Analytica Chimica Acta 1051 (2019) 32–40

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

# Sparse NIR optimization method (SNIRO) to quantify analyte composition with visible (VIS)/near infrared (NIR) spectroscopy (350 nm-2500 nm)

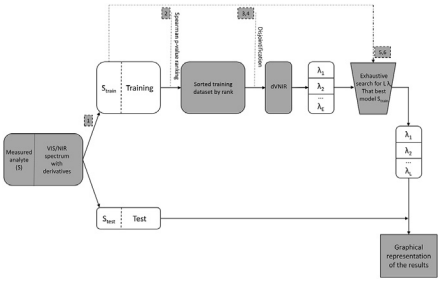
Yonatan Peleg [a, b], Shai Shefer [a], Leon Anavy [c], Alexandra Chudnovsky [a], Alvaro Israel [d], Alexander Golberg [a, **], Zohar Yakhini [b, c, *]

[a] School of Environment and Earth Sciences, Tel Aviv University, Israel
[b] School of Computer Science, IDC Herzliya, Israel
[c] Department of Computer Science, Technion Israel Institute of Technology, Haifa, Israel
[d] Israel Oceanographic and Limnological Research, The National Institute of Oceanography, Haifa, Israel

## HIGHLIGHTS

- A Sparse NIR Optimization method (SNIRO) for selecting a given number of significant wavelengths from spectra was developed.
- The computed complexity time and the accuracy of SNIRO was compared to Marten's test, to forward selection test and to LASSO.
- SNIRO was used to determine protein content in corn flour and meat, and octane number in diesel using public NIR datasets.
- SNIRO was used to determine the glucose content in the green seaweed *Ulva* sp.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Article history:*
Received 12 August 2018
Received in revised form
14 November 2018
Accepted 21 November 2018
Available online 23 November 2018

*Keywords:*
VIS/NIR spectroscopy
Chemometrics
Seaweeds
*Ulva* sp.
Multivariate analysis
Imaging
Sparse linear regression
Diesel octane number

## ABSTRACT

Visual-Near-Infra-Red (VIS/NIR) spectroscopy has led the revolution in high-throughput phenotyping methods used to determine chemical and structural elements of organic materials. In the current state of the art, spectrophotometers used for imaging techniques are either very expensive or too large to be used as a field-operable device. In this study we developed a Sparse NIR Optimization method (SNIRO) that selects a pre-determined number of wavelengths that enable quantification of analytes in a given sample using linear regression. We compared the computed complexity time and the accuracy of SNIRO to Marten's test, to forward selection test and to LASSO all applied to the determination of protein content in corn flour and meat and octane number in diesel using publicly available datasets. In addition, for the first time, we determined the glucose content in the green seaweed *Ulva* sp., an important feedstock for marine biorefinery. The SNIRO approach can be used as a first step in designing a spectrophotometer that can scan a small number of specific spectral regions, thus decreasing, potentially, production costs and scanner size and enabling the development of field-operable devices for content analysis of complex organic materials.



* Corresponding author. School of Computer Science, IDC Herzliya, Israel.
** Corresponding author.
*E-mail addresses:* agolberg@tauex.tau.ac.il (A. Golberg), zohar.yakhini@idc.ac.il (Z. Yakhini).

https://doi.org/10.1016/j.aca.2018.11.038
0003-2670/

## 1. Introduction

Near-Infra-Red (NIR) and Visual-NIR (VIS/NIR) spectroscopy, in the 350–2500 nm spectral range, is a widely used method for analyzing compounds [1,2]. NIR and other imaging techniques are currently used for several important applications such plant phenotyping, geological applications, food industry, and agriculture [3]. VIS/NIR spectroscopy can be performed in two modes: whole spectrum and discrete wavelength selection [4]. Discrete wavelength spectrophotometers have the advantage of being simple to use and relatively cheap to develop, in comparison to whole spectrum devices, due to the low cost of photodiodes and narrow band light emitting diodes (LEDs) [5,6]. Furthermore, their design can easily be miniaturized and packaged [7] to support robust and efficient field work [8]. Wavelengths (λs) can be selected either by using filters that screen for narrow bands or by using LEDs that directly produce narrow bands [5,9]. An adverse result of analyzing specific wavelengths is that their application is reduced to analytes that absorb in the distinct selected spectral zones, whereas whole spectrum instruments are applicable to a broad range of analytes [10]. A major advantage of VIS/NIR spectroscopy over standard chemical analysis methods is the speed of analysis. Therefore, many applications of VIS/NIR can benefit from portable devices that can be deployed in multiple locations in the field [5]. This demand emphasizes the importance of developing low-cost, but precise, devices. Such devices will be based on preselecting the application specific wavelengths [7].

Complex samples chemistry analysis with low-cost portable devices and multivariate analytics requires new methods for rapid selection of informative wavelengths from the whole VIS/NIR spectrum read [11,12]. One approach to extract the analytical information embodied in the VIS/NIR spectra is based on a wide range of multivariate analysis methods that relate specific variables (in this case VIS/NIR spectrum components) to sample properties, for example sugar or protein concentration [4,13–17]. Multivariate analysis methods include 2D correlation plots [18], partial least-squares regression (PLSR) [19], principal component analysis (PCA) [20], support vector machines [21], neural networks [22] and other machine learning approaches.

Most methods that address wavelengths selection often do so by using indirect statistical approaches. For example, the Marten's method is based on the standard deviation of the regression coefficients calculated from the cycles of leave-one-out cross validation [13,14]. Coefficients are rejected based on the normal cumulative distribution function and whether they fall within a pre-determined boundary. Another approach for feature selection is the PLSR method. Studies have been done on fruit juice [23], tomatoes [24] and wheat straw [25] to predict sugar and salt [26] concentration using PLS, and the models produced vary according to the number of factors chosen. Another approach is using slopes across different spectral ranges as indicators of change in chemical constitutes, combined with PLS analyses as done for fresh and dry vegetation pasture, to determine protein content [27]. Other studies conducted on grapes [28] used principal component analysis (PCA) to reduce the number of variables, and multiple linear regression (MLR) for the variable selection. These PCA based approaches, while seeking to reduce the number of explanatory variables, work with rotated and combined dimensions. The resulting low dimensional explanatory vectors, therefore, cannot be directly measured by a discrete VIS/NIR device of the corresponding wavelength multiplicity.

One shortcoming of the discussed above methods is that they only afford indirect control over the number of selected λs. In the context of enabling field-operable measurement devices it is necessary to control the number of λs upon which to base the measurement. The goal of the method presented in this study is to optimize the selection of a pre-determined number of significant λs, with respect to the information that can be inferred about a given analyte composition in a given type of sample. That is – to develop an optimization process that takes as input a target analyte type, A (e.g glucose), and a target sample type, T (e.g corn), as well as training whole spectrum data from multiple samples, and produces an efficient discrete spectra approach for measuring A in samples of type T. The Sparse NIR Optimization (SNIRO) process was developed to serve as a first step in designing and potentially constructing field-operable discrete wavelength spectrophotometers.

