Crop Canopy Nitrogen Concentration Algorithm

By: ASU Carbon Mapper Land and Ocean Team

Version 1.0

1. Algorithm Description

The objective of the Crop Canopy Nitrogen Concentration (CCNC) algorithm is to retrieve mass-based Nitrogen (N) concentration (in % by mass) of crop top canopy from Tanager imaging spectroscopy data (Dai et al., to be submitted). Input data can be any visible-to-shortwave infrared (VSWIR) imaging spectroscopy surface reflectance data covering the 420-2440 nm wavelength range. This algorithm was designed to work with spectral resolutions around 5 nm, but finer or coarser resolutions can be accommodated via band resampling. Input data is expected to be integer-valued % reflectance scaled such that 0 and 100% are 0 and 10,000, respectively, though other data types and gains can be specified.

The algorithm applies an array of pre-calibrated and validated spectral coefficients to input data and produces a single-band N concentration map. This N retrieval algorithm/model is an empirical one, developed from linking laboratory-measured N concentration with sensor-collected imaging spectroscopy data through partial least squares regression (PLSR). The coefficients for this algorithm were developed using GAO data with 0.6 m spatial resolution, but they were later tested with simulated 30 m resolution simulated Tanager data.

The algorithm is designed to work on well-lit image pixels dominated by photosynthetic vegetation, and two spectral filters to remove unfit pixels are applied to the reflectance data during the each run of the algorithm: Input pixels should have Normalized Difference Vegetation Index (NDVI) values no lower than 0.7 and reflectance value higher than 25% at the band representing 1070 nm. Under these conditions, the averaged validation R^2 of model fit was 0.71 and RMSE was 0.56 % N, which easily met the minimal performance target of RMSE <= 1% N and nearly met the nominal target of RMSE <= 0.5% N.

2. Benchmark Data

Input data include VSWIR imaging spectroscopy data collected by Global Airborne Observatory (GAO) during a suite of campaigns in 2022. As reference data, we collected leaf samples in the field and subsequently processed these in a laboratory to derive N concentrations of the crop foliage at known locations as within a few days of the airborne data collection.

Acquisition details

To obtain crop foliage samples, we visited both research and commercial farms in four of the five major U.S. farming regions (USDA 2022), including the South, the Midwest, the Plains, and the West (Figure 1). Overall samples were collected across eight sites located in six states

(Table 1).

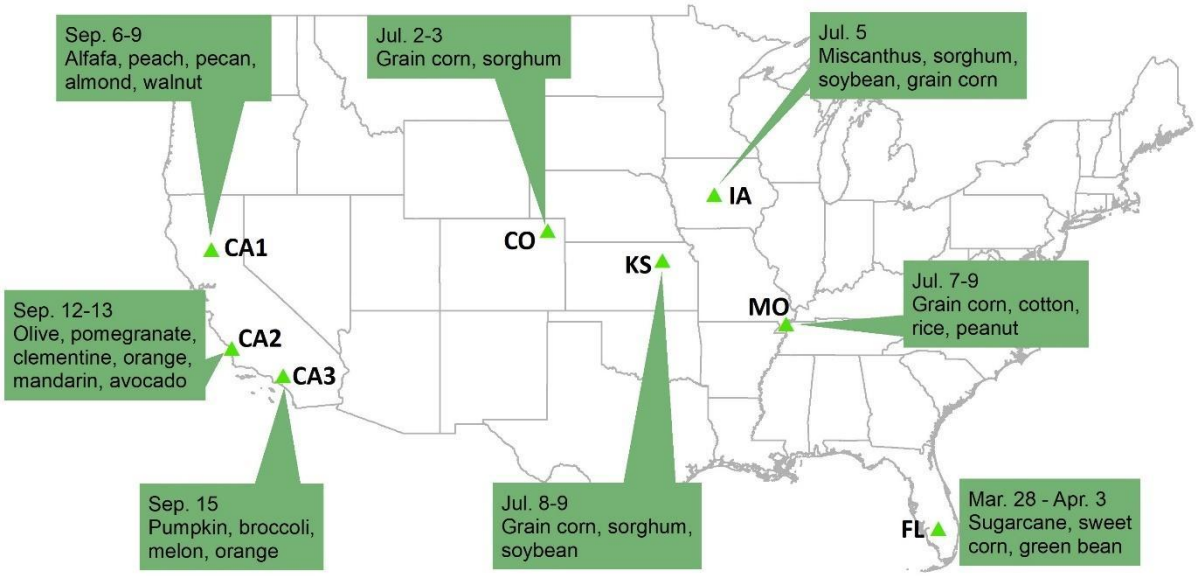

Figure 1. Geographical locations of study sites in the contiguous United States, with relative field sampling dates and crop species labeled. All fieldwork was conducted in the year 2022.

Table 1. Sampling sites, field and image data collection dates. See Section 3.2 for more details.

Site	Farm Name	Latitude	Longitude	Field Dates	GAO Dates
FL	Rouge River Farm	26.69	-81.17	03/28 – 04/03	03/28-04/10
CO	NEON STER Site	40.48	-103.01	07/02 – 07/03	07/02
IA	Iowa State University Farm	42.00	-93.70	07/05	07/09
KS	Kansas State University Farm	39.21	-96.59	07/08 – 07/09	07/10
MO	University of Missouri Farm	36.41	-89.42	07/07 – 07/09	07/10
CA1	Chico State University Farm	39.68	-121.82	09/06 – 09/09	09/04
CA2	Cal Poly San Luis Obispo Farm	35.30	-120.67	09/12 – 09/13	09/03
CA3	Cal Poly Pomona Farm	34.04	-117.82	09/15	09/06

Sampled species included staple crops such as corn and soybean as well as cash crops such as Miscanthus and orchard fruits. A complete list of species collected at each site and their relative sample sizes can be found in Table 2. The complete dataset included 471 individual foliage samples, representing 24 species collected from the eight sites

Table 2. Sample size of each species at all sites.

Crop	Site	Sample Count
Sugarcane	FL	24
Sweet corn	FL	15
Green bean	FL	16
Grain corn	CO	20
	IA	20
	KS	15
	MO	50
Sorghum	CO	20
	IA	20
	KS	20
Cotton	MO	20
Rice	MO	20*
Peanut	MO	15
Soybean	IA	20
	KS	20
Miscanthus	IA	20
Alfalfa	CA1	15
Peach	CA1	15
Almond	CA1	15
Pecan	CA1	15
Walnut	CA1	15
Olive	CA2	15
Pomegranate	CA2	15
Clementine	CA2	5
Mandarin	CA2	15
Avocado	CA2	15
Orange	CA2	8
	CA3	7
Pumpkin	CA3	5
Broccoli	CA3	5
Melon	CA3	5

*Only one sample location was covered by GAO imagery.

