figure 10: Time series forest feature importance

For wheat, results are mixed: Zn is best modelled by the time series forest with MIR spectra. Fe is best modelled using the wet chemistry as the predictor variables, however, when using the spectral data exclusively, the time series forest outperforms random forest. When looking at teff, MIR spectra data seems to predict crop Zn better than the wet chemistry baseline with time series forest showing the best performance (accuracy: 0.79). For Fe, the prediction seems to be more difficult, the methods generally perform worse than for Zn. Note that no confidence intervals on the performance were computed in order to assess whether observed performance differences between the different methods are statistically significant, a binomial confidence interval (or Wald interval) would be appropriate, but the test set is relatively small.

Time series forest allows the computation of feature importance graphs based on the feature importance computed by the individual trees of the ensemble and the random intervals that the individual trees used for extracting the features. Feature importance graphs may help to better understand regions of spectra and their relationship to crop nutrients, and relating the performance of the model back to the fundamental chemistry in the sample.

4.5 Additional strategy - using soil wet chemistry to predict crop nutrition

Referring to the scheme for data set 2 (figure 6), Strategy C could be reached in two ways. Direct prediction of a crop's nutritional value from the soil spectra (strategy C), or, in-direct prediction of crop nutrition through soil spectra. One such indirect method was using IR data to predict the available elemental composition of the soil (strategy B), then using these predictions to further predict the nutrition of the crop. To determine whether this would be plausible, models predicting crop nutritional values from available elemental concentration wet

chemical data were explored.

The content of each element in the crop was predicted separately. A data frame containing the complete wet-lab data was used to fit a linear model to predict the square-root transformed elemental content of the crop. The linear model was then improved by applying the stepwise Akaike Information Criterion (AIC) model selection algorithm, as implemented by the 'step()' function in R.

Performance of the model improved by splitting the data set according to crop type. There were 10 crop types in total Barley:41;Field pea:1; Finger millet:4; Maize:62; Pearl millet:1; Rice:7; Sorghum:25; Teff:134; Triticale:14, Wheat:110 however, because the number of data points for each was limited, the focus was on the two species with most data - wheat and teff.

The results of these models are shown in table 11. Of the modelled elements, Se shows the highest correlation between soil concentration and uptake into the plant biomass. While the F-statistic is supplied, the significance of the results were not explored or confirmed.

Table 7: Prediciting crop nutrition from soil available elemental composition

Element/Crop	r^2	Residual	F-statistic
Ca; Teff	0.329	3.30	3.72
Ca; Wheat	0.282	2.32	3.67
Fe; Teff	0.396	3.96	4.49
Fe; Wheat	0.381	1.30	3.68
Se; Teff	0.782	0.10	22.6
Se; Wheat	0.672	0.08	6.45
Zn; Teff	0.401	0.34	4.18
Zn; Wheat	0.625	0.39	7.72
K; Teff	0.495	3.06	5.84
K; Wheat	0.576	3.11	5.94
P; Teff	0.201	3.58	2.59
P; Wheat	0.615	3.45	6.62

4.6 Dataset 2 summary

Dataset 2 was the most explored dataset during the study group and the largest section of this report. To more effectively split the work between the team members, the problem was split into three smaller tasks: A- modelling the total elemental concentration in soil, B- modelling the bio-available elemental concentration in soil and C-predicting the elemental uptake into crop biomass.

Using SVR and RF good results on the total elemental concentration of two targets, Ca and Se were obtained with minimal model optimisation. Despite the ease of modelling these targets, minimal time was spent in improving these models or expanding to other nutrients as the total elemental concentration is not as useful as the bio-available elemental concentration.

The available elemental concentration of 10 targets was modelled using SVR; the SVR employed grid-search cross validation to optimise hyperparameters. Overall results were promising particularly for nutrients of interest like P and

Se; overall more work is required to elude the best modelling strategies for avaialable elemental concentration

Strategy C was predicted to be the most challenging, given other possible factors that could effect crop elemental concentration besides the soil. A variety of regression algorithms were explored such as functional regression (fPCA-Regression and kernel regression) and SVR, as well as novel algorithms like time-series forests. The use of two different metrics, (MSE and r^2) made direct comparisons difficult, however, results suggest that functional regression approaches were the most effective particularly for predicting Se. This problem was also approached as a binned classification problem using time-series forests, superior predictions of two targets (Zn and Fe) were obtained with r^2 values ranging from 0.4 - 0.86 for wheat and teff.