The framework developed herein, SNIRO, takes as input the desired number of wavelengths, L, as determined, e.g, by engineering device considerations. It then seeks an optimal combination of L wavelengths by using sparse linear regression. SNIRO is deployed over a distributed computing platform (Azure, Microsoft, WA).

Sparse linear regression seeks linear models that use only a small number of explaining variables. In general, finding the sparsest solutions to an underdetermined linear systems is NP hard [29]. StOMP and other related techniques [30] heuristically address a related task – find the sparsest near solution to an underdetermined system. The Westad-Martens uncertainty test (MUT) [13] is also designed to select an adequate number of explaining variables in a given system, but provides no direct control of L, as above. Note that the task addressed in this paper, in the context of VIS/NIR and inference of analyte levels, is finding the best approximate solution of an underdetermined system, using a predetermined number of columns (L). While this task is strongly related to sparse exact solutions or to solutions with fixed approximation bounds [31], it is not the same. The current work is driven by a fixed constraint on the number of parameters expected to be affordable for a field operable device. Solutions produced by StOMP or Westad-Martnes, while intrinsically more efficient, are not necessarily useful in this context, as they are driven by accuracy and may produce solutions with too many non-zero coefficients, implementation of which is hardly possible in low-cost devices.

The SNIRO approach uses one of two processes to select the wavelengths. The first starts with a correlation based disjointification to reduce the set of wavelengths to be considered and then exhaustively searches the best subset there. The other is a forward selection approach tailored to the task at hand. To test and validate SNIRO the process was first applied to several public datasets including protein content in corn [32] and meat [33], and diesel octane number [34]. Furthermore, we used SNIRO to model glucose content in *Ulva* sp. macroalgae, using data that was specially produced for this work. For diesel octane number we demonstrate a measurement based on 5 VIS-NIR wavelengths that has a Spearman correlation of 0.96 to the actual octane number ($p$-value $< 0.001$), on a test set of 80 samples.

## 2. Methods and data

Data was obtained from various sources (see Methods – data). The raw data was organized in a matrix in which each row represents a sample and each column represents a wavelength measurement or an inferred derivative. Let NIR represent the mXn matrix containing all the absorbance spectra and let S be the target data, or the response vector, with every entry corresponding to a single sample (Fig. 1). Therefore, every column of the input matrix VIS/NIR (Fig. 1) represents the spectral absorbance measurement results at one of the wavelengths used or an inferred derivative (we index columns by λ). Table 1 contains an explanation for each of the variables used in the method. We provide code that implements

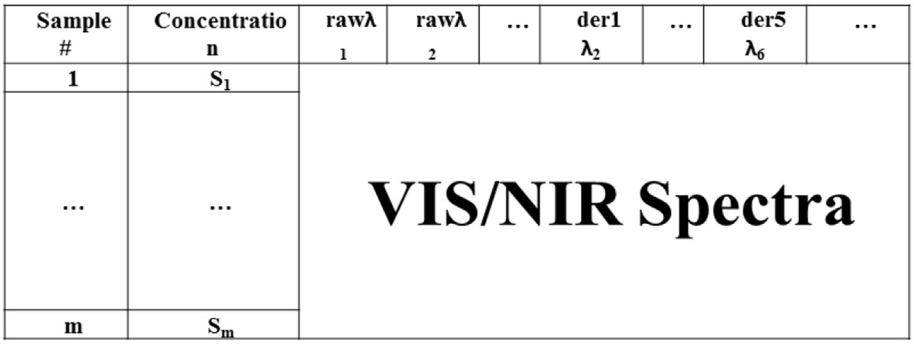



**Fig. 1.** m × n matrix representation of the VIS/NIR spectra (measured in samples of type T) and the response vector S (the analyte A measured in the same samples). These represent the input to the learning process.

**Table 1**
Variables used in SNIRO.

| Variable name | Symbol | Explanation |
|---|---|---|
| Analyte Type | A | The type of analyte targeted by VIS/NIR spectra model. For example, glucose. |
| Target Sample | T | The product containing the analyte. For example, corn flour. |
| Wavelength | λ | A wavelength or inferred derivative. |
| Number of wavelengths | n | The number of λs in the full spectra. |
| VIS/NIR Spectra | VNIR | A matrix with n columns (λs) and m rows (samples). |
| Analyte Concentration | S | The target data trying to model. A vector of length m. |
| Training Length | m | The number of samples in the training dataset. |
| Spearman P Value | $p(\lambda, S)$ | The Spearman P value of a wavelength with the analyte concentration. |
| Disjointification λs | E | The number of λs for the disjointification process to allow through. |
| VIS/NIR disjointified matrix | dVNIR | A matrix with E columns and m rows. |
| Solution | C | The sparse solution vector to the underdetermined sparsity constrained system: dVNIR*C=S |
| Solution Length | L | The length of the solution vector, C. The number of wavelengths desired for the solution. |
| Set of changing λs | Λ | A set of L λs. Thus, a matrix of dimension m × L. |
| Best set of λs | $\Lambda^*$ | The best set of L λs which minimize the TRE. |
| Predicted Values | PV | A vector of length m of predicted values, a putative prediction of S. |
| Pearson Correlation | $\tau(\lambda_1, \lambda_2)$ | The Pearson Correlation between two different λs. |
| Shuffled Analyte (random response) | $S_{shuffle}$ | A shuffled analyte vector. |
| Best random set of λs for a random response | $\Lambda^*_{shuffle}$ | The best set of L λs produced using a randomized S vector which minimize the TRE. |
| Random Solution for a random response | $C^*_{shuffle}$ | The sparse solution vector to the underdetermined system: dVIS-NIR*C=$S_{shuffle}$ |

SNIRO at https://github.com/yakhinigroup/sniro.git.

### 2.1. SNIRO – selection of the best wavelengths

SNIRO infers a small set of λs to form the basis of a linear model for the analyte content quantification in a given sample type. The SNIRO framework includes a training/test-data approach to support model validation. Model results can be translated into a measurement architecture based on a small number of VIS/NIR wavelengths. SNIRO can be applied to any sample type of interest and will lead to feasibility assessment and to the potential development of an efficient VIS/NIR device that allows analyte quantification in that context, based on the inferred wavelengths that may obviously depend on the specific analyte context. The main steps of SNIRO are as follows (Fig. 2):

