doi: 10.1255/nirn.1522

NIR news 2014 Data Analysis Challenge: results

Juan Antonio Fernández Pierna and Philippe Vermeulen
Walloon Agricultural Research Centre (CRA-W), 24 Chaussée de Namur, 5030 Gembloux, Belgium

You will remember, I am sure, the challenge which we set you in the last issue of 2014 dedicated to the memory of Jim Burger. This used a dataset supplied by Juan Antonio Fernández Pierna and Philippe Vermeulen of CRA-W, Belgium, a dataset which they had actually generated and used in a real application. Somewhat to our disappointment, we only received two responses to our challenge, and neither of the scientists involved belong to our normal NIR community. They were Xavier Hadoux and Taylor Glenn. However, it seems that they have added to the repertoire of strategies and methods available to imaging NIR spectroscopists since some of the approaches they used were not familiar to the CRA-W scientists. The approaches taken together with the reported results are described and discussed in this article. Given that we only had the two applicants and they both produced interesting and improved solutions, we have decided to award the prize jointly—Ed.

The damage caused by nematode infestations of sugar beet roots eventually leads to a reduction in sugar yield. The size of this reduction is related to the number of cysts present. The current challenge was to detect and quantify by hyperspectral NIR imaging the presence of cyst nematodes on sugar beet root samples.

For this experiment, 20 sugar beet plants with different levels of resistance were grown in a soil support, spread in plastic plates and infested with nematodes. The number of cysts in each sample was independently counted by optical microscopy. Then, one image for each plant was acquired with a pushbroom NIR hyperspectral imaging system. All the images consisted of lines of 320 pixels acquired at 209 wavelength channels (1100–2400 nm) with a spectral resolution of 6.3 nm and result from averaging 32 scans at each line. Around 300 lines (100,000 pixels) were acquired for each sample.

Images 1 to 14 include reference values and are to be used to develop models for quantifying cysts. Samples 1 to 4 also include two RGB images (A: original image; B: image with cysts identified in red). Images 15 to 20 are blind samples, which are to be used to check whether the developed model is consistent (reference values are not included). A summary of the available samples is provided in Table 1.

The challenge was to provide estimates of the number of cysts for a blind test set (Images 15–20).

Two participants reported results which are summarised in Table 2 and Figure 1. In this commentary, these results are compared to those published by CRA-W,[1] suppliers and originators of the original dataset.

Both participants gave very good results and in all cases, the predicted values for the tolerant plants were lower than for their susceptible counterparts.

The originality of the two approaches proposed by the participants is the application of methods generally used on the images acquired by spatial or aerial remote devices. That was not the case for the results published by Fernandez *et al.*[1]

The next sections give a more detailed description of the overall approach and methodology used. Additionally, for sample 19 which is a tolerant plant, RGB images and predicted images indicate, for each approach, the locations of the predicted cysts.

CRA-W

For detection and possible quantification, a complete spectral library, including spectra from the background (including a water feedstrip and a plastic box), soil support, roots and cyst nematodes,

Table 1. Summary details of sample set.

Sample number	Number of cysts	RGB image	NIR image	Set
1	24	yes	yes	Cal
2	49	yes	yes	Cal
3	70	yes	yes	Cal
4	82	yes	yes	Cal
5	33	no	yes	Cal
6	35	no	yes	Cal
7	43	no	yes	Cal
8	50	no	yes	Cal
9	51	no	yes	Cal
10	51	no	yes	Cal
11	55	no	yes	Cal
12	66	no	yes	Cal
13	76	no	yes	Cal
14	77	no	yes	Cal
15			yes	Test
16			yes	Test
17			yes	Test
18			yes	Test
19			yes	Test
20			yes	Test

Table 2. Summary of results.