5 Dataset 3

Dataset 3 is the largest dataset in terms of sample size, comprising of roughly 2000 samples. The Spectral data for this dataset was taken using a spectrophotometer that analyses both NIR and MIR data; however, the instrument is specialised for taking MIR data. As well as the spectral data, wet chemical data on the total and bio-available elemental concentration is available; the lack of crop nutritional data limits the approaches to modelling soil physicochemical properties only. Two approaches were explored in dataset 3: (i) Direct modelling of the soils total and available elemental composition, and, (ii) memory based learning technique (Ramirez-Lopez *et al.* (2013)) that was evaluated on data set 3, but also used to predict results from dataset 2 using interpolation.

5.1 Predicting soil elemental composition from spectral data

This approach attempts to build upon findings from approach 4.2 and apply them to dataset 3. The overall goal was to predict the total and available concentration of elements within the soil from the spectral data. The MIR range of the spectral data was used exclusively, the wavenumbers pertaining to the NIR range (4000 - 12000 cm^{-1}) were dropped. Like section 4.2, both transposed z-score and first-order derivative spectral pre-processing strategies were explored. The target feature (Ca) was not log-normalised prior to modelling.

Method development was applied to Ca in the first instance. The data set was split into training and test data sets, 70% and 30% respectively; the test set was drawn at random but kept consistent within this approach. The modelling strategies consisted of linear SVR and random forests in the first instance, however, these results were lower what was observed in section 4.2, (table 8). After this initial explore other strategies were used: Due to the geographical variation within the dataset the country of origin was included as a one-hot encoded categorical feature, and a different algorithm using boosted decision trees (light gradient boosted machines (LGBM)). Note that the categorical data was

used exclusively for the boosted decision tree algorithm and was not explored for SVR or RF models.

Table 8: Comparing multiple algorithms for the prediction of Ca in dataset 3

Model	Data	Pre-processing	$RMSE$
SVR	MIR	z-score	4.19
SVR	MIR	derivative	3.62
RF	MIR	z-score	4.82
RF	MIR	derivative	4.21
LGBM	MIR + country	z-score	3.87
LGBM	MIR + country	derivative	3.55

The results from the method development suggest that LGBM combined with categorical information on the country is better able to model and predict the concentration of elements within the soil. This approach was expanded to all of the elemental data available for dataset 3. The same train-test split as the method development was used, 30% of the data was selected randomly and withheld as the test set, the exact split was kept consistent throughout the approach. Models were trained iteratively for each target variable. The Pearson correlated co-efficient (r^2) was selected as a metric for direct comparison between the different elements modelled. The results for the available and total element concentration predictions are given separately in tables 9 and 10 respectively. The models were repeated twice using different IR pre-processing procedures, z-score and first-order derivative, just like the method development, the target features were not log normalised prior to modelling.

Overall the dataset showed a good propensity to be modelled with typical r^2 values around 0.8-0.9. First-order derivative was frequently the most successful pre-processing strategy. Available P and exchangeable K, as well as total Pb, were the worst performing elements, with r^2 below 0.6.

Table 9: Results of predicting available elemental concentration in soil

Element	Z score	Derivative	Best pre-processing strategy
Caex	0.94	0.95	Derivative
eCEC	0.95	0.95	Derivative
Am Ox-Al	0.93	0.94	Derivative
Am Ox-Fe	0.88	0.91	Derivative
pbi	0.89	0.9	Derivative
Mgex	0.87	0.89	Derivative
Se 78	0.84	0.88	Derivative
AmOx-P	0.78	0.85	Derivative
AmOx-Mn	0.77	0.81	Derivative
Naex	0.74	0.73	z-score
Olsen P	0.35	0.57	Derivative
Kex	0.59	0.56	z-score

One end goal of this project was the prediction of element composition from MIR spectra of the soil. Using the above model, a rough pipeline was demon-