1. Divide the data into a training dataset (of length m) and test dataset at a ratio of 4:1. Choose samples for the test dataset which best represent the distribution of values in the measured target data, S. In the implemented software, the data were split into test and training sets by using the R function 'sample.split' which preserves relative ratios of different labels/values in S.
2. Compute Spearman p-values, $p(\lambda, S)$, for all λs in the initial training data. This step is performed on a concatenated dataset of the derivatives up to the 5th derivative (Fig. 1). Note that higher derivatives require larger λ windows which may affect the measuring device. This step is similar to StOMP's Matched Filter step [35]. While StOMP uses the filter step to select columns for the next iteration, we use it to set the starting point of our search based on the computational power available to the procedure. In this work, Derivatives were calculated using Savitzky-Golay method [36].
3. Determine a desired maximum number of λs to use in the next steps of the calculations, denoted by E (depends on the expected actual computing time and on the available processing power). Note that λs can either be raw VIS/NIR measurement columns or derivatives of different orders.
4. Perform a disjointification process to obtain a nearly orthogonal linear system [m × E] for the training data. Thus at the end of this step we have an m × E under-determined system to be solved: $dVNIR \cdot C = S$ (see pseudo code in supplementary). Disjointification is a heuristic iterative orthogonalization process designed to find a small (size controllable by the process) set of co-ordinates/variables/measurements that are maximally orthogonal. Disjointification results in a set of co-ordinates (in our case wavelengths) that is a strict sub-set of the original co-ordinates/measurements. This is a very important distinction from classical methods such as PCA and SVD. While PCA is much more efficient in finding a low dimensional representation of the data it results in co-ordinates (or axes) that are linear combinations of the original measurements. The same is true for SVD. In many applications, such linear combination cannot be directly measured. Since we are interested in a practicable set of

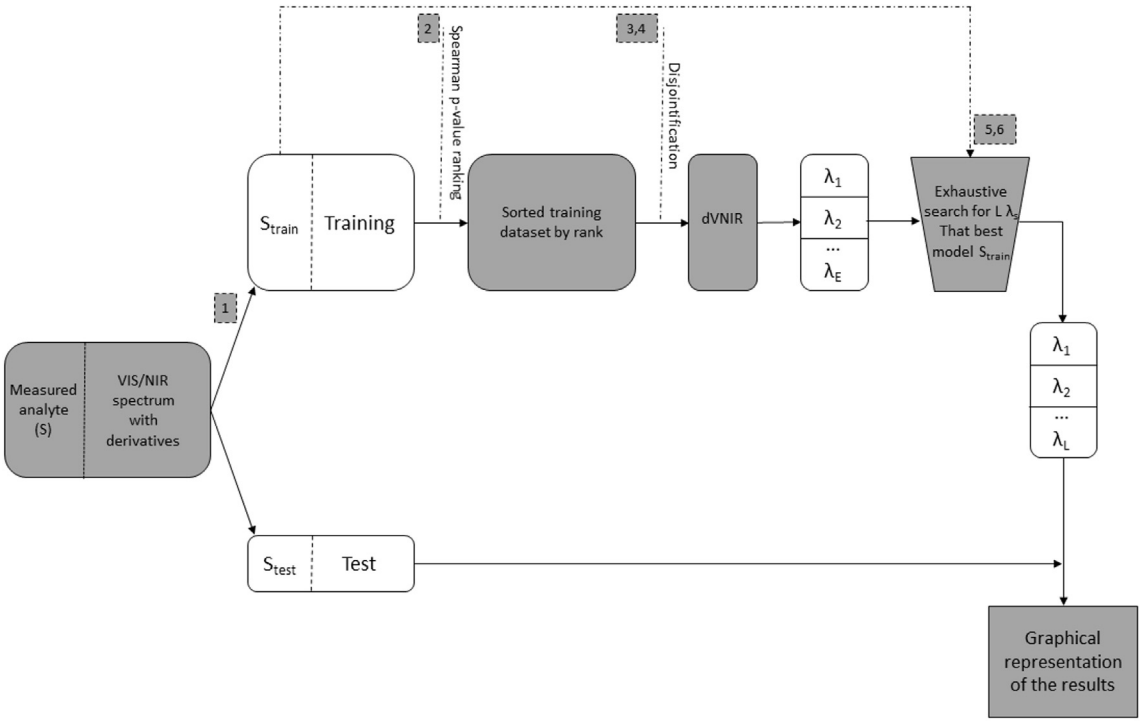



**Fig. 2.** Flow chart for the entire SNIRO method for L<<E.

wavelengths (and/or derivatives) we must work with the original co-ordinates. Disjointification yileds, therefore, a less efficient but fully practicable set of variables to represent the data.

5. We seek a sparse solution C, where the number of non-zero values of C are dependent on the number of λs in the desired solution/device (the sparse size of the solution vector C is denoted by L). To solve the system perform an exhaustive search and generate $\binom{E}{L}$ over-determined systems, $\Lambda \cdot C = S$, each of which uses L columns of dVNIR. Each such over-determined system is $m \times L$. Solve each one of these using a pseudo inverse procedure. L depends on processing time and power as well as the envisioned final measurement process.
6. The λs that yield the minimal Total Relative Error (TRE), denoted $\Lambda^*$, together with the associated solution $C^*$, are selected as the results of the process. The vector of predicted values (PV) for the analytes of interest is given by: $PV = \Lambda^* \cdot C^*$.

TRE for PV and the actual measured value for the analyte (denote by S) is defined as follows:

$$TRE = \frac{100}{m}\sqrt{\sum_{i=1}^{m}\left(\frac{S_i - PV_i}{PV_i}\right)^2} \quad (1)$$

7. Report results on the test dataset and validate against random controls (see comment below for details). To visualize the validation, produce an $R^2$ fit such as in Fig. 3a showing measured against predicted values and reporting: TRE, Spearman correlation, $R^2$, best λs and C vector. $R^2$ is calculated by the following formula:

$$R^2 = 1 - \frac{\sum (S_i - PV_i)^2}{\sum\left(S_i - \bar{S}\right)} \quad (2)$$

RMSE was calculated to compare SNIRO results to other literature results using the formula:

$$RMSE = \sqrt{1 - r^2}SD_y \quad (3)$$

Where $SD_y$ is the standard deviation of the actual values.

#### 2.1.1. Comments on permutation testing and controls

To control against over-fitting, shuffle the vector S from Step 2 and perform the entire process again. Use the resulted $\Lambda^*_{shuffle}$ and $C^*_{shuffle}$ on the test dataset. We expect to see a much higher TRE and lower Spearman correlation on the test dataset when comparing using the λs produced from the shuffled process $\Lambda^*_{shuffle}$ to those produced from the real data $\Lambda^*$. Such a difference indicates that the primary results were not random. Produce a density histogram as seen in Fig. 3b showing the distribution of 1000 TRE calculations of a randomized S vector against the actual S vector. This is required to determine the quality of approximation expected from shuffling the actual values in S. This quality will differ depending on the variance of the analyte data. When the number of wavelengths considered gets too large the selection process may over-fit the training data. In this case, it will approximate a randomly shuffled analyte vector. In addition, analyte data with low variance may also be well approximated by random shuffles of itself. Fig. 3b exemplifies benchmarking against random controls that addresses these two issues. The histogram represents TREs for shuffles of S. The red star represents the results of the entire learning process when applied to shuffled data.