		#Cysts Ref	#Cysts CRA-W		#Cysts Participant 1		#Cysts Participant 2	
	Code		Pred	diff.	Pred	diff.	Pred	diff.
Tolerant	1	24	8	17	21	3	23	1
Tolerant	2	49	12	37	40	9	47	2
Susceptible	3	70	79	–9	79	–9	65	5
Susceptible	4	82	82	0	71	11	72	10
Tolerant	5	33	24	9	45	–12	33	0
Tolerant	6	35	15	20	33	2	36	–1
Tolerant	7	43	27	16	61	–18	40	3
Tolerant	8	50	17	33	46	4	55	–5
Tolerant	9	51	49	3	53	–2	53	–2
Susceptible	10	51	26	25	45	6	50	1
Susceptible	11	55	63	–8	47	8	52	3
Susceptible	12	66	32	34	61	5	61	5
Susceptible	13	76	53	23	74	2	74	2
Susceptible	14	77	71	7	79	–2	74	3
Tolerant	15	24	20	4	36	–12	28	–4
Susceptible	16	60	29	31	56	4	48	12
Tolerant	17	37	5	32	37	0	34	3
Susceptible	18	69	71	–2	61	8	65	4
Tolerant	19	49	31	18	55	–6	33	16
Susceptible	20	80	80	0	77	3	60	20
Mean susceptible		69	58		65		62	
Mean tolerant		40	21		43		38	
RMSEC Cal				20.7		8.1		3.9
RMSEP Test				19.7		6.7		11.8

was built by selecting around 500 pixels in each region of interest on the images of 10 plants (five tolerant and five susceptible). In total, more than 2000 spectra were used to build the SVM discrimination models. The spectra dataset was pre-processed using smoothing (window = 5), SNV and first derivative Savitzky–Golay (window = 5, polynomial = 2). A dichotomist classification tree was built including the following steps:

(1) Detection of pixels in the image, showing a higher absorbance around 1690 nm than around 1970 nm, corresponding to the conveyor belt;
(2) Detection of pixels in the image, detected as soil support, roots or cysts by the SVM model "background vs soil support+root+cyst";
(3) Detection of pixels in the image, detected as roots or cysts by the SVM model "soil support vs root+cyst");
(4) Detection of pixels in the image, detected as cysts by the SVM model "root vs cyst";
(5) Removal of the pixels classified as outliers according to rules based on the comparison of absorbance at several wavelengths; the cysts showing a lower absorbance around 1734 nm than around 1715 nm and 1765 nm;
(6) Application of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method to study the neighbourhood of the pixels detected as cysts in step (5). Using this technique, pixels within 1 pixel of each other and with a minimum of two neighbour pixels were placed in a single class and identified as a cyst. Pixels that did not meet these conditions were identified as outliers. Once the models were constructed and validated, the complete discrimination tree, including the three equations and the spectral rules, was applied successively to all the pixels in the images of the 20 plants in order to estimate the number of pixels detected as cysts by surface unit.

This method is described in detail in Reference 1.

T. Glenn (USA)

A single prototype target signature for the cysts was extracted from one of the calibration images and used in a target detection algorithm in each of the images.

The detection prototype spectrum was extracted from the CC4 image by using the pixel value at row 92, column 143.

The best results were found by setting a high threshold to achieve a low false-positive rate. However, this resulted in fewer of the cysts being detected than were given as the calibration number. To provide an estimate of the population, a scale factor was estimated.