Table 10: Results of predicting total elemental concentration in soil

Element	Z score	derivative	Best pre-processing strategy
Al	0.96	0.98	Derivative
Ca	0.96	0.97	Derivative
Fe	0.95	0.96	Derivative
Zn	0.89	0.93	Derivative
Mg	0.9	0.92	Derivative
Co	0.88	0.91	Derivative
K	0.87	0.91	Derivative
Se 78	0.84	0.88	Derivative
Ni	0.86	0.87	Derivative
Mn	0.81	0.86	Derivative
Mo 95	0.88	0.86	z-score
Cd 114	0.83	0.85	Derivative
Cu	0.76	0.76	Derivative
Cr	0.69	0.74	Derivative
X	0.65	0.71	Derivative
Na	0.73	0.7	z-score
S	0.54	0.67	Derivative
As 75	0.66	0.65	z-score
Pb	0.46	0.45	z-score

Notes: The table shows the r2 of the true and test values for both pre-processing strategies (z-score and first order derivative).

strated as a proof of concept. A boosted decision tree model was trained on 1400 samples, this was used to predict the element composition of a test sample with known composition. The predictions were plotted in a bar chart along with the true values in a figure below (figure 11). Laying out the information in this fashion could be used to give a farmer a quick overview of their soil elemental composition and help decide what nutrients the soil should increased with for that crop to provide adequate nutrition and mitigate hidden hunger.

5.2 Memory-based learning (weighted PLSR)

Memory-based learning (MBL) is an algorithm that has been described in Ramirez-Lopez et al. (2013) within the context of soil spectral datasets. Ramirez-Lopez et al. (2013) describe that as opposed to most machine learning algorithms, it does not attempt to derive a general target function. MBL fits a weighted average partial least squares as a local model based on nearest neighbours from a given similarity matrix. At each local partition, the final predicted value is a weighted average of the predicted values from the different PLSR models.

The similarity metric used was the Mahanalobis distance between sample locations in a principal component space. The weighted PLSR was iteratively fitted with varying groups of nearest neighbours, a sequence from 40 to 150 by an increment of 10. The final number of nearest neighbours considered was decided based on the lowest RMSE from the cross-validation as a function of the number of neighbours. The validation method used was leave-nearest neighbour

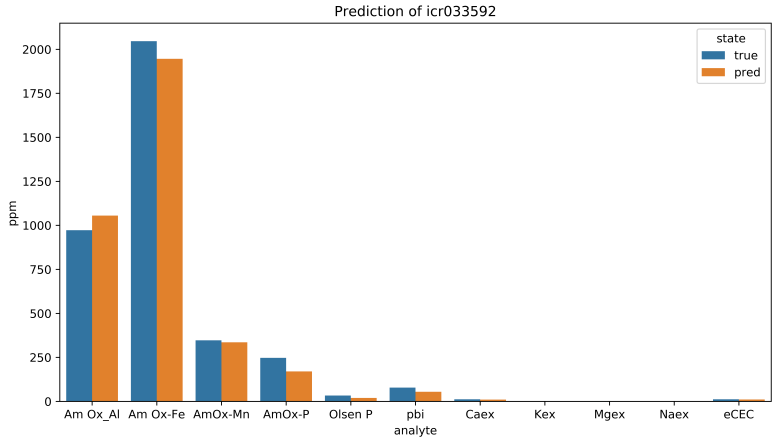


Figure 11: True vs predicted value for bio-available elemental concentration in sample ICR033592

out validation where the closest neighbour is excluded, and a local model is fitted using the remaining neighbours. An example of this for Se is given in figure 12.

Minimum and maximum number of principal components have been constrained to 4 and 17, respectively in order to minimize the risk of over-fitting. For exploratory purposes, DS3 has been split into a training and testing dataset based on a random selection of rows. Testing consists of 25% of the total dataset (n = 500) hence training (n = 1501).

Table 11: Results of memory based learning approach

Element	r^2	$RMSE$
Fe OxEx (ppm)	0.82	1825
MS-Se (ppm)	0.79	0.30

Random sub-setting may have led to overestimation of the prediction performance. For example, there might be a way of sub-setting the data to reduce bias in the testing dataset. Main sources of bias could occur from spatial correlation. For example, based on longitude and latitude but also within the sampling design (clusters of samples within the design at a given location). One way of dealing with this could be to compute a training/testing split using sampling techniques such as balanced sampling or Latin hypercube sampling.