### 2.2. Data

#### 2.2.1. Overview of datasets

Four different datasets that vary in sample size and spectral range were used in the study, all on which SNIRO and other methods were applied for comparison. The datasets for diesel, corn

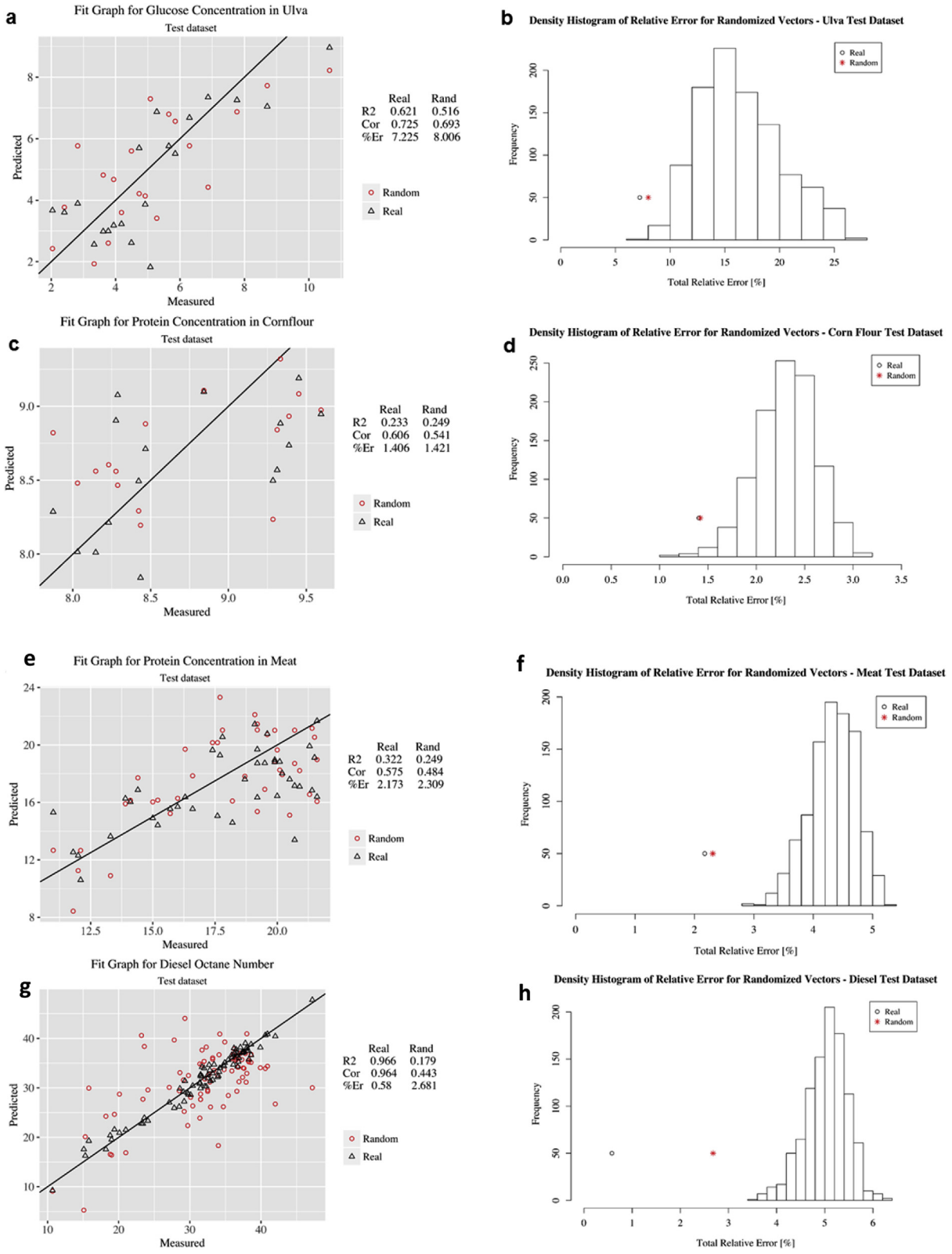



**Fig. 3.** Test results and comparison of test results to the randomized process for the glucose concentration in the *Ulva* sp. dataset (**a** and **b**), protein concentration in the corn flour (**c** and **d**), meat (**e** and **f**) and diesel octane number (**g** and **h**).

flour and meat are open source and available online. The dataset for *Ulva* sp. glucose content was specially produced for this work, as detailed below, and is available for use upon request. Data used for the determination of protein concentration in Meat was recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850 – 1050 nm by the Near Infrared Transmission (NIT) principle. Each sample contains finely chopped pure meat with different moisture, fat and protein contents. Table 2 details the sample size and spectral range of each dataset.

#### 2.2.2. Macroalgae Ulva fasciata L. glucose measurement with high pressure ion chromatography and VIS/NIR spectrometry

*Ulva fasciata* L. is a green marine macroalga of worldwide distribution found in the intertidal and shallow waters within the Israeli Mediterranean shores *Ulva* sp. are of particular interest as a feedstock for biorefineries [37–42] because of their high growth rates and fermentable carbohydrate content [39,43–45]. Significant effort is required to develop and select macroalgal species and strains with specific properties tailored for food, chemicals or fuel applications [46]. VIS/NIR spectrometry could enable rapid selection of strains with the required chemical composition, for example high glucose content for biofuel fermentation [47]. The sufficient goal for this early stage of this approach development is to differentiate high and low glucose samples.

For the current study, specimens were taken from stocks maintained at a seaweed collection at Israel Oceanographic & Limnological Research, Haifa, Israel (IOLR). Cultivation trials were conducted in an outdoor setting at IOLR. Single 5 $cm^2$ pieces were placed in 15 × 15 cm plastic net baskets. A total of 100 baskets, with a single thallus per basket, were divided into 10 groups (5 groups of nutrients with two complete replicates), 10 baskets each, and tied to 40-l fiberglass tanks supplied with running seawater and aeration. With nutrients application (Supplementary Information Table S2), the water exchange was stopped for 24 h to allow the absorption. Total cultivation time for all 4 groups was 4 weeks from 8 to 29 November 2015.

Following the growth period in the outdoor tanks, each thallus was dried separately at 60 °C for 48 h until constant weight and ground into powder manually in a mortar with liquid nitrogen. For hydrolysis, 0.05 g of *Ulva* sp. powder was mixed with 2 ml sulphuric acid (5%) in 10-ml plastic tubes and autoclaved at 120 °C for 45 min. Next, 500 mM of a phosphate buffer (Sigma, Israel) was added into the above mix, and the resulting hydrolysate was neutralized with 3 M sodium hydroxide (NaOH) to pH 7. Dionex ICS-5000 (Thermo Fischer Scientific, CA) was used to quantify glucose in the hydrolysates. Carbopac MA1 (Thermo Fischer Scientific, MA) and its corresponding guard column were used for separation. An electrochemical detector with AgCl as reference electrode was used for detection. A trinary solvent system was used for elution (Table S1, Supplementary information). The column temperature was kept at 30 °C and flow rate 0.25 ml $min^{-1}$. Calibration curves were made for glucose to determine its concentration in hydrolysates.