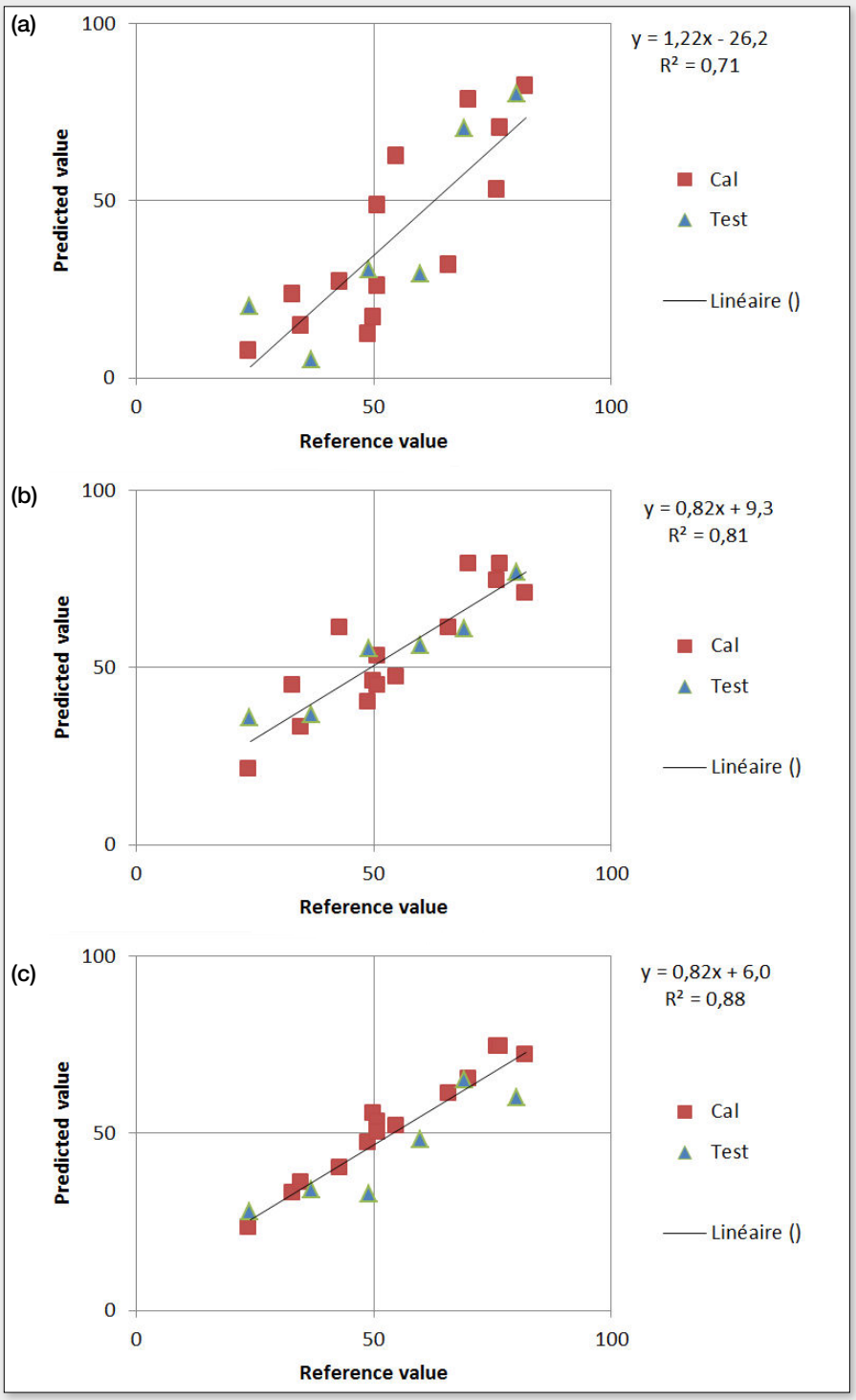

Figure 1. Predicted versus actual results reported by (a) Fernandez *et al.*,[1] (b) participant 1 and (c) participant 2. By comparing the three results, those obtained by participant 2 gave the best determination coefficient (0.88) and the best *RMSEC* calculated on the calibration set (3.9). However, the *RMSEP* calculated on the six samples of the test set was three times higher (11.8). The results from participant 1 gave quite a good determination coefficient (0.81) and a very low *RMSEP* (6.7). The published results (Fernandez *et al.*)[1] indicated a lower determination coefficient (0.71) and higher *RMSEP* (19.7).

- Reduce dimensionality to 50 bands by hierarchical band merging based upon mutual information, this is done mostly to reduce processing time, but it also slightly improves performance.
- Difference along spectral dimension, this slightly reduces the noise sensitivity.
- Use the ACE (Adaptive Cosine Estimator) target detection statistic where the mean and covariance are estimated from the entire image.
- Mask off detection into a region of interest. This is mostly to avoid the "hotspot" found within top 50–75 rows of the image that causes several false detections.
- Threshold the detection output to form detection blobs.
- Select regional maxima of semi-thresholded detection output as the hit locations.
- Estimate true population count from detections.
- Multiply detection count by a scale factor to estimate the population count.

For best performance, a detection threshold was selected that gave detection counts which best correlated to the calibration population counts. The threshold was found by seven-fold (leave two out) cross-validation of the calibration set. The threshold that minimised squared error of the population estimate in cross-validation was selected. Then, with the threshold level selected, a simple scale factor estimator was trained on the entire calibration set.

The best correlation between detection counts and population counts was found by setting the detection threshold to 0.55, this implied a scale factor of 1.48 to estimate the population count.