Sampling and collection methods

While in the field we endeavored to cover wide N concentration ranges within each crop species when collecting foliage samples. We did so by identifying crops in varied growing stages and conditions, either by consulting with local farmers and researchers or through visual interpretation. For herbaceous species, we searched for sample locations where crop conditions looked homogenous. For trees, we identified individual plants to sample from. We randomly clipped four to twenty fully grown leaves, depending on leaf size, from sunlit tops of plant crowns. Clipped foliage was immediately sealed in polyethylene bags and stored on ice in coolers to preserve moisture. GPS readings were recorded at the sample locations using

Arrow Gold RTK Global Navigation Satellite Systems receiver with estimated horizontal positioning precision <= 1cm RMSE. At each site and for each species, we normally collected 15 samples from a single field, or 20 samples in total from multiple fields if available, except for grain corn at site MO, where N experiments were being conducted and we gathered 50 samples. The complete dataset included 471 individual foliage samples, representing 24 species collected from eight sites in six states (see Section 3.2 for more details).

Physical sample processing methods

Foliage samples were processed right after we finished the entire field collection of the day. The general procedures were as follows: We first carefully wiped off dust, water or any other non-plant particles, and removed petioles from each leaf. Next, we placed foliage in open paper bags, dried them at 65 °C for at least 48 hours, so that they would not mold during transportation. After that, all samples were mailed to our lab in Tempe, AZ and stored in the drying oven at 65 °C for at least another 72 hours. Then the dried foliage was ground using a 20 mesh Wiley mill and powdered to finer particles with a high throughput homogenizer (Troemner Inc., Thorofare, NJ, USA). Sample N concentration (%) was determined by flash combustion in a conventional elemental analyzer (PE 2400; PerkinElmer Inc., Waltham, MA, USA).

Airborne data processing

GAO data collected for this study contained 428 channels of spectral information covering the wavelength range between 345 nm and 2488 nm, with a spectral resolution of 5 nm. Flight elevation was around 600 m above the ground, which resulted in a ground sampling distance of 60 cm. Light detection and ranging (LiDAR) data that were simultaneously collected with the imaging spectroscopy data were processed in canopy surface elevation maps. These surface maps were used along with precise position and orientation data from the flight were used to ortho-georeference the images with 60cm spatial resolution.

The GAO data were initially processed to calibrated radiance, which was run through the ACORN v6.0 (Atmospheric CORrection Now; AIG LLC; Boulder, CO) atmospheric correction software to retrieve surface reflectance. We trimmed the wavelengths at far ends of detection and removed regions with high levels of atmospheric absorption, resulting in 320 bands of data spanning the wavelengths 420-1330, 1500-1775, and 2030-2440 nm.

3. Algorithm Development

Spectral filters

To minimize spectral variations caused by plant biophysical conditions (Asner 1998), as well as contributions from non-photosynthetic vegetation (NPV) and substrate materials (Dashti et al. 2019), we filtered the surface reflectance data with a narrow band Normalized Difference

Vegetation Index (NDVI; near-infrared: 858 nm; red: 648 nm) mask. All pixels with NDVI value no lower than 0.7 were included in further analyzes (Fig. 2). To mitigate potential artifacts introduced by sun-sensor-canopy geometry (i.e., shade), we applied a minimum brightness threshold where pixels should have a reflectance value greater than 25% at 1072 nm wavelength (Asner et al., 2018). We used GPS records to extract VSWIR spectra corresponding to plant locations and applied brightness-normalization for the candidate spectra before further statistical analysis (Feilhauer et al., 2010). These filtered N-spectrum pairs (n = 307) were the data input for the following statistical analyzes.

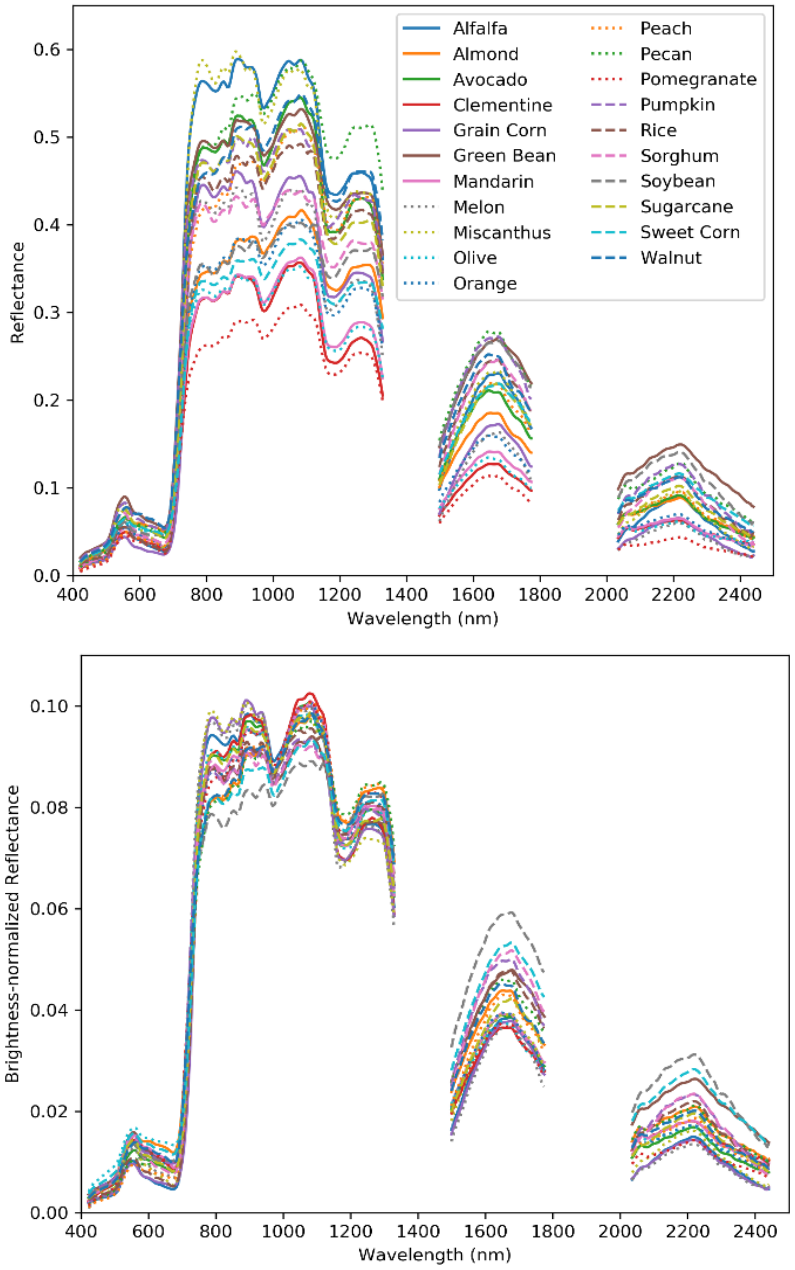

Figure 2. Mean crop spectra of original (top) and brightness normalized (bottom) pixels used for analysis following NDVI and brightness filtering.

Modeling

We used partial least-squares regression (PLSR) to link field-sampled, lab-measured N records to airborne imaging spectroscopy data collected by GAO (Asner et al., 2015; Chadwick & Asner, 2016; Martin et al., 2018). To avoid overfitting of PLSR results, we set the number of latent orthogonal vectors as determined by minimizing the Prediction Residual Error Sum of Squares (PRESS) statistics (Asner et al., 2015; Chen et al., 2004). We applied two levels of iterations in the model procedures in the calibration-validation mode (Fig. 3). At the basic level, for all N-spectrum pairs engaged in the regression analysis (i.e., original dataset), we randomly set aside 25% as global holdout dataset (i.e., validation), which was not included in model development. For each iteration of the fitting procedure, we randomly selected 1/3 of the remaining records (i.e., calibration) as the testing dataset, and 2/3 as the training dataset. Thus, the ratio of training/testing/validation data was roughly 0.5/0.25/0.25.