For spectral analysis, dried thalli were scanned by Fieldspec Analytical Spectral Devices (ASD) Full-Range (FR) spectrometer (Analytical Spectral Devices, Boulder, CO, USA) at three locations on each thallus [61]. The FR spectrometer samples a spectral range of 350–2500 nm (http://www.asdi.com). The instrument uses three detectors spanning the visible and near infrared (VNIR, comprising a Si photodiode array) and shortwave infrared (SWIR1 and SWIR2, comprising two separate InGaAs photodiodes). All samples were measured in the laboratory by attaching the High Intensity ASD Contact Probe ('potato') device to the sample and extracting an average of 40 readings, using bare fiber and self-probed illumination. The "potato" was set on a stable tripod base and maintained in a constant position at a nadir-looking angle. For all measurements, we used a Spectralon standard white reference panel (Spectralon, Labsphere Inc. www.labsphere.com) in the same geometry as a white reference to enable conversion of the measurement data into reflectance values (see Excel Table in supplementary for spectral data).

**Table 2**
Description of datasets used.

| Dataset | Samples | Spectral range [nm] |
|---|---|---|
| Diesel - Octane Number [34] | 395 | 948–1550 |
| Corn Flour –Protein [32] | 80 | 1132–2498 |
| Meat – Protein [33] | 215 | 850–1050 |
| *Ulva* sp. - Glucose | 100 | 250–2500 |

### 2.3. Distributed computing

The exhaustive processes in SNIRO required significant computational time and resources. Thus, calculations were conducted on several platforms using parallel computing. Code was written on R, a language and environment for statistical computing and graphics (https://www.r-project.org). Configurations used for performing calculations include:

- 16 GB, 8 Core Microsoft Windows lab computer
- 8 GB, 4 Core Macbook Pro personal laptop
- 2 Microsoft Azure 32 Core Virtual Desktops operated via the Microsoft Azure service: https://azure.microsoft.com/en-us/
- 16 Core Server at the Technion, Haifa

## 3. Results

Results for using SNIRO on the datasets, in the description below, are divided into calibration and prediction datasets. We had chosen to implement SNIRO with the parameters L and E equal to 5 and 100 respectively based on data analysis performed on the Ulva dataset. More information regarding the data analysis can be found in the supplementary material. Coefficients calculated for the model $VNIR \cdot C = S$ along with the wavelengths found are reported using a notation scheme in which, for example, 350d3 indicates the 3rd derivative at wavelength 350. If the raw data is used, the wavelength will appear without a derivative. Table 3 displays the results for the four datasets used in this study. $R^2$ and random distribution graphs for the corn flour, meat and diesel datasets can be found in the supplementary.

### 3.1. Determination of glucose in Ulva sp.

Dataset produced for this work includes VIS/NIR spectrum for 100 samples and glucose composition for each one. Prediction of glucose content in *Ulva* sp/is paramount for the development of algae-based bio-refineries as it enables faster screening based on a simple form of measurement. Fig. 3a and b displays the results of the test dataset using SNIRO to predict glucose concentration in *Ulva* sp.

### 3.2. Determination of protein in corn flour

Prediction of nutritional values of corn is important due to the ever-growing demand for the product in many different industries such as food, energy and livestock. Thus NIR has been used as a fast and accurate method for determining nutritional value of corn [48]. To date, many methods used to predict corn analytes include PLS and PCR [49]. Fig. 3c and d displays the results of the test dataset using SNIRO to predict protein concentration in corn flour.

**Table 3**
Summary of SNIRO analysis on various datasets. $R^2$ and Total Relative Error (TRE) of regression models are shown for the Calibration (training) and Prediction (test) datasets. Wavelengths ($\Lambda^*$): the best set of 5 λs which minimize the TRE. Coefficients: coefficients for each of the best 5 λs for the linear regression model. Samples from the prediction group were not included in the calibration. N is the number of samples.

| Ulva sp. glucose | Calibration (N = 80) | | Prediction (N = 20) | | | Corn protein | Calibration (N = 64) | | Prediction (N = 16) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | TRE [%] | $R^2$ | TRE [%] | | | $R^2$ | TRE [%] | $R^2$ | | TRE [%] |
| | 0.61 | 2.48 | 0.62 | 7.23 | | | 0.76 | 0.34 | 0.23 | | 1.41 |
| Wavelengths | 528d1 | 519d2 | 2287d4 | 537d1 | 407d2 | Wavelengths | 1760d1 | 2262d2 | 2034d2 | 1356d4 | 2468d1 |
| Coefficients | 3318 | −13674 | 16695 | 2052 | 1061 | Coefficients | −12570 | −7899 | 22948 | 36405 | 2180 |
| Meat protein | Calibration (N = 174) | | Prediction (N = 41) | | | Diesel octane number | Calibration (N = 315) | | Prediction (N = 80) | | |
| | $R^2$ | TRE [%] | $R^2$ | TRE [%] | | | $R^2$ | TRE [%] | $R^2$ | | TRE [%] |
| | 0.56 | 0.90 | 0.32 | 2.17 | | | 0.96 | 0.28 | 0.97 | | 0.58 |
| Wavelengths | 904d1 | 928d1 | 958d3 | 956d4 | 1046 | Wavelengths | 992d1 | 1018d1 | 1266d1 | 1016d2 | 1022d2 |
| Coefficients | −743 | −1129 | −5178 | 6081 | 9 | Coefficients | 3518 | −2600 | 1282 | 6449 | 9977 |

### 3.3. Determination of protein in meat

NIR spectroscopy has been successfully applied to the quantitative determination of major constituents (moisture, fat and protein) in meat and meat products [50]. Statistical methods used to model these constituents include multiple linear regression, partial and modified partial least square (PLS), principal components (PCR) and also techniques that allow for non-linear relationships such as neural networks [51]. Fig. 3e and f displays the results of the test dataset using SNIRO to predict protein concentration in meat.

### 3.4. Determination of octane umber in diesel

Diesel Octane Number (ON) and other fuel specifications dictate several attributes necessary for operation in vehicles. Determining the ON of fuels using a Cooperative Fuels Research (CFR) engine costs over $200,000, requires trained personnel to operate and takes 20 min [52]. In an effort to reduce testing costs, researchers sought out more cost-effective and faster noninvasive optical techniques for determining ON, among other fuel specifications, by way of statistical analysis. Vibrational spectroscopy, such as infrared absorption (IR), has proved to be a reliable method for fuel characterization. Multivariate analysis methods used to determine fuel specifications include Genetic Inverse Least Squared [53], PCR and PLS [54]. Fig. 3g and h displays the results of the test dataset using SNIRO to predict protein concentration in meat.