Transfer modelling from dataset 3 to dataset 2

To assess the robustness of the model calibration based on DS3 training, the MIR spectra of DS2 were first linearly interpolated and subsampled to equal wavelength intervals; the target variable used was exchangeable cations (CEC) an important variable for healthy plant development. Dataset 3 was split into

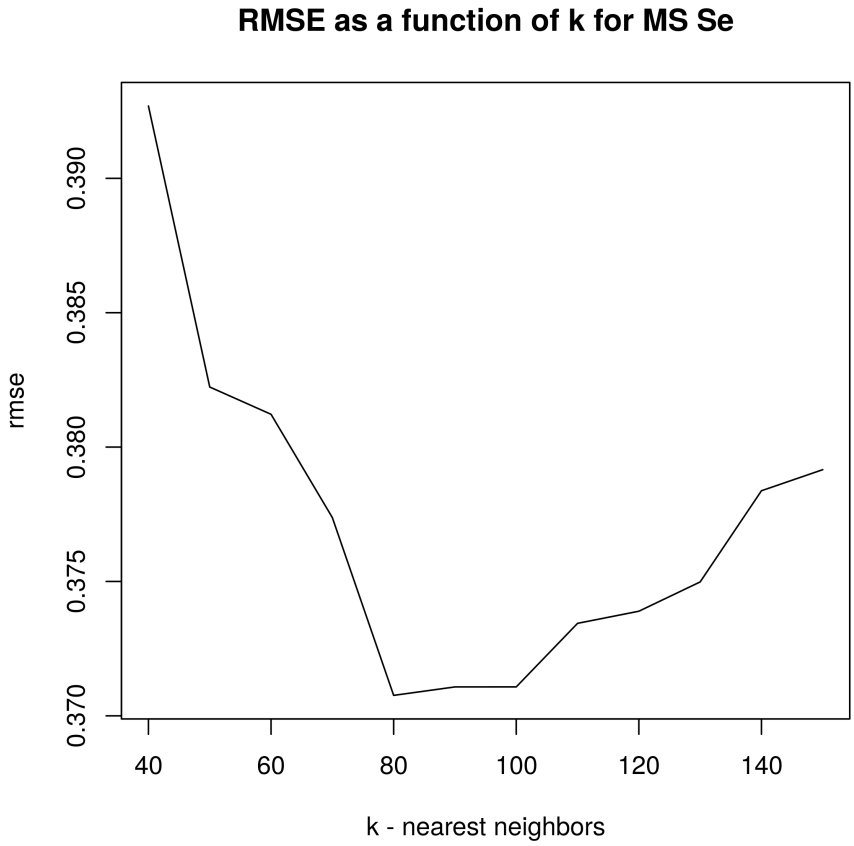


Figure 12: Nearest neighbours vs RMSE

a training and test data set; the model was trained and then used to generate predictions on dataset 3 test set, as well as the entirety of dataset 2 using the interpolated MIR spectra. Table of results given in table 12.

Table 12: Memory based learning transfered to dataset 2

Dataset	r^2	$RMSE$
CEC DS3 (Test)	0.92	4.15
CEC DS2	0.46	7.56

Transferring the model to DS2 confirms lack of generalisation, however, the model seems to give reasonable results for CEC compared to soil nutrients. But the RMSE is arguably too large to be sufficient for decision making based on predicted CEC. It is also worth noting that the wet chemical data for datasets 2 and 3 may have been calculated under different conditions / procedures; this may influence results.

5.3 Dataset 3 summary

Dataset 3 was the largest dataset with around 2000 paired soil-spectra samples; no crop data was available for this dataset. Two strategies were explored: modelling total and available elemental concentration through various regressional analyses and a memory based learning approach utilising PLSR.

For the regressional analyses three algorithms were utilised, SVR, RF and LGBM as was two different spectral pre-processing strategies z-score and derivitisation. The most effective combination was first order derivitisation combined with LGBM which managed to obtained r^2 values above 0.8 for 21 out of the 30 targets.