We repeated the model development procedures (i.e., splitting between training and testing as well as PLSR fitting) 500 times. In each iteration, calibrated PLSR results (i.e., iteration model) were tested with the testing data, and performance statistics (i.e., coefficient of determination, or R^2 and Root Mean Square Error, or RMSE, as well as the RMSE normalized to the range of values reported, or nRMSE) were generated for both training and testing datasets. After the 500 iterations were completed, training and testing performance statistics were averaged. The models with better performance statistics during validation (validation R^2 higher than the mean R^2 of the 500 iterations and validation RMSE lower than the mean RMSE of the 500 iterations) were also averaged to produce the mean model (Chadwick & Asner, 2016). Then the mean model was applied to the validation dataset to get validation performance statistics.

Next, at the higher level, all procedures described in the basic level were iterated 50 times. The training, testing and validation performance statistics were averaged for each of these 50 iterations, as were the PLSR results. In other words, final training and testing performance statistics as well as PLSR results were first averaged from the 500 iterations at the basic level, and then averaged from the 50 iterations at the higher level, whereas final validation performance statistics were averaged solely from the 50 iterations at the higher level. In total, we ran the PLSR-PRESS model 25000 (500*50) times.

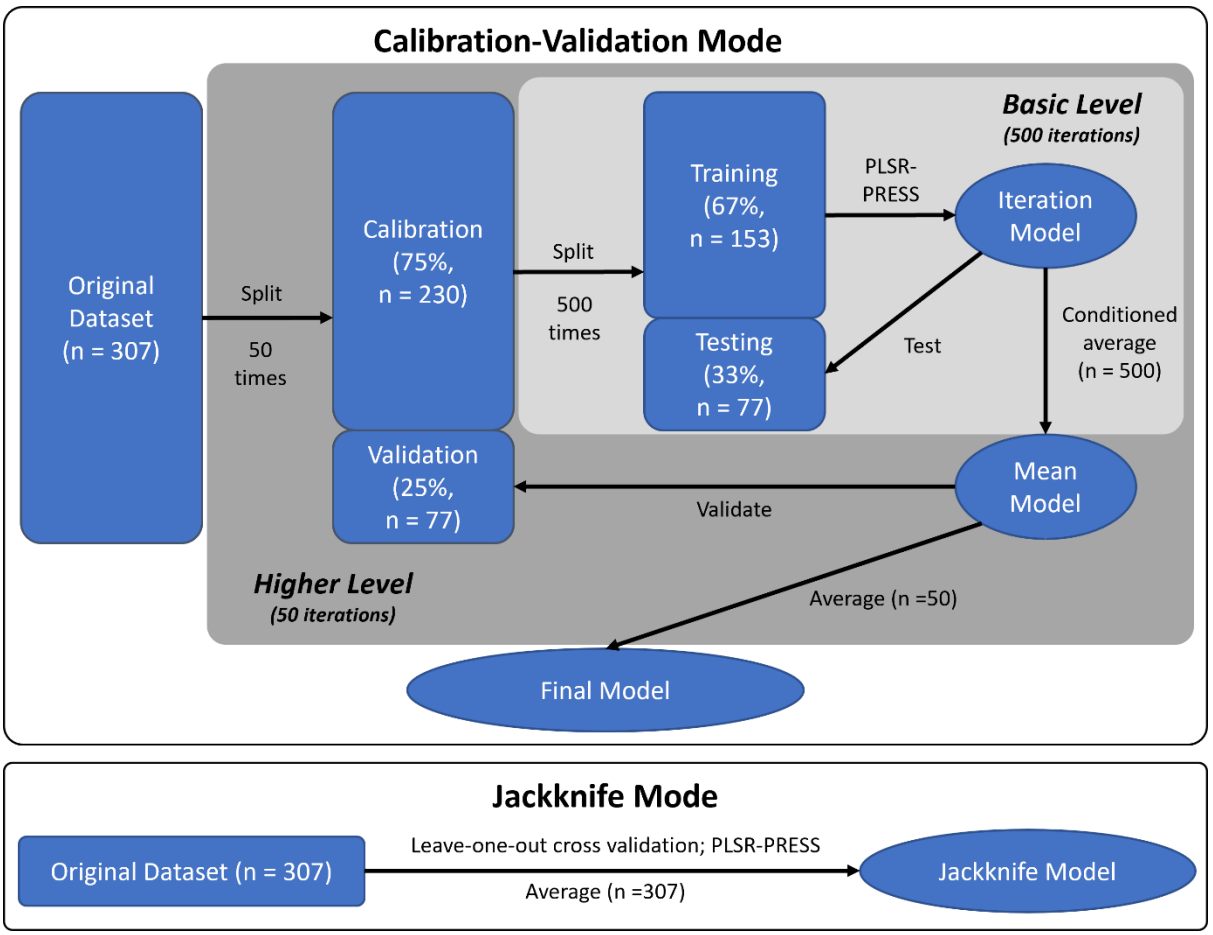

Figure 3. Two modes of model development: calibration-validation mode and jackknife mode.

To further test the robustness of the final model, we developed a jackknife model through leave-one-out cross validation based on the original dataset (Fig. 3). The performance statistics as well as the model results were averaged from these leave-one-out models. These results were compared with those of the final model described above. We also calculated the variable influence on projection (VIP) to explore the importance of specific wavelengths for predicting N concentrations (Burnett et al., 2021; Wold et al., 2001).

Performance and validation results

The filtered and brightness normalized VSWIR data calibrated well against the N concentration data observed in the laboratory, with averaged R^2 values of 0.81, 0.65, and 0.71 and RMSE of 0.47, 0.65, 0.56 % N, respectively for the training, testing and validation data sets (Table 3). The normalized RMSE values (RMSE divided by the range of N concentrations) were between 0.10 and 0.14. The standard deviations of the 50 iterations for the training and testing R^2 and RMSE were all close to 0.02. Whereas for validation, the standard deviations for the R^2 and RMSE were 0.05 and 0.04 % N, respectively. The leave-one-out Jackknife model produced similar results, with R^2 and RMSE values of 0.78 and 0.49 % N respectively. The spectral coefficients of jackknife PLSR, as well as the VIP were almost identical to those of the averaged final model in the calibration-validation mode (Fig. 4).

Table 3. Average performance statistics (with standard deviation in parentheses) for the PLSR-PRESS iterations. Normalized RMSE (nRMSE) is calculated as the RMSE divided by the range of N concentrations. Training and testing statistics were first conditionally averaged from the 500 iterations at the basic level, and then averaged from the 50 iterations at the higher level (Fig. 3). Validation statistics were averaged from the 50 iterations at the higher level.