X. Hadoux (France)

- Camera radiometric calibration. The first row of the first image was used to estimate the difference in spectral responses between the columns. The white background represented in the first row which was, of course, supposed to be homogeneous. For the columns in which the root was overlapping the white background, the estimation was made using another image. A relative correction (to the left column) was thus performed by subtracting the difference between each column to the left column.
- Reflectance calibration. To compensate for the lighting differences between images, a logarithm transformation of the radiance image $L_{i,j}(\lambda)$ (i and j are the pixel coordinates and λ the wavelength) was performed. After centring, the additive part log $E(\lambda)$ cancelled out thus leading to images that are independent of the lighting conditions.
- Also, because of low values (log), spectral bands <10 and >180 were removed.

continued on page 15

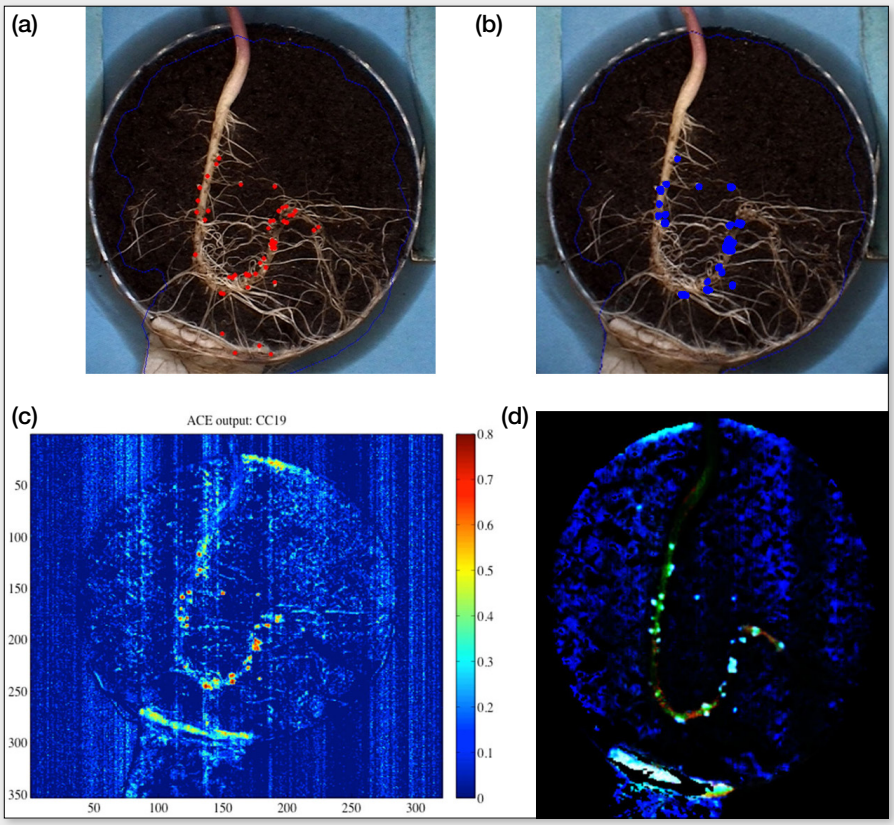

Figure 2. (a) RGB image, (b) predicted image (CRA-W), (c) predicted image (Glenn), (d) predicted image (Hadoux).

continued from page 13

- Background removal. Specific combinations of wavelengths were manually chosen. The removed parts of the images were: the white background, the pot, some roots with very low spectral values (dead?) and other soil related pixels with very low spectral responses.
- Classification. Pixels that corresponded to cysts had to be manually extracted. Some pixels (the obvious one) were first extracted according to four classes: cyst, taproot, fibrous and soil.

A PLS-LDA (LDA on PLS scores) was performed using five LV and three discriminant vectors (DV). The resulting LDA scores (that well separate different classes) were then used to create a false-RGB image with enhanced contrast between classes.

The complete ground truth creation using the first four images was then manually made with the help of these RGB images.

- Classification Models. A single model was not satisfactory to correctly classify the cysts because of high similarity between spectral responses of every class. I thus used three different PLS-LDA models to increase the discriminatory power:
  - Model 1: PLS-LDA four classes with five LV and four DV.
  - Model 2: PLS-LDA two classes (cyst versus the rest) with seven LV and one DV.
  - Model 3: PLS-LDA two classes (cyst versus taproot) with four LV and one DV.
- Class decision (model fusion). Probability maps were computed of observing class cyst for with the three models. Then, because outliers were present in the three models (but were usually situated at different spatial positions), the cyst were detected on the product of the three probability maps.

Reference

1. J.A. Fernández Pierna, P. Vermeulen, O. Amand, A. Tossens, P. Dardenne and V. Baeten, "NIR hyperspectral imaging spectroscopy and chemometrics for the detection of undesirable substances in food and feed", *Chemometr. Intell. Lab. Syst.* **117,** 233–239 (2012). doi: http://dx.doi.org/10.1016/j.chemolab.2012.02.004