## 4. Discussion

### 4.1. Comparison to other computational approaches

*Ulva* sp. and diesel datasets were used to compare SNIRO to other statistical approaches. TRE and Spearman correlation in a test dataset were used as the basis of the comparison. We report running times when relevant. Note that the number of wavelengths that are required to implement any given result is an important parameter in the context of this work. That is – if comparable performances are obtained by SNIRO with 5 λs (L = 5) and by Martens with 8 λs then, from a practical device perspective, the 5 λs solution is far superior and justifies the additional computational resources. The comparative study covered the following methods as benchmarks:

- Martens [13,14]. Several cutoffs were used for some of the datasets as indicated.
- Forward Selection Search (FSS) [55]. This greedy approach selects the next wavelength based on the best matching to the previously selected λs. Namely: we start with the best single wavelength and iteratively add wavelengths to the previously selected ones. Thus, this method is computationally efficient [56]. However, due to its greedy nature, it does not necessarily guarantee the best combination of λs. Once a variable is added to the model using this method, it cannot be removed. Thus, when searching for the best combination of L out of E variables, an exhaustive search of all such combinations is guaranteed to produce the best results, whereas performing a forward selection search may not find the best combination.
- LASSO [57]. LASSO is an appropriate method to compare to SNIRO since it allows the user to indirectly control for the number of wavelengths selected. Thus, we turned the regularization coefficient of LASSO so that it returns 5 wavelengths. In the Ulva dataset, LASSO was not able to predict the highest and lowest values of sugar concentration. For example, the value of 9 was predicted as 5.5, thus lowering its $R^2$ and increasing its TRE.

For each of the methods described above, a best set of NIR parameters was inferred to measure TRE and $R^2$. Results are reported in Table 4 and chosen wavelengths and their corresponding constants are reported in Table S3 of the Supplementary material.

A second comparison, working with published studies that used the same datasets as used here, was performed and is reported in Table 5.

### 4.2. Performance of SNIRO compared to literature methods

The underlying function of SNIRO is to produce a pre-

**Table 4**
Comparison of different methods. CO– Cutoff value used in Marten's Test. Calibration is for the training set and prediction is for the test set. $R^2$ and Total Relative Error (TRE) of regression models are shown.

| | Method | L (number of λs) | Calibration | | Prediction | |
|---|---|---|---|---|---|---|
| | | | $R^2$ | TRE [%] | $R^2$ | TRE [%] |
| *Ulva* sp. | Marten's CO = 0.1 | 8 | 0.41 | 4.98 | 0.41 | 8.24 |
| | Forward Selection | 5 | 0.52 | 2.52 | 0.55 | 6.77 |
| | LASSO | 5 | 0.41 | 4.40 | 0.32 | 8.16 |
| | SNIRO | 5 | 0.61 | 2.48 | 0.62 | 7.23 |
| Diesel | Marten's CO = 0.44 | 12 | 0.76 | 0.78 | 0.74 | 1.61 |
| | Forward Selection | 5 | 0.96 | 0.29 | 0.97 | 0.51 |
| | LASSO | 5 | 0.95 | 0.41 | 0.96 | 0.79 |
| | SNIRO | 5 | 0.96 | 0.28 | 0.97 | 0.58 |

**Table 5**
Comparison of results to other publications on the test dataset.

| Dataset | Method | $R^2$ | RMSE | $R^2$ - SNIRO |
|---|---|---|---|---|
| Diesel - Octane Number | GILS[a] | 0.866 [53] | | 0.97 |
| | PCR[b] | 0.991 [54] | | |
| | PLS | 0.985 [54] | | |
| Corn Flour -Protein | PCR | | 0.14[c] [49] | $R^2$ - 0.23 |
| | PLS | | 0.15[c] [49] | RMSE - 0.46 |
| Meat - Protein | PLS | 0.51 [58] | | 0.32 |
| *Ulva* sp.- Glucose | SNIRO | | | 0.62 |

[a] Genetic Inverse Least Squared (GILS).
[b] Principal Component Regression (PCR).
[c] Root Mean Squared Error (RMSE) of prediction on test dataset.

determined number of wavelengths that best model the concentration of a specific analyte. In comparing the results attained by SNIRO to other publications, it is important to note that most them produced more than 5 λs. SNIRO prioritizes the number of wavelengths above the accuracy of the results, which is something that other methods do not do. SNIRO fared well compared to other literature methods as can be seen in Table 4. Furthermore, SNIRO fared well against other methods that were implemented in this study.

Nonetheless, SNIRO required extensive computational power. Under the conditions studied here, we took an average of 4.5 h to produce results. In comparison, LASSO and Marten's Test took an average of 2 min to produce results, and forward selection took an average of 20 min. SNIRO serves as a first step in potentially developing an efficient field-operable discrete wavelength spectrophotometer. Thus, if accuracy standards are the major concerns, a SNIRO calculation should be performed as part of the design process. Moreover, a single run may suffice for the production of an efficient field operable device. That said, if computational time is a major bottleneck, FSS would be a good choice for the design process. Further investigation of the tradeoff between SNIRO and FSS including the comparison of the actual output (see below) is an important point for future studies with more data.

As indicated above, in the Ulva dataset, LASSO was not able to predict the highest and lowest values of sugar concentration. For example, the value of 9 was predicted as 5.5, thus lowering its $R^2$ and increasing its TRE. This shortcoming of LASSO is because it is not designed to work with a fixed number of wavelengths, rather to regularize the number of non-zero coefficients [31].

In addition to the accuracy of the methods, it is also worth noting the differences in the actual output. Table S3 in the supplementary reports the all relevant selected wavelengths. In principle, the sets returned by the different methods are not identical. Some overlaps between FSS and SNIRO exist, as can be expected. The LASSO output is completely different.

### 4.3. Study considerations

Model accuracy and stability are represented by the prediction quality in the test dataset and not necessarily by the training performance. Over-fitting of the data and bias-variance tradeoffs should be taken into consideration whilst analyzing the results. The number of samples used in each dataset directly influenced the accuracy of the results. Datasets with relatively small numbers of samples may cause over-fitting. In general, the more samples available, the better the models prediction [1]. Furthermore, it is necessary for the analyte content to have a large enough variance to assure the models validity for future samples [1]. The importance of sample size and variance is exemplified by comparing the diesel dataset and corn flour dataset. The diesel dataset had 395 samples with a variance of 44.4 and its model provided the most accurate results in the study. On the other hand, the corn dataset had 79 samples with a variance of 0.23 and its model fared the least accurate.

The histograms in Fig. 3 represent the TRE calculated for shuffled versions of the response vector, compared to the real response vector. Note that for a response vector with small variance, these will be very small, as is the case for the corn flour dataset. As an additional comparison to random data, we also ran SNIRO to predict a shuffled response vector. The performance of this run is indicated by the red star. The conclusion of this analysis is that SNIRO TRE on the real response vector is to the left of both the histogram and the red star, affirming the validity of the selected models.

The quality of the scanning device and method used to scan may also affect the quality of the results [59,60]. For the *Ulva* sp. dataset, multiple measurements of the same sample were taken in order to assess the variance of the scans. It was evident that in many cases multiple scans of the same sample produced different spectral fingerprints. In order to decrease the effect of this variability it is important to produce multiple scans of each sample and to average the data.