Stage	**R^2**	**RMSE (% N)**	**nRMSE**
Training	0.81 (0.02)	0.47 (0.02)	0.10
Testing	0.65 (0.02)	0.65 (0.02)	0.14
Validation	0.71 (0.05)	0.56 (0.04)	0.12

A VIP value higher than 0.8 typically indicates wavelength of high importance to model prediction (Wold et al., 2001). The plotted VIP indicated that all wavelength regions were significant for the prediction of N concentration (Fig. 4). According to the threshold of 0.8, wavelengths of 550-900 nm, 1100-1350 nm, the whole SWIR I (1500-1800 nm), as well as the beginning and end of SWIR II (around 2050 nm and 2400 nm) were important contributors to our PLSR models. Specifically, the “red edge” region roughly between 700-750 nm appeared to be the most significant in our investigation (Smith et al., 2003). We fitted the final model with all 307 N-spectrum pairs and compared its performance for different crop species (Fig. 5). We found that the model tended to overestimate the N concentration of sugarcane, melon and nearly all avocado samples, meanwhile underestimating almost all sweet corn and peach ones.

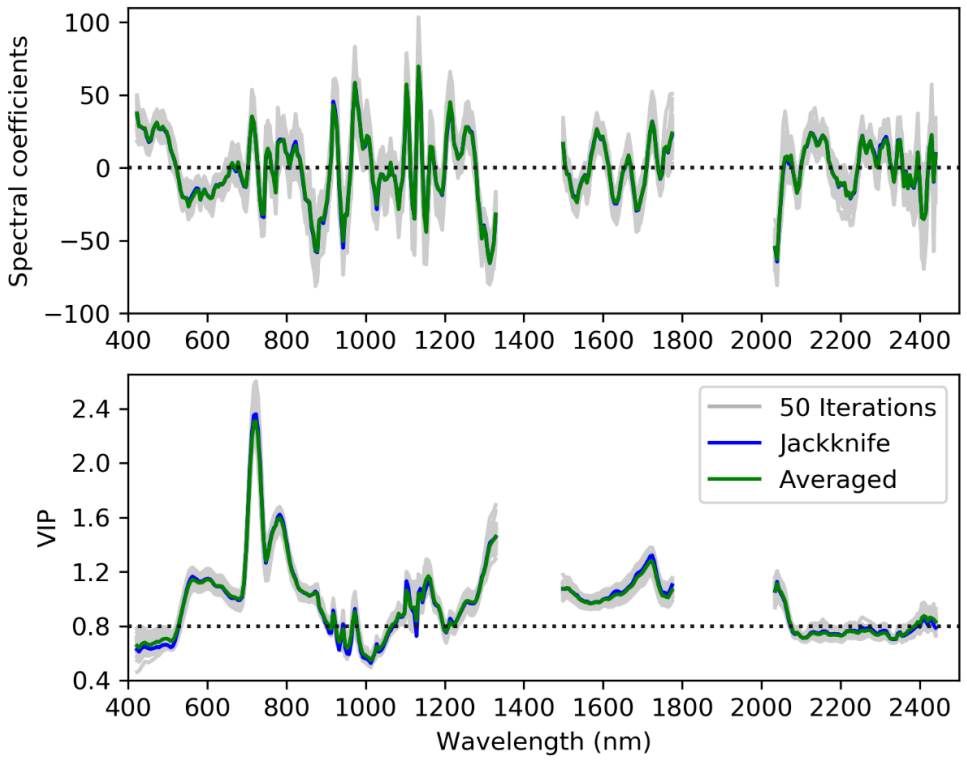

Figure 4. PLSR spectral coefficients and variance influences projection (VIP) of the 50 iterations and their average (i.e., the final model) in the calibration-validation mode, as well as PLSR coefficients generated in the Jackknife mode. Dashed line in the VIP figure indicates a suggested threshold value of 0.8 (Wold et al., 2001). The spectral coefficients of the final model (i.e., averaged) were overlaid onto those of the jackknife model and they were almost identical.

We applied the final model generated in the calibration-validation mode to the GAO data and generated N concentration maps for selected crops (Fig. 6). Note that the maps in the figure didn't capture the full extent of the original VSWIR data. Instead, we mapped spatial subsets where photosynthetic crops dominate the image area. Besides, NDVI and brightness filtered pixels corresponding to non-plant objects and shadows were not mapped. Row crop patterns can be detected for orchard trees, soybean and sweet corn, where white pixels separate different rows in the maps.

