Peer Reviewed Paper openaccess Paper Presented at IASIM 2018, June 2018, Seattle, WA, USA

Comparison of spectral selection methods in the development of classification models from visible near infrared hyperspectral imaging data

Aoife A. Gowen,[a] Jun-Li Xu[b]* and Ana Herrero-Langreo[c]

[a]UCD School of Biosystems and Food Engineering, University College of Dublin (UCD), Belfield, Dublin 4, Ireland.
https://orcid.org/0000-0002-9494-2204

[b]UCD School of Biosystems and Food Engineering, University College of Dublin (UCD), Belfield, Dublin 4, Ireland. E-mail: junli.xu@ucd.ie,
https://orcid.org/0000-0002-4442-7538

[c]UCD School of Biosystems and Food Engineering, University College of Dublin (UCD), Belfield, Dublin 4, Ireland.
https://orcid.org/0000-0003-3258-6248

Applications of hyperspectral imaging (HSI) to the quantitative and qualitative measurement of samples have grown widely in recent years, due mainly to the improved performance and lower cost of imaging spectroscopy instrumentation. Data sampling is a crucial yet often overlooked step in hyperspectral image analysis, which impacts the subsequent results and their interpretation. In the selection of pixel spectra for the calibration of classification models, the spatial information in HSI data can be exploited. In this paper, a variety of sampling strategies for selection of pixel spectra are presented, exemplified through five case studies. The strategies are compared in terms of the proportion of global variability captured, practicality and predictive model performance. The use of variographic analysis as a guide to the spatial segmentation prior to sampling leads to the selection of representative subsets while reducing the variation in model performance parameters over repeated random selection.

Keywords: hyperspectral imaging, data sampling, classification, spatial, variographic analysis

Introduction

A key step in the successful implementation of hyperspectral imaging (HSI) applications is the development of robust calibration models. The HSI data cube, or "hypercube", is information rich, typically containing hundreds of thousands of spatially co-registered spectra. This abundance of data, while advantageous from a modelling perspective, is also challenging to deal with. On the one hand, in HSI we are rapidly presented with a large number of spectra with which to characterise samples; on the other, such a large volume of data is computationally burdensome and large portions of it may be redundant. A further challenge, in the development of calibration models from HSI data, is that reference or measured values for a property of interest are not usually available at a pixel level. In conventional spectroscopy, it is common to use a single bulk sample value, i.e. an

Correspondence
Jun-Li Xu (junli.xu@ucd.ie)

Received: 23 November 2018
Revised: 7 January 2019
Accepted: 7 January 2019
Publication: 17 January 2019
doi: 10.1255/jsi.2019.a4
ISSN: 2040-4565

Citation
A.A. Gowen, J.-L. Xu and A. Herrero-Langreo, "Comparison of spectral selection methods in the development of classification models from visible near infrared hyperspectral imaging data", *J. Spectral Imaging* **8,** a4 (2019). https://doi.org/10.1255/jsi.2019.a4

average value characterising a homogenous sample, matched to a single spectrum representing the entire sample. The direct translation of this approach in HSI would be to calculate a mean spectrum from the HSI data representing each sample for which a response (*Y*) was measured. This is the approach followed in many research works where the objective is to apply HSI for quantitative analysis. Examples include the prediction of oil concentration in corn kernels,[1] quality prediction of strawberries,[2] sugar content in apples,[3] and quality of mushrooms.[4] However, it is also possible to build predictive models using multiple pixel spectra or multiple mean spectra, calculated from different regions of a hypercube, matched to a single *Y* value. In the case of discriminant analysis, the advantage of using multiple spectra, rather than a single spectrum for each image, is more obvious, in that it makes the spectral variation associated with each class implicit in the discriminant model which generally improves model robustness.

Representative subset selection is a common problem in large datasets. Daszaykowski *et al.*[5] compared several uniform and cluster-based designs for the selection of representative spectra from databases of NIR spectra.[5] Uniform designs, such as the Kennard and Stone (KS) algorithm[6] aim to cover the data space uniformly, while cluster-based designs involve an initial clustering of the data followed by representative subset selection. In terms of uniform selection methods, they found that the "OptiSim" algorithm, proposed by Clark,[7] was more computationally efficient than KS, with a running time 3–4 times faster than achieved with KS, while providing a similar distribution of selected spectra, as observed in principal components 1 vs 2 score plots. This method requires the input of several user-defined parameters including a threshold distance between dissimilar objects, the number of objects to be selected and the size of the subset. In terms of cluster-based designs, they observed that some small clusters risked being underrepresented when a small total number of clusters was selected. In order to avoid this risk, a higher total number of clusters could be used. This could be through minimising the *k*-means cost function or a preliminary density-based clustering step in the principal component score space can be carried out prior to the application of *k*-means within each cluster, although this would increase the computational burden.