## 5. Conclusion

We developed SNIRO as a method for modeling Glucose in *Ulva* sp. based on its VIS/NIR spectral fingerprint. Model performance is directly influenced by the accuracy of the input data. Thus, reliable results primarily require high-quality scans and a large number of samples. For this reason SNIRO was able to produce excellent results for the tested datasets, and most notably for Diesel Octane Number. The total relative error (TRE) demonstrated for the datasets investigated herein, often surpassed the accuracy of the devices used to measure the target analyte values (for example 0.58% for Diesel, 1.41% for corn flour), attesting to the effectiveness of the SNIRO method. Under the given computational restraints, SNIRO was able to provide confident results, and can adapt to other computational platforms by having direct control over the parameters L and E. However, a full analysis of the data is required prior to running SNIRO in order to define the appropriate parameters as indicated in Section 3. Comments on the data analysis methods used in this study can be found in the Supplementary material of the paper. Thus, SNIRO is fully adaptable to the user's requirements and to the available computing resources. It has a dynamic structure that allows the input of desired limitations (for example, limiting the wavelength distance between each selected λ). In particular, the required number of output wavelengths is taken as input by SNIRO and should be provided by the user depending on the needs of the field operation.

Using SNIRO we demonstrated the effectiveness of selecting a limited number of wavelengths for the development of a field-operable spectral canning device.

## Acknowledgements

We would like to thank Microsoft Azure, Microsoft Corporation USA , for its generous and efficient help on operating the tools needed for our work. We thank the Israel Ministry of Energy and Israel Ministry of Science and Technology for their contribution to the study, and the research groups of Alex Golberg in Tel Aviv University and Zohar Yakhini in IDC and Technion for valuable comments and discussions.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at

https://doi.org/10.1016/j.aca.2018.11.038.