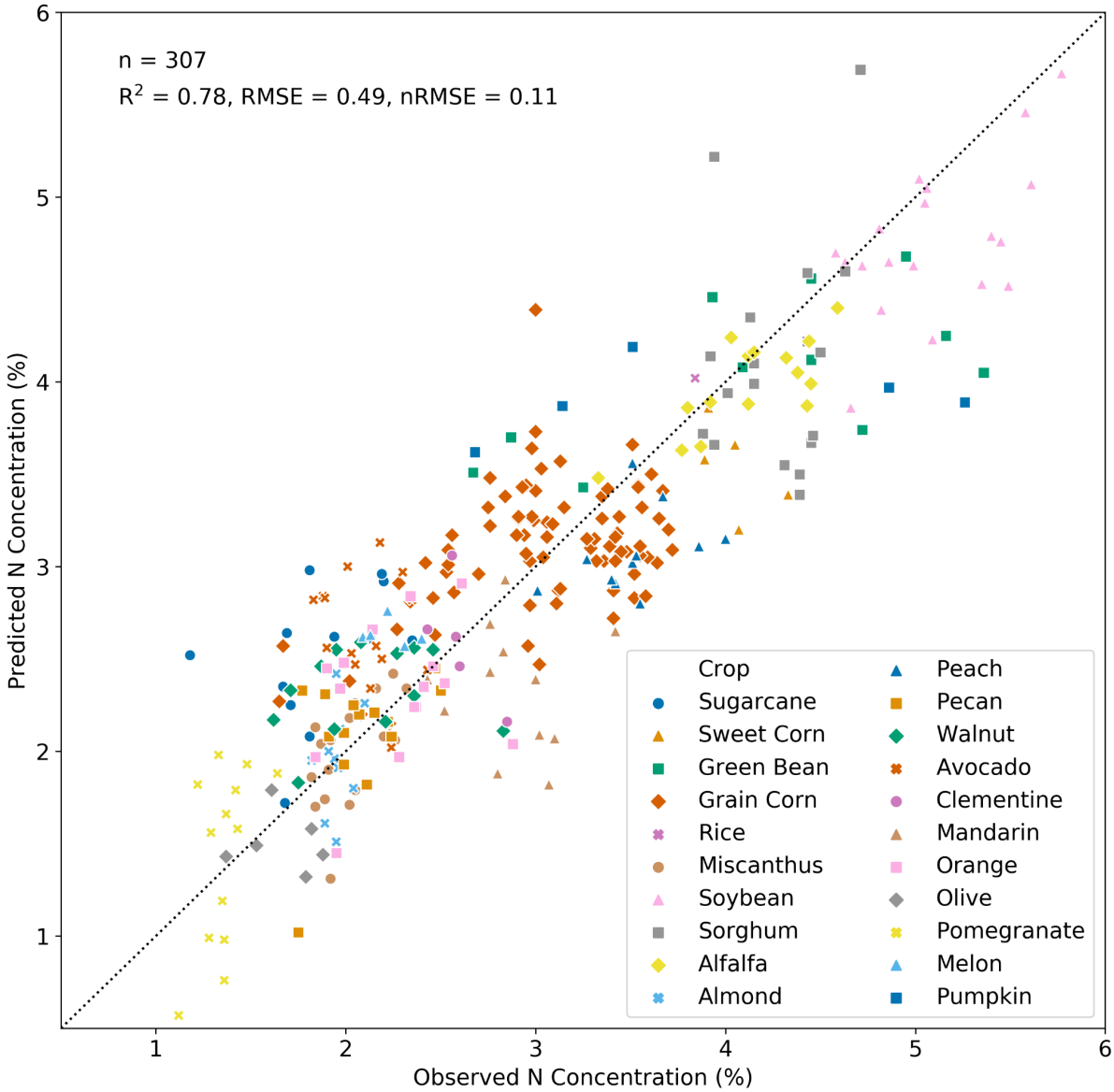

Figure 5. Scatter plot of observed (in the lab) and predicted (through PLSR of imaging spectroscopy data) N concentrations, fitting with the final model in the calibration-validation mode. The dashed gray line is 1:1.

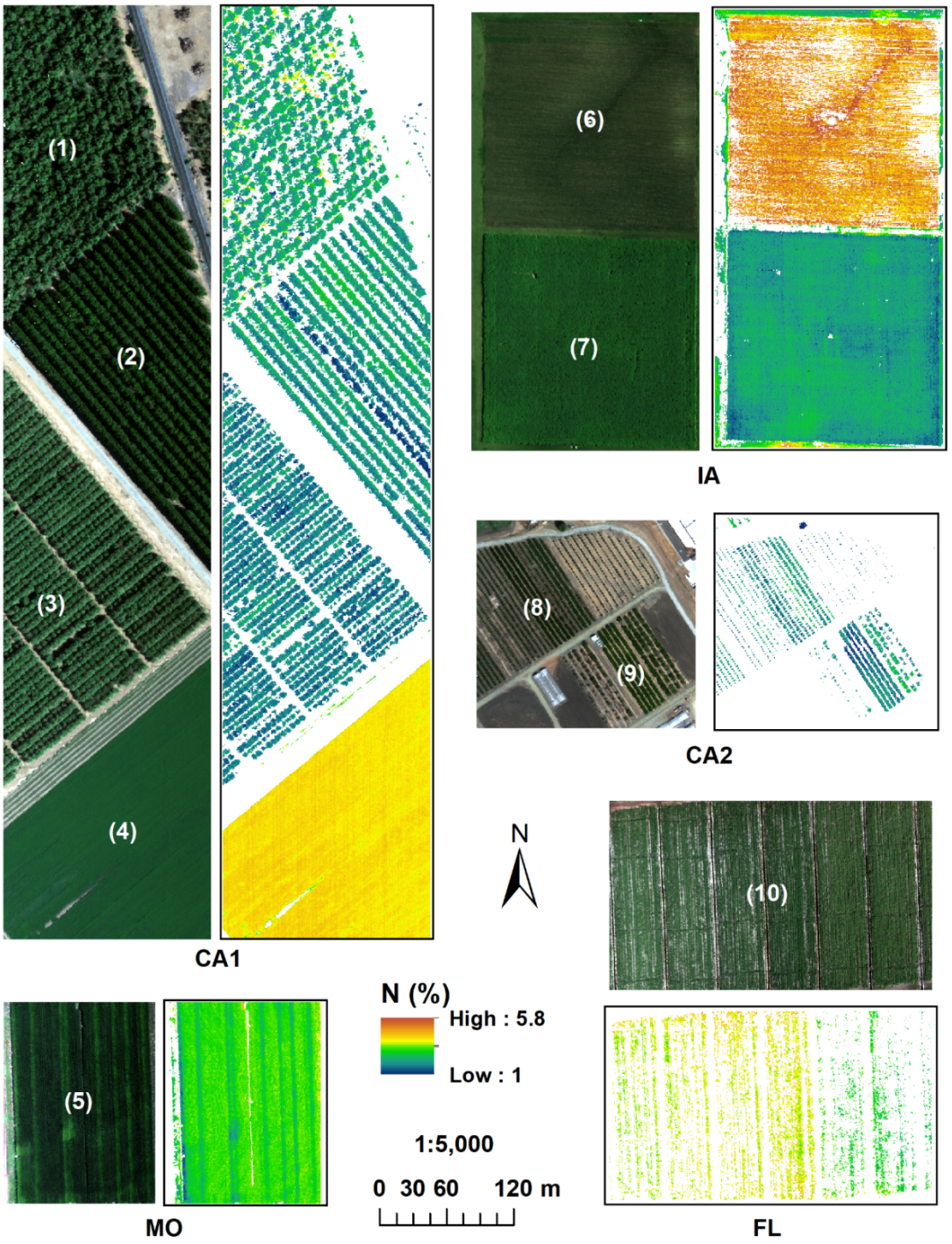

Figure 6. Selected N concentration maps displayed with relative natural color three-band composite GAO data (Red: 657 nm; Green: 567 nm; Blue: 487 nm). Maps were labeled with site ID (specified in Fig.1) and crop species numbers. NDVI and brightness filtered pixels presented no data and were displayed in white. Species list: (1) walnut; (2) pecan; (3) almond; (4) alfalfa; (5) grain corn; (6) soybean; (7) Miscanthus; (8) orange; (9) pomegranate and (10) sweet corn.

The N concentrations of different species corresponded well with laboratory measurements. N fixers (i.e., soybean and alfalfa) and herbaceous crops (i.e., sweet corn and grain corn) showed higher N values than trees. Pomegranate presented the lowest N concentrations in deep blue whereas at the opposite end of the display spectrum, soybean had

the highest N concentrations in brown. In addition, intra-species variations can be detected in the grain corn field at site MO where N fertilization experiments were being conducted (Fig. 6). The blue vertical stripes in the figure correspond to where crops were under-fertilized compared to the other areas showing in green.

Assumptions and limitations

As we mentioned above, we applied a NDVI filter to GAO data to minimize the influences of substrate materials. We assumed that the correspondent spectral profiles of the filtered 164 samples in the field included significant contributions from either NPV or substrate soil. It has been pointed out that empirical methods might lead to unreliable interpretations of N concentration in arid ecosystems, especially when leaf area index (LAI) is low and bright background soil contributes greatly to the total canopy radiation budget (Dashti et al., 2019).

To test the final model on NPV or soil dominated pixels, we divided the 164 N-spectrum pairs into five categories based on NDVI values: [0.1, 0.3), [0.3, 0.4), [0.4, 0.5), [0.5, 0.6) and [0.6, 0.7), with 43, 17, 33, 33, 38 samples in these categories, respectively. Then we applied the spectral coefficients of the final model to the samples in these categories to get performance statistics (Fig. 8). In category [0.6, 0.7), although R^2 was still close to that of category [0.7, 1], RMSE almost doubled, indicating less satisfactory prediction accuracy. As NDVI decreased, the coefficient of correlation (i.e., precision) was still statistically significant but much weaker, and prediction accuracies were far from applicable. We thus don't recommend applying the spectral coefficients in the final model to NPV or soil dominated image pixels.