In common with selection of spectra from traditional spectroscopic data, in order to optimally select spectra from HSI data for inclusion in a calibration model, it is necessary to select the minimum number of spectra that can be considered representative of a given object. The subset of selected spectra should represent all relevant sources of variability in the dataset. However, hyperspectral images obviously contain additional spatial information. Therefore, selection of spectra from hyperspectral images can be framed as a spatial sampling problem and thus concepts from the Theory of Sampling may be applied. For instance, applying variographic analysis to spatial data facilitates quantisation of correlation between spatially congruent intensity values.[8] This type of analysis could be useful in defining regions over an image from which spectra may be selected. Variogram analysis has previously been used in HSI, for the evaluation of maize plants.[9] It was shown that variogram parameters (nugget, sill and range) derived from single-band images could be used as explanatory variables for the prediction of spider mite infestation and drought stress. However, to the best of our knowledge, variographic analysis has not been used as a tool for selecting representative spectra from HSI data. The objective of this work is to investigate the effects of incorporating sample variation in the calibration model on robustness of calibration models. A variety of sampling strategies for selection of pixel spectra are presented, exemplified through a number of real and synthetic datasets. The strategies were compared in terms of:

1) the resultant representativeness of the selected spectra,
2) resultant performance of classification models,
3) computational time.

Materials and methods

Hyperspectral imaging instrumentation

Diffuse reflectance images of various samples (described below) were obtained using two pushbroom hyperspectral imaging systems (DV Optics Ltd, Padua, Italy): an NIR system (wavelength range 950–1650 nm, pixel size approximately 320 × 320 μm, described further in Reference 10) and a Vis-NIR system (wavelength range 450–950 nm, pixel size approximately 170 × 170 μm, described further in Reference 11). Direct reflectance spectra were used for subsequent data analysis.

Data analysis

All data analysis was carried out using Matlab (release R2014b, The MathWorks, Inc., Natick, MA, USA) incorporating functions from the Image Processing and Statistics toolboxes and additional functions written in-house.

Spectral selection

Five different spectral sampling strategies were evaluated, as described below. For each dataset, a separate image of each class was obtained, denoted the "class image". Spectra were selected from the class image, and spectra selected from each class were combined in order to build a calibration set of data for modelling. For each of the selection methods involving random selection (i.e. Strategies 1, 3, 4 and 5 described below), the procedure of selection was repeated 100 times. In order to have a fair comparison of the methods, the desired number of spectra to select from each class was defined by the regular grid method (Strategy 4), as follows. The grid method subsets the class image into an $N \times N$ grid, one spectrum is selected from each rectangle of the grid. The total number selected is counted (this is the desired number of spectra to select) and the percentage of pixels that this corresponds to is calculated. For strategies 3 and 5, the percentage is used to define the number of pixels selected from each cluster or variogram-based rectangle of the grid. In instances where the percentage was not an integer, the number was rounded up to the nearest integer using the "ceil" function in Matlab.

Strategy 1: Random selection. This, the simplest and most commonly used strategy, involved unfolding the non-background spectra of the class image, randomly permuting the order of the spectra (using the randperm.m function in Matlab) and subsequently selecting the desired number of spectra from the permuted list.

Strategy 2: Kennard and Stone selection. This strategy involved applying the Kennard and Stone algorithm (KS)[6] to select a desired number of the spectral points that optimally span the spectral variation in the dataset. It can be summarised in the following steps:

(i) the first selected spectrum is that nearest the mean spectrum,
(ii) the next selected spectrum is that furthest from (i),
(iii) the next selected spectrum is that which is furthest from (i) and (ii) and
(iv) the procedure is continued until the desired number of spectra are selected.

Strategy 3: Spectrally stratified random selection. This strategy involves applying a clustering algorithm (*k*-means in this case) to the data to initially estimate the number of clusters in the class image. As a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation), silhouette values range from −1 to +1, where a high value indicates that the object is well matched to its own cluster and far away from the neighbouring clusters. The mean silhouette value of each individual class image is computed for 2–5 clusters and the optimal number of clusters is determined with the maximum mean silhouette. Following selection of the number of clusters, the desired number of spectra were randomly selected from each cluster.