The weighted PLSR model also showed promising results with r^2 values around 0.8 for predictions of Fe and Se. Additionally this approach was used to transfer a model between datasets, specifically calibrating a model on dataset 3 and applying it to dataset 2. While a relatively r^2 of 0.46 was obtained, minimal time was spent on this approach. It is suspected that more work is needed to fully evaluate the potential of transferred learning between datasets.

6 Conclusion

The variety of study datasets for this challenge enabled the exploration of multiple approaches towards tackling hidden hunger. Of the three main objectives (section 1.3), the first two were pursued the most; predicting crop elemental biomass from IR spectra, and predicting soil elemental composition from IR data. Under the controlled conditions of dataset 1 it was possible to predict many elemental components within the biomass to a high degree of accuracy with typical r^2 values around 0.8. When predicting the crop biomass content of

samples from a non-controlled dataset, dataset 2, the results were significantly lower with typical r^2 values around 0.2.

Unlike predicting crop biomass, the prediction of soil composition from IR spectra was more successful and yielded significantly higher r^2 values typically in the range of 0.6 - 0.9. A variety of models prediction soil elemental composition were implemented for datasets 2 and 3. Tree-based algorithms such as boosted trees and random forest as well as SVR performed well and are discussed in detail in this report. A potentially novel model applied to IR data was time-series forests, these are tree-based algorithms that segments the spectra into pieces thus taking some serial correlations into account as well as extracting meaningful features from the data. Typical r^2 values for time series forests were between 0.5 - 0.8. Other techniques such as Bayesian networks and memory based learning were also explored in this work but need more time to be fully evaluated.

6.1 Future Work

Some ideas for future directions to explore are:

- Transfer learning across datasets, e.g. using fitted models from dataset 2 to make predictions on dataset 3
- Fitting and evaluating models trained on all crops using indicator (dummy) variables
- Trying out methods which better capture the serial correlation of the spectra (e.g. wavelet transform)
- Using a smarter interval selection using for example correlation between wavenumber and target variable of interest
- Weighting additional covariates by cost/difficulty of measurement in order to optimise another model which bridges from soil availability to crop nutrient content with the minimal economic cost. Modelling such additional covariates may also yield insights into the systems biology of the plants' nutrient uptake process, leading to a better understanding of soil types and more efficient fertilisation. Once this insight is gained, more effort can be directed towards developing *e.g.* more efficient clay composition measures to supplement the spectral readings

Acknowledgements

Dataset 1 was generated by C. Thomas, Rothamsted Research (https://www.rothamsted.ac.uk/). Dataset 2 was provided by S. Haefele (http://www.geonutrition.com/). Dataset 3 was provided by S. Haefele (http://africasoils.net/). C. Thomas, K. Hassall and A. Dowsey provided guidance during the DSG.

Bibliography

1. Deng, Houtao, et al. "A time series forest for classification and feature extraction." *Information Sciences*, **239**, 142–153 (2013).

2. Febrero Bande, Manuel, and Manuel Oviedo de la Fuente. "Statistical computing in functional data analysis: The R package fda. usc.", *Foundation for Open Access Statistics* (2012).

3. Fenton, N., and Neil, M. "Risk assessment and decision analysis with Bayesian networks." *CRC Press.* (2012).

4. Ferraty, Frédéric, and Vieu, Philippe "Nonparametric functional data analysis: theory and practice." *Springer Science & Business Media,* (2006).

5. Silverman, B. W., and Ramsay, J. O. . "Applied functional data analysis: methods and case studies." *New York: Springer,* (2002).

6. Ordóñez, Celestino, *et al.* "Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach.", *Chemometrics and Intelligent Laboratory Systems*, **173**, 41–50 (2018).

7. Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Dematte, J.A.M., Scholten, T., "The spectrum-based learner: A new local approach for modeling soil vis-NIR spectra of complex datasets.", *Geoderma*, 195–196, 268–279 (2013).

8. Hengl T, Heuvelink GBM, Kempen B, Leenaars JGB, Walsh MG, Shepherd KD, et al. "Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions." *PLoS One* (2015).