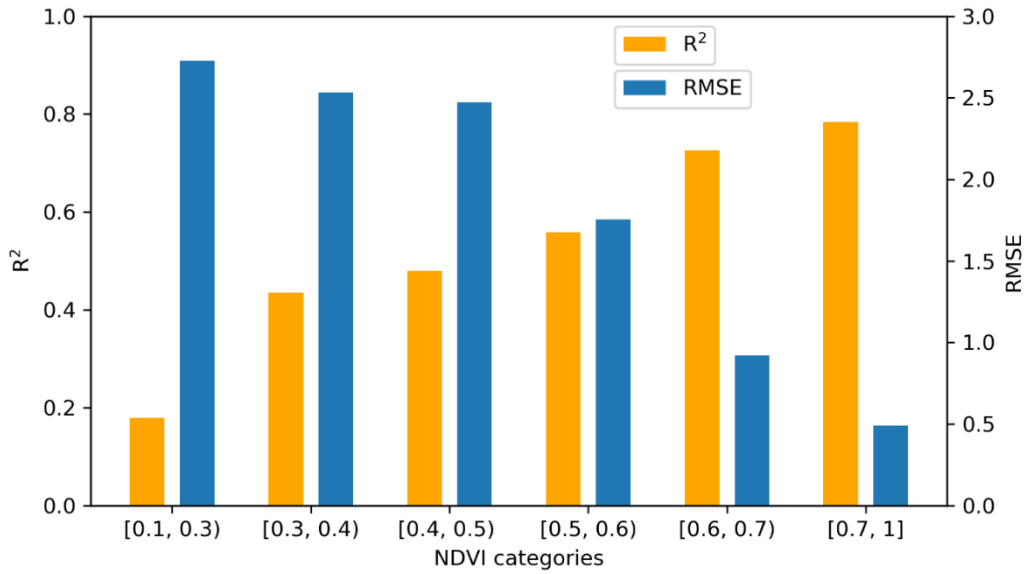

Figure 8. Performance statistics when applying the PLSR coefficients calibrated from green vegetation dominated pixels (i.e., category [0.7, 1]) to pixels in other NDVI categories.

The NDVI and brightness filters efficiently excluded NPV and soil, as well as shadow areas in the image. During field sampling, we also intentionally looked for homogeneous canopies for each specific sample plot and avoided locations that had photosynthetic vegetation other than the target crop. However, the filters applied in our research are not likely to exclude potential contributions from other green vegetation, which can also present high NDVI values in the image. For example, understory grasses and shrubs may be detected if the top canopy of the trees is not dense enough, or in other words, the effective photon penetration depth (EPPD) of the canopy is sufficiently deep to capture understory vegetation (Asner 2008). This might be a larger problem for organic farms, where herbicides are prohibited and weeds tend to coexist with crops. In this case, if the target crop is about the same height with the weeds, neither NDVI, brightness or canopy height filter is likely to exclude weed spectra from the crop spectra, and these contaminated spectra may not correspond well with the model developed here.

Algorithm scalability

We tested the scalability of the algorithm on simulated 30 m Carbon Mapper data. We first resampled the GAO data (0.6 m resolution; Fig. 9 (1)) to 30 m (Fig. 9 (5)) and applied spectral coefficients to the resampled data to get the N concentration map (Fig. 9 (4)). We also resampled the original N map (0.6 m resolution; Fig. 9 (2)) to 30 m (Fig. 9 (3)). We compared both 30 m resolution N maps through pairwise linear regression (Fig. 10). The R^2 of 0.97 and RMSE of 0.20 indicate high similarities between both N maps. However, pixels along crop field edges tend to produce higher discrepancies, as shown in the purple circles in Fig. 10. These pixels captured sufficient photosynthetic vegetation signals to pass the NDVI and brightness filter, but also significant fractions of non-green-vegetation objects that affect the accuracy of the algorithm. Potential solutions are higher NDVI thresholds or more elaborate spectral filtering to ensure that target pixels contain sufficient dominance of photosynthetic vegetation.