## References


[1] P. Williams, K. Norris, Near Infrared Technology in the Agricultural and Food Industries. Near-infrared Technology in the Agricultural and Food Industries, second ed., 2001.
[2] W.F. McClure, Near-infrared spectroscopy: the giant is running strong, Anal. Chem. (1994), https://doi.org/10.1021/ac00073a002.
[3] L. Li, Q. Zhang, D. Huang, A review of imaging techniques for plant phenotyping, Sensors 14 (2014) 20078–20111.
[4] M. Blanco, I. Villarroya, NIR spectroscopy: a rapid-response analytical tool, TrAC Trends Anal. Chem. (Reference Ed.) 21 (2002) 240–250.
[5] T.S. Yeh, S.S. Tseng, A low cost LED based spectrometer, J. Chin. Chem. Soc. 53 (2006) 1067–1072.
[6] R. Civelli, et al., A simplified, light emitting diode (LED) based, modular system to be used for the rapid evaluation of fruit and vegetable quality: development and validation on dye solutions, Sensors 15 (2015) 22705–22723.
[7] R.A. Spragg, Encyclopedia of spectroscopy and spectrometry, Encyclopedia of Spectroscopy and Spectrometry (2017), https://doi.org/10.1016/B978-0-12-803224-4.00088-1.
[8] J.G. Schnable, et al., Portable LED-array VIS-NIR spectrophotometer/nephelometer, Field Anal. Chem. Technol. 2 (1998) 21–28.
[9] D.R. Albert, M.A. Todt, H.F. Davis, A low-cost quantitative absorption spectrophotometer, J. Chem. Educ. 89 (2012) 1432–1435.
[10] R. Laudien, G. Bareth, R. Doluschitz, Comparison of remote sensing based analysis of crop diseases by using high resolution multispectral and hyperspectral data – case study: Rhizoctonia solani in sugar beet, in: Proc. 12th Int. Conf. Geoinformatics – Geospatial Inf. Res. Bridg. Pacific Atl, 2004, pp. 670–676.
[11] M. Arakawa, Y. Yamashita, K. Funatsu, Genetic algorithm-based wavelength selection method for spectral calibration, J. Chemom. 25 (2011) 10–19.
[12] Z.-M. Liu, R. Zhang, G.-M. Zhang, K.-Q. Chen, Wavelength variable selection method in near Infrared Spectroscopy based on discrete firefly algorithm, Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy Spectr. Anal. 36 (2016).
[13] F. Westad, H. Martens, Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression, J. Near Infrared Spectrosc. 8 (2000) 117–124.
[14] M. Forina, S. Lanteri, M.C.C. Oliveros, C.P. Millan, Selection of useful predictors in multivariate calibration, Anal. Bioanal. Chem. 380 (2004) 397–418.
[15] J.C.J.N. Miller, J.C.J.N. Miller, Chemometrics for analytical chemistry, Anal. Chem. (2005), https://doi.org/10.1198/tech.2004.s248.
[16] B.K. Lavine, Workman, J. Chemometrics. *Analytical Chemistry* (2013), https://doi.org/10.1021/ac303193j.
[17] B.K. Lavine, Workman, J. Chemometrics. *Anal. Chem.* (2012), https://doi.org/10.1021/ac303193j.
[18] R. Darvishzadeh, et al., LAI and chlorophyll estimation for a heterogeneous grassland using hyperspectral measurements, ISPRS J. Photogrammetry Remote Sens. 63 (2008) 409–426.
[19] S.P. Serbin, D.N. Dillaway, E.L. Kruger, P.A. Townsend, Leaf optical properties reflect variation in photosynthetic metabolism and its sensitivity to temperature, J. Exp. Bot. 63 (2012) 489–502.
[20] M.F. Dreccer, L.R. Barnes, R. Meder, Quantitative dynamics of stem water soluble carbohydrates in wheat can be monitored in the field using hyperspectral reflectance, Field Crop. Res. 159 (2014) 70–80.
[21] C. Römer, et al., Robust fitting of fluorescence spectra for pre-symptomatic wheat leaf rust detection with Support Vector Machines, Comput. Electron. Agric. 79 (2011) 180–188.
[22] U. Seiffert, F. Bollenbeck, H.P. Mock, A. Matros, Clustering of crop phenotypes by means of hyperspectral signatures using artificial neural networks, in: 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, WHISPERS 2010 - Workshop Program 1–4, IEEE, 2010, https://doi.org/10.1109/WHISPERS.2010.5594947.
[23] J.C. Tewari, V. Dixit, B.K. Cho, K.A. Malik, Determination of origin and sugars of citrus fruits using genetic algorithm, correspondence analysis and partial least square combined with fiber optic NIR spectroscopy, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 71 (2008) 1119–1127.
[24] A.M.K. Pedro, M.M.C. Ferreira, Nondestructive determination of solids and carotenoids in tomato products by near-infrared spectroscopy and multivariate calibration, Anal. Chem. 77 (2005) 2505–2511.
[25] J. Lindedam, et al., Near infrared spectroscopy as a screening tool for sugar release and chemical composition of wheat straw, J. Biobased Mater. Bioenergy 4 (2010) 378–383.
[26] N. Goldshleger, A. Chudnovsky, R. Ben-Binyamin, Predicting salinity in tomato using soil reflectance spectra, Int. J. Rem. Sens. 34 (2013) 6079–6093.
[27] R. Lugassi, A. Chudnovsky, E. Zaady, L. Dvash, N. Goldshleger, Estimating pasture quality of fresh vegetation based on spectral slope of mixed data of dry and fresh vegetation-method development, Rem. Sens. 7 (2015) 8045–8066.
[28] C. Jarén, J.C. Ortuño, S. Arazuri, J.I. Arana, M.C. Salvadores, Sugar determination in grapes using NIR technology, Int. J. Infrared Millimet. Waves 22 (2001) 1521–1530.
[29] B.K. Natarajan, Sparse approximate solutions to linear systems, SIAM J. Comput. 24 (1995) 227–234.
[30] D.L. Donoho, Y. Tsaig, I. Drori, J.L. Starck, Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit, IEEE Trans. Inf. Theor. 58 (2012) 1094–1121.
[31] A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Rev. 51 (2009) 34–81.
[32] Cargill Corn, Flour Data, 2005.
[33] Tecator. Meat Data (1993). http://lib.stat.cmu.edu/datasets/tecator.
[34] Southwest Research Institute, Diesel Data, 2005.
[35] D.L. Donoho, Y. Tsaig, I. Drori, J.L. Starck, Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit, IEEE Trans. Inf. Theor. 58 (2012) 1094–1121.
[36] W.H. Press, S.A. Teukolsky, Savitzky-golay smoothing filters, Comput. Phys. 4 (1990) 669.
[37] D. Aitken, C. Bulboa, A. Godoy-Faundez, J.L. Turrion-Gomez, B. Antizar-Ladislao, Life cycle assessment of macroalgae cultivation and processing for biofuel production, J. Clean. Prod. 75 (2014) 45–56.
[38] A. Bruhn, et al., Bioenergy potential of Ulva lactuca: biomass yield, methane production and combustion, Bioresour. Technol. 102 (2011) 2595–2604.
[39] A. Golberg, et al., Proposed design of distributed macroalgal biorefineries: thermodynamics, bioconversion technology, and sustainability implications for developing economies, Biofuels, Bioprod. Biorefining 8 (2014) 67–82.
[40] S. Kraan, Mass-cultivation of carbohydrate rich macroalgae, a possible solution for sustainable biofuel production, Mitig. Adapt. Strategies Glob. Change 18 (2013) 27–46.
[41] H. van der Wal, et al., Production of acetone, butanol, and ethanol from biomass of the green seaweed Ulva lactuca, Bioresour. Technol. 128 (2013) 431–437.
[42] A.J. Wargacki, et al., An engineered microbial platform for direct biofuel production from brown macroalgae, Science 335 (2012) 308–313.
[43] E. Vitkin, A. Golberg, Z. Yakhini, BioLEGO — a web-based application for biorefinery design and evaluation of serial biomass fermentation, Technology 03 (2015) 89–98.
[44] L. Korzen, et al., An economic analysis of bioethanol production from the marine macroalga *Ulva* (Chlorophyta), Technology 03 (2015) 114–118.
[45] L. Korzen, I.N. Pulidindi, A. Israel, A. Abelson, A. Gedanken, Marine integrated culture of carbohydrate rich Ulva rigida for enhanced production of bioethanol, RSC Adv. 5 (2015) 59251–59256.
[46] N. Robinson, P. Winberg, L. Kirkendale, Genetic improvement of macroalgae: status to date and needs for the future, J. Appl. Phycol. 25 (2013) 703–716.
[47] S. Shefer, A. Israel, A. Golberg, A. Chudnovsky, Carbohydrate-based phenotyping of the green macroalga Ulva fasciata using near-infrared spectrometry: potential implications for marine biorefinery, Bot. Mar. 60 (2017) 219–228.
[48] Y. Fang, X. Chengwei, H. Dan, Analysis and estimate of corn quality by near infrared reflectance (NIR) spectroscopy, in: 2011 Symposium on Photonics and Optoelectronics, SOPO 2011 1–4, IEEE, 2011, https://doi.org/10.1109/SOPO.2011.5780611.
[49] D.A. Burns, E.W. Ciurczak, Handbook of Near-infrared Analysis, 2008, https://doi.org/10.1021/ja015320c.
[50] N. Prieto, R. Roehe, P. Lavín, G. Batten, S. Andrés, Application of near infrared reflectance spectroscopy to predict meat and meat products quality: a review, Meat Sci. 83 (2009) 175–186.
[51] M. Prevolnik, M. Čandek-Potokar, D. Škorjanc, Ability of NIR spectroscopy to predict meat chemical composition and quality - a review, Czech J. Anim. Sci. 49 (2004).
[52] S.R. Daly, K.E. Niemeyer, W.J. Cannella, C.L. Hagen, Predicting fuel research octane number using Fourier-transform infrared absorption spectra of neat hydrocarbons, Fuel 183 (2016) 359–365.
[53] D. Özdemir, Near infrared spectroscopic determination of diesel fuel parameters using genetic multivariate calibration, Petrol. Sci. Technol. 26 (2008) 101–113.
[54] O. Soyemi, M. Busch, K. Busch, Multivariate analysis of near-infrared spectra using the G-programming language, J. Chem. Inf. Comput. Sci. 40 (2000) 1093–1100.
[55] K.Z. Mao, Orthogonal forward selection and backward elimination algorithms for feature subset selection, IEEE Trans. Syst. Man Cybern. B Cybern. 34 (2004) 629–634.
[56] J.M. Sutter, J.H. Kalivas, Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection, Microchem. J. 47 (1993) 60–66.
[57] C. Wang, Y. Yao, H. Liu, J. Wang, Rapid compositional analysis of sawdust using sparse method and near infrared spectroscopy, in: 26th Chinese Control and Decision Conference, CCDC 2014, IEEE, 2014, pp. 4487–4492, https://doi.org/10.1109/CCDC.2014.6852972.
[58] M. Králová, Z. Procházková, A. Saláková, J. Kameník, L. Vorlová, Determination of Meat Quality by Near-infrared Spectroscopy, 2006.
[59] Y. Dixit, et al., Developments and challenges in online NIR spectroscopy for meat processing, Compr. Rev. Food Sci. Food Saf. 16 (2017) 1172–1187.
[60] P. Paz, M.T. Sánchez, D. Pérez-Marín, J.E. Guerrero, A. Garrido-Varo, Evaluating NIR instruments for quantitative and qualitative assessment of intact apple quality, J. Sci. Food Agric. 89 (2009) 781–790.
[61] A. Chudnovsky, A. Golberg, Y. Linzon, Monitoring complex monosaccharide mixtures derived from macroalgae biomass by combined optical and microelectromechanical techniques, Process Biochem. 68 (2018) 136–145.