Team Members

Timo Breure is a PhD student at Cranfield University and Rothamsted Research. His research is in the use of VNIR/MIR/XRF soil spectroscopy in precision agriculture. He contributed to the report by computing weighted PLSR models to predict soil nutrient status across 18 countries in Africa.

Chris U. Carmona is a doctoral researcher in Statistical Machine Learning at the University of Oxford. His current research is focused in developing Bayesian methods designed to learn from multiple sources of information under difficult conditions, such as model mis-specification and copious missing data. His main contribution on this report was to design and implement the predictive models (Lasso and random forest) for the controlled experiment data (dataset 1), targeting crop nutrient as response and using the spectrometry of the soil, pH levels, and fertilisation treatments as predictive features (see sec 3.1 and 6.2).

Samuel Ellick is a PhD student at the University of Bristol. His research is in the field of chemometrics and metabolomics. He contributed to the report by modelling elemental composition of soils from mid-range infrared spectra, including the use of gradient boosted decision trees to predict available nutrients in dataset 3. Additionally Sam was the final report editor.

Ben Evans is a Senior Research Associate in the School of Psychological Science, University of Bristol. His research is in the domains of computational neuroscience, deep learning and machine learning where his current focus is on making deep CNNs more human-like and robust in their visual perception. For this project he contributed exploratory data analysis of Data Set 1 and built regression models from the MIR spectra for both crop and available soil nutrient concentrations for Data Set 2. He also designed the git repository structure and formulated an overall framework and strategy for conceptualising and tackling the research questions of the project.

Ali Fahmi is a PhD student in Computer Science at Queen Mary University of London. He is a member of the PamBayesian project working on diagnosis and treatment of rheumatoid arthritis using causal Bayesian networks. His area of research focuses on decision support, Bayesian networks, and causal inference. In this report, he contributed to build a causal Bayesian network model based on experts' knowledge and trained the model using available data.

Stephan Haefele is a soil scientist and agronomist with a long experience in international agricultural research focused on rice, wheat and soil quality in Africa and Asia. He conducted research for his PhD and PostDoc at the West Africa Rice Development Association, and worked then for ten years at the International Rice Research Institute. His next position was at the University of Adelaide where his main focus was plant phenotyping in wheat systems, and in 2017 he started a position at Rothamsted Research. There he leads the dry spectral laboratory and his recent work has concentrated on the use of XRF and MIR for the analysis of soil, plant and fertilizer samples. His group has now a focus on soil health, investigating a range of indicators and soil health indices. Other research activities are around nutrient use efficiency in arable systems, agronomic biofortification for micronutrients and approaches to upscaling of agronomic advice (agronomy to scale). He authored or co-authored 103 peer reviewed publications, has currently 4660 citations and an h-index of 33. He is contributing to several ongoing projects funded by the BBSRC (ASSIST, S2N) and the Bill and Melinda Gates Foundation (AfSIS, iSDA, GeoNutrition).

Kirsty Hassall is a Senior Statistician at Rothamsted Research with a keen interest in addressing the statistical and mathematical problems arising from complex biological datasets, in particular the interplay between observed data and the underlying biological processes. Recent work has focused on the incorporation of expert opinion in quantifying soil health through Bayesian Belief Networks and efficiently capturing zone information for use in managing agricultural landscapes. As a chartered statistician (CStat), she has a strong track record of providing statistical expertise into a wide range of projects. Kirsty conceived the data challenge problem and provided guidance throughout the study week.

Markus Loning is a PhD student at UCL and an Enrichment Student at The Alan Turing Institute. His research focuses on machine learning with time series/panel data and toolbox development. He contributed to the report by implementing and evaluating crop nutrient predictions from spectral data, including the time series forest regression and classification algorithms.

Diego Perez Ruiz is a Research Associate in Statistics at the department of Mathematics at the University of Manchester. He holds a PhD in Statistics and a MSc in Probability and Statistics. His research focuses in the field of Functional Data and Nonparametric statistics. In this report, he contributed by implementing and evaluating crop (micro and macro) nutrients predictions from spectra data following a functional data approach. He also proposed and implement classification algorithms, including the functional data regression and classification algorithms.