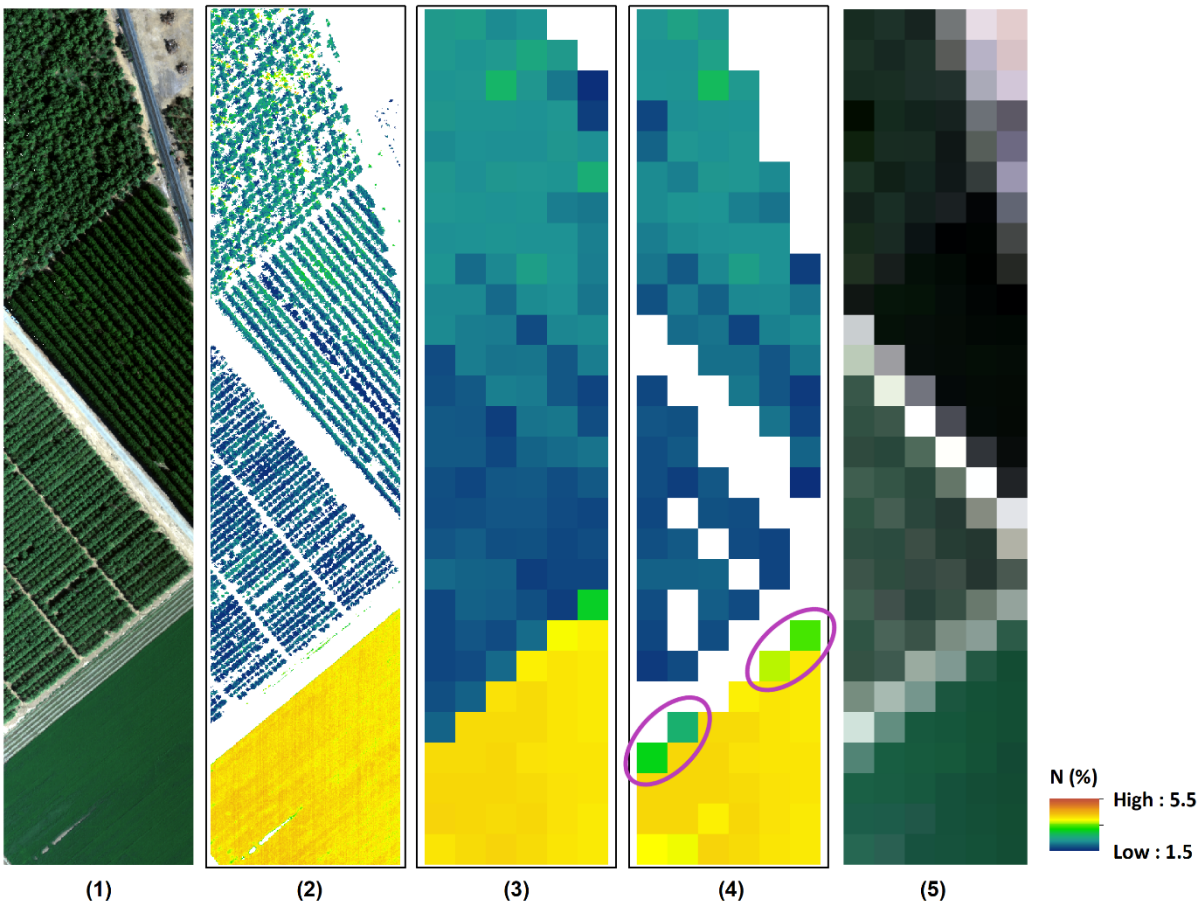

Figure 9. Five maps depicting the same spatial extent of site MO in Fig. 6: (1) natural-color composite of GAO data at 0.6 m resolution; (2) N map at 0.6 m resolution; (3) 30 m resolution N map resampled from 0.6 m resolution N map; (4) 30 m resolution N map by applying spectral coefficients to the resampled GAO data; (5) natural-color composite of resampled 30 m resolution GAO data. Algorithm scalability was tested by comparing maps (3) and (4). Purple circles indicate high discrepancies at crop field edges.

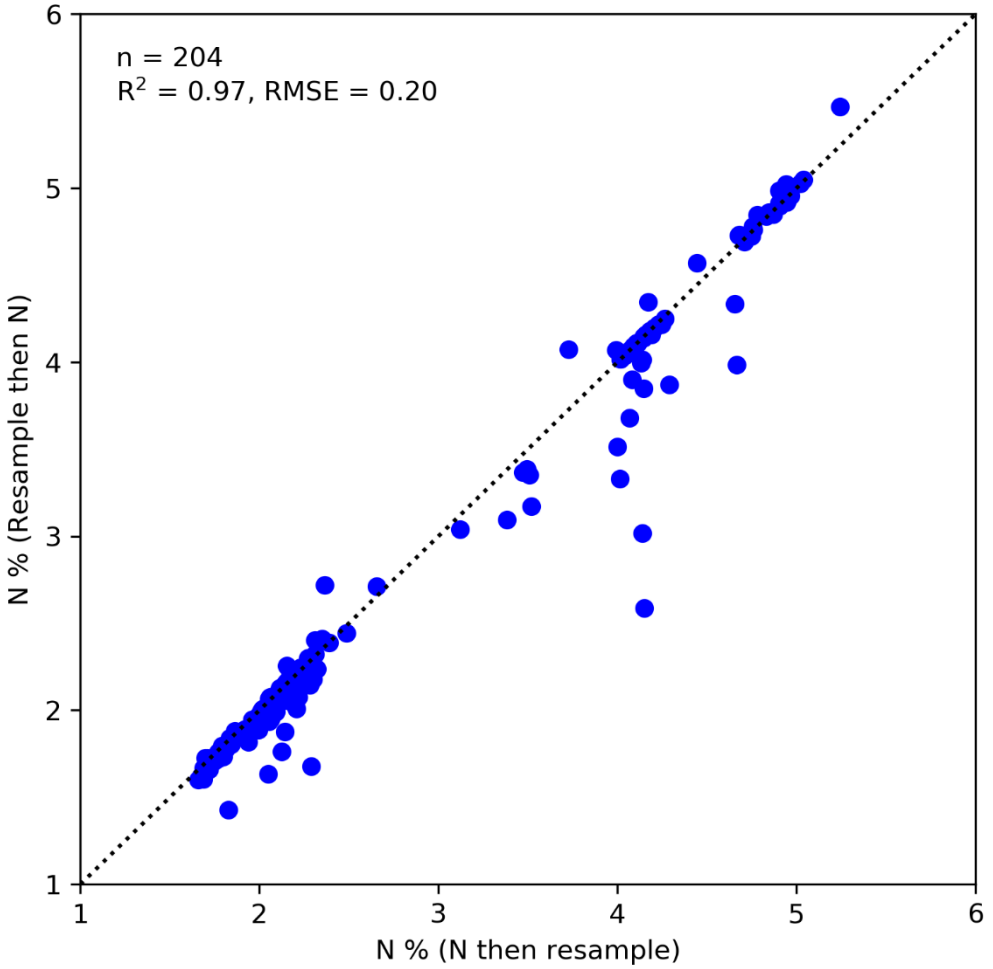

Figure 10. Comparing the similarities between 30 m resolution N map resampled from 0.6 m resolution N maps (i.e., N then resample) and 30 m resolution N map derived from resampled GAO data (i.e., resample then N). The dashed line is 1:1.

4. Codebase

Organization

Code can be found in the GitHub repository at:
https://github.com/CMLandOcean/NitrogenRetrieval

There are two subfolders, code and coefficients. The code folder contains two python scripts, one to fit and rebuild PLSR coefficients (`fit_plsr_coefficients.py`) and one to apply coefficients to a reflectance map (`apply_plsr_coefficients.py`). The coefficients folder contains a CSV database of the coefficients from the fit of PLSR to GAO reflectance data described in this report.

Usage

`apply_plsr_coefficients.py`

To use the apply script is fairly straightforward:

```
apply_plsr_coefficients.py [-h] [-bright_min BRIGHT_MIN] [-bright_max
BRIGHT_MAX] [-ndvi_min NDVI_MIN] [-nodata NODATA] [-format FORMAT] [-co CO]
[-rescale RESCALE] refl_dat_f output_name chem_eq_f
```

The three required positional arguments are:

refl_dat_f	ENVI-formatted reflectance map
output_name	Output file name, should match given -format
chem_eq_f	CSV file containing fields for chem name, transform, intercept, and 214 coefficients

Optional arguments are:

-h, --help	Print help info
-bright_min -bright_max	Thresholds on brightness (vector norm of bands with non-zero PLSR coefficients) to remove input spectra from consideration. Default 1.5 and 9.9, specify -1 to remove.
-rescale	If input reflectance is not scaled 0-1, then default brightness thresholds will not work. This value will be multiplied by input reflectance values.
-ndvi_min	Threshold of NDVI to remove spectra from consideration - default is 0.7, specify -1 to remove.
-nodata	Output map will have this value representing "no data" - defaults to -9999.
-format	GDAL-recognizable shortname for output data format - defaults to "GTiff"
-co	GDAL-recognized creation options for the specified data format

Note that a wavelength/fwhm database is included in the code folder and must be located in the same folder as the apply_plsr_coefficients.py script in order for the script to work.

fit_plsr_coefficients.py

To use the build script is much less straight-forward. This script implements a resampling-based approach to fitting PLSR with minimal tuning. There are multiple ways of dividing the supplied data into training, testing and validation sets. All settings are specified

with a JSON-formatted config file. An optional validation set (called global test set in the script) can be specified (as a proportion using parameter test set holdout) that is never included in the training of the data. Samples not in the validation set are iteratively divided into a training and testing set for each of a specified number of iterations. Division at each iteration can be in one of two modes:

Bootstrap mode The number of iterations is user-specified (using `iterations parameter`) and for each iteration the samples are randomly selected into the two sets with replacement with a proportion specified in `iteration holdout` being placed into the testing set. If data are clustered, a minimal number of samples per cluster in the training set can be specified with config parameter `samples per cluster`.