Darya Shchepanovska is a PhD student at the University of Bristol at the Centre for Computational Chemistry. Her research deals in the modelling of light induced reactions that occur in the troposphere. She was the group facilitator and contributed by building linear models of crop nutrition from wet lab data.

Cathy Thomas is a postdoctoral researcher at Rothamsted Research, focusing on geophenomics which uses innovative dry spectral techniques for the analysis of soils and crop tissues to understand how fertilisers can be tailored to the specific conditions of soils. She gained her PhD in Crop Science from the University of Nottingham in 2016.

Prof Andrew Dowsey is Professor of Population Health DataScience at the University of Bristol, and group leader of the BioSPI Laboratory. He is a Turing Fellow and holds research programmes in data science methodology for mass spectrometry omics, antimicrobial resistance epidemiology, and animal biometrics. Prof Dowsey was the University of Bristol data science lead for this Data Study Group.

Supplementary

6.2 Lasso feature importance graphs - Dataset1

Relevance of predictors As one of our covariates in the models was the spectra of the soil, we are able to identify frequencies that are more informative for the nutrition of the plants.

6.3 SVR hyperparemeter tuning - Dataset 2

A train/test split of 3:1 was used. The following parameters were explored with a 5-fold cross-validated grid search 'GridSearchCV'. The following kernels were explored: linear, rbf, sigmoid and poly with the accompanying hyperparame-

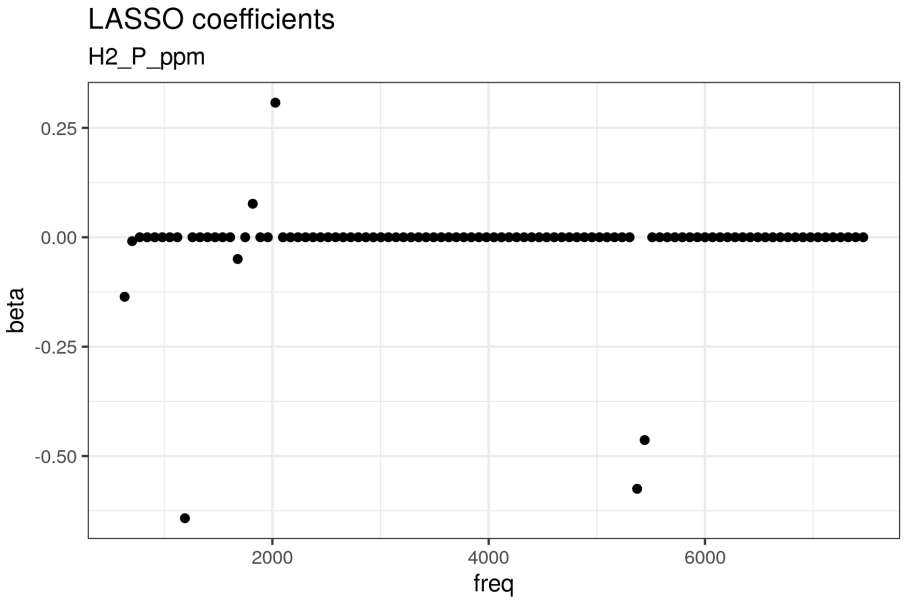


Figure 13: Lasso Model coefficients predicting P (dataset 1)

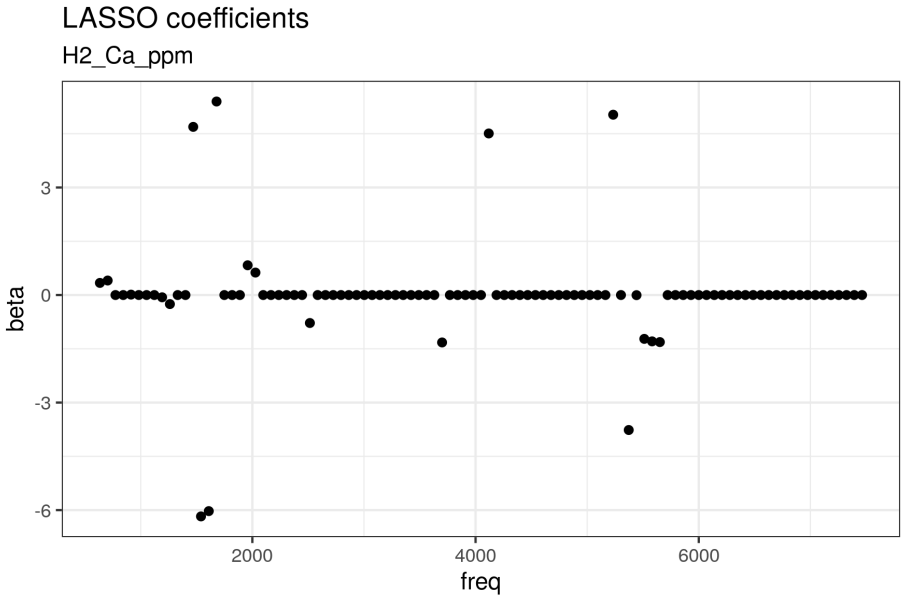


Figure 14: Lasso Model co-coefficients predicting Ca (dataset 1)

ters[1] :

- Cs = np.logspace(-3, 3, 19)
- gammas = np.logspace(-5, -1, 13)
- coef0s = [0.1*i for i in range(1, 10)]
- degrees = list(range(1, 4))

Table 13 shows the SVR parameters used in data set 2 regression approaches to predicting available soil composition (section 4.3).

Table 13: Prediction results of the available element concentration in soil from support vector regression with hyper parameters

Element	Process	Spectra	r^2	Hyperparameters
Fe	Fe_OxEx_mg_kg	Raw	0.656	rbf, $C = 1000.0$, $\gamma = 0.00001$
P	pbi	Raw	0.869	linear, $C = 0.00464$
K	K_CaClEx_mg_kg	Raw	0.16	'C': 0.01, 'kernel': 'linear'
P	P_CaClEx_mg_kg	Raw	0.2	'C': 0.01, 'kernel': 'linear'
Fe	Fe_CaClEx_mg_kg	Raw	0.0566	'C': 0.464, 'gamma': 0.00464, 'kernel': 'rbf'
Zn	Zn_CaClEx_mg_kg	Raw	0.455	'C': 10.0, 'gamma': 2.15e-05, 'kernel': 'rbf'
Se	Se_ugkg_KH2PO4	Raw	0.271	'C': 0.001, 'kernel': 'linear'
Se	Se_ugkg_KNO3	Raw	0.314	'C': 0.01, 'kernel': 'linear'
Se	Se_ugkg_TMAH	Raw	0.717	'C': 100.0, 'gamma': 1e-05, 'kernel': 'rbf'
Zn	ZnDTPA_mgperkg	Raw	0.293	'C': 215.4, 'gamma': 2.15e-05, 'kernel': 'rbf'
Fe	Fe_OxEx_mg_kg	1der	0.717	'C': 46.4, 'gamma': 4.64e-05, 'kernel': 'rbf'
P	Olsen_P_mg_kg	1der	0.521	'C': 100.0, 'gamma': 1e-05, 'kernel': 'rbf'
P	pbi	1der	0.9	'C': 21.5, 'gamma': 1e-05, 'kernel': 'rbf'
K	K_CaClEx_mg_kg	1der	0.316	'C': 21.5, 'gamma': 1e-05, 'kernel': 'rbf'
P	P_CaClEx_mg_kg	1der	0.372	'C': 21.5, 'gamma': 1e-05, 'kernel': 'rbf'
Zn	Zn_CaClEx_mg_kg	1der	0.436	'C': 10.0, 'gamma': 1e-05, 'kernel': 'rbf'
Se	Se_ugkg_KH2PO4	1der	0.33	'C': 10.0, 'gamma': 1e-05, 'kernel': 'rbf'
Se	Se_ugkg_KNO3	1der	0.396	'C': 0.001, 'kernel': 'linear'
Se	Se_ugkg_TMAH	1der	0.586	'C': 1.0, 'gamma': 1e-05, 'kernel': 'rbf'
Zn	ZnDTPA_mgperkg	1der	0.292	'C': 0.00215, 'kernel': 'linear'

[1]After some initial exploration of the parameter space, the sigmoid and poly kernels were removed from further model fitting in order to avoid over-fitting and extended training times.

turing.ac.uk
@turinginst