Jackknife mode The number of iterations is computed based on available sample size and number of samples (or clusters of samples) per iteration defined as a negative integer in config parameter `iteration holdout`. All samples (or clusters) are divided into *n* groups, and each group acts as the test set for one iteration. Thus, each group is tested once and only once against a model built from the other groups.

The user specified a maximum number of components (as `max components` - akin to principal components analysis) to consider in the `PLSRegression` model fit in each iteration. Within an iteration, and for each number of components from `1` to `max components`, the PLSR model is fit using a 10-fold cross validation approach, allowing the computation of the PRESS statistic. The optimal number of components is selected using the number associated with the maximum value of the PRESS statistic. The model is fit again to the full training set using the optimal number of components, and the coefficients from this model are retained and applied to the iteration test set to get an RMSE. After all iterations, a across-iteration median RMSE is computed, and the coefficients for all iterations that had RMSE > $RMSE_{med}$ are averaged by band to get a global coefficient set. These coefficients are applied to the global test / validation set if there is one to get a validation RMSE and R^2.

The command line call looks like this:

```
fit_plsr_coefficients.py [-h] settings_file output_dir
```

Which will apply the settings in a JSON-formatted `settings_file` and write all results to the specified output directory (`output_dir`). The settings file has the following entries:

`csv file` CSV file with all data - should have a row for each input spectra. Spectra can be grouped into clusters if there is a column identifying cluster ID. The spectral data should be column-wise with a column

	for each band in sequential order. Columns names can have a preface, e.g. `B001, B002 ...`, which can be removed to get a band index as an integer.
`band preface`	What precedes numbers in the band columns - usually `"B"`
`chems`	List of chems to fit, e.g. `["LMA"]`
`chem transforms`	Matching list of transforms "log", "sqrt", "square", "inverse" or "none" - like `["none"]`
`cluster col`	Field that contains cluster id (assuming multiple pixels in each cluster). If not cluster-based, use `"none"`
`bad bands`	List of (1-based) band numbers removed before normalization e.g. `[1,2,3,4,211,212,213,214]`
`ignore columns`	List of quoted column names that should be ignored in the analysis, e.g. `["ID","SomeStat"]`
`brightness normalize`	`true/false` - do brightness normalization (Suggested)
`ndvi minimum`	Minimum NDVI value, -1 for no limit
`ndvi maximum`	Minimum NDVI value, -1 for no limit
`ndvi red band`	Band representing red for NDVI computation, e.g. `34`
`ndvi nir band`	Band representing red for NDVI computation, e.g. `46`
`brightness maximum`	Maximum NDVI value, `-1` for no limit
`brightness minimum`	Minimum NDVI value, `-1` for no limit
`iterations`	Number of iterations of the algorithm - for each iteration, the fraction of clusters specified in `"iteration holdout"` will be used to make a test set and the model will be fit on the rest of the data (except that specified in "test set holdout"). Use `-1` to use jackknife mode, i.e. if negative value in `"iteration holdout"`
`iteration holdout`	Fraction of data used for validation at each iteration - can be `0` for no validation set, use negative values (i.e. `-n`) to use jackknife mode, where `n` clusters are held out each time, and each iteration is mutually exclusive.
`test set holdout`	Fraction of data used for global holdout test set - can be `0` for no global holdout set, ignored in jackknife mode.
`samples per cluster`	Specify a minimum number of samples per cluster for training data

	using bootstrap mode, use `-1` for no limit
`max components`	Maximum number of components checked with PRESS stat
`use degen removal`	`true/false` - use procedure to find and remove degenerate (less significant) input features. Similar to the A4 function in autopls. (*Untested and unused for this analysis)
`scale features`	`true/false` - fit scaler to input features (Not suggested, as the scaler would need to be saved to be used for applying fitted coefficients at a later point)
`n jobs`	Number of parallel jobs run by sklearn functions
`random seed`	Seed value for random number generator for reproducible results (`-1` for random seed)

Example configuration files for the bootstrap and jackknife modes used in the fitting of coefficients in this report can be found in the `config` folder in the repository.

5. References

Asner GP. 1998. Biophysical and biochemical sources of variability in canopy reflectance. *Remote Sensing of Environment* 64: 234-253.

Asner GP. 2008. Hyperspectral remote sensing of canopy chemistry, physiology, and biodiversity in tropical rainforests. In *Hyperspectral Remote Sensing of Tropical and Sub-Tropical Forests*.

Asner GP, Martin RE, Anderson CB, Knapp DE. 2015. Quantifying forest canopy traits: Imaging spectroscopy versus field survey. *Remote Sensing of Environment* 158: 15-27.

Asner GP, Martin RE, Keith LM, Heller WP, Hughes MA, Vaughn NR, Hughes RF, Balzotti C. 2018. A spectral mapping signature for the Rapid Ohia Death (ROD) pathogen in Hawaiian forests. *Remote Sensing* 10: 404.

Burnett AC, Anderson J, Davidson KJ, Ely KS, Lamour J, Li Q, Morrison BD, Yang D, Rogers A, Serbin SP. 2021. A best-practice guide to predicting plant traits from leaf-level hyperspectral data using partial least squares regression. *Journal of Experimental Botany* 72: 6175–6189.

Chadwick KD, Asner GP. 2016. Organismic-scale remote sensing of canopy foliar traits in lowland tropical forests. *Remote Sensing* 8: 87.

Chen S, Hong X, Harris CJ, Sharkey PM. 2004. Sparse modeling using orthogonal forest regression with PRESS statistic and regularization. *IEEE Transactions on Systems, Man, and Cybernetics* 34: 898–911.

Dai J, Jamalinia E, Vaughn NR, Martin RE, König M, Hondula KL, Calhoun C, Heckler J, Asner GP. To be submitted. A general methodology for the quantification of crop canopy nitrogen across diverse conditions and species using airborne imaging spectroscopy.

Dashti H, Glenn NF, Ustin S, Mitchell JJ, Qi Y, Ilangakoon NT, Flores AN, Silván-Cárdenas JL, Zhao K, Spaete LP, de Graaff M-A. 2019. Empirical methods for remote sensing of nitrogen in drylands may lead to unreliable interpretation of ecosystem function. *IEEE Transactions on Geoscience and Remote Sensing* 57: 3993-4004.

Feilhauer H, Asner GP, Martin RE, Schmidtlein S. 2010. Brightness-normalized partial least squares regression for hyperspectral data. *Journal of Quantitative Spectroscopy and Radiative Transfer* 111: 1947–1957.

Martin RE, Chadwick KD, Brodrick PG, Carranza-Jimenez L, Vaughn NR, Asner GP. 2018. An approach for foliar trait retrieval from airborne imaging spectroscopy of tropical forests. *Remote Sensing* 10: 199.

United States Department of Agriculture. 2022. ARMS III Farm Production Expenditure Regions. Retrieved from https://www.nass.usda.gov/Charts_and_Maps/Farm_Production_Expenditures/reg_map_c.php on Feb. 10, 2023.

Wold S, Sjöström M, Eriksson L. 2001. PLS-Regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58: 109–130.