Strategy 4: Spatially stratified random selection: user-defined rectangular grids. This strategy employs spatial information by subsetting the class image into an $N \times N$ grid, where N ranged from 5 to 10. The horizontal and vertical sizes of the grid rectangles were estimated as the lowest nearest integer to the image dimensions divided by N. From each rectangle of the grid, one pixel spectrum is randomly selected. For any given value of N for any class image, the total number of spectra selected using the grid method was set as the desired number of spectra in strategies 1, 2, 3 and 5.

Strategy 5: Spatially stratified random selection: variogram-defined rectangular grids. This method is similar to Strategy 4, except the size of the grid rectangles is estimated using a variographic approach, as follows. The North–South and East–West directional semi-variogram (defined in Equations 1 and 2, respectively) of the mean hyperspectral image is calculated, and the lag distance at which the semi-variogram intersects the global variance of the image is selected as the grid size. In cases there was no intersection, or where the intersection point occurred at a distance greater than half of the domain size the grid size was set to half of the domain size.

North–South (vertical):

$$\gamma(h) = \frac{1}{2M}\sum_{i=1}^{M}\left[G(x,y) - G(x,y+h)\right]^2 \quad (1)$$

East–West (horizontal):

$$\gamma(h) = \frac{1}{2M}\sum_{i=1}^{M}\left[G(x,y) - G(x+h,y)\right]^2 \quad (2)$$

where γ is the semi-variance and M is the total number of pairs at a distance of h, while $G(x, y)$ is the pixel intensity of image at (x, y).

Comparison of sampling techniques

Each sampling strategy was compared in terms of representativeness, time required (using an Intel® Core™ i5-7600 CPU @ 3.50GHz; x64-based processor, 8 GB RAM) and resultant classification performance. Representativeness of selected spectra was estimated using the RMS (root mean squared) statistic.[12] The RMS statistic for each selected spectrum is defined in Equation 3, where y_{ij} is the spectrum of the j^{th} sample, $\overline{y_i}$ is the mean spectrum of n selected spectra of an image and n refers to the number of wavelengths.

$$RMS_j = \sqrt{\sum_{i=1}^{n} \frac{\left(y_{ij} - \overline{y_i}\right)^2}{n}} \quad (3)$$

Partial least squares discriminant analysis (PLSDA)[13] was used to build models to classify the samples in each dataset. A calibration set of data was generated using each of the sampling strategies, where each spectrum was assigned a class number depending on which class it represented. This resulted in, for each example dataset, five calibration sets (i.e. one for each sampling method). The number of latent variables for each model was fixed at eight for each dataset, to prevent variations in model performance arising from variations in the number of latent variables. A single common **test set** was created from the remaining spectra that were not used in any of the calibration sets, to enable comparison of model performance. In addition, a **prediction image** containing a mixture of all classes was used for model evaluation. For each of strategies 1, 3, 4 and 5, the model resulting in a classification performance closest to the mean (out of the 100 randomised runs for grid $N = 10$) was selected and applied to the mixture image. Model performance was evaluated by % correct classification (% CC) and geometric mean (G-mean) metrics. G-mean indicates the balance between classification performances on the majority and minority classes by taking into consideration both sensitivity (accuracy of the positive object) and specificity (accuracy on the negative object).

Datasets

In this study, five datasets were used to compare the performance of different spectral selection methodologies in terms of representative spectral subset selection and classification model performance. These datasets are described in detail below and representative colour images of them are shown in Figure 1. Mean spectra of all samples were displayed in Figure S1 (see Supplementary Material). For each dataset, a classification model was built on the selected spectra to classify pixels in the dataset.

Dataset 1: Nuts and dried fruit

Hyperspectral images of a mixture of dried fruit and nuts were obtained using the Vis-NIR hyperspectral imaging system. Images of one individual nut or fruit sample were used for model calibration and the model was tested using an image containing each type of sample. The samples were placed on a black piece of sandpaper for imaging. The background was removed from each image by thresholding. Due to the varying colours of the samples with respect to the background, different strategies were used to maximise the contrast between the sample and background. For background removal of the almond sample, the ratio of images at 825 nm and 495 nm was thresholded by setting all pixels >4 to 1 and the remaining pixels to 0. The peanut was separated from the background by applying principal components analysis (PCA) to the image and setting to 1 all pixels in the PC 2 score image that were less than zero. For the remaining samples, PCA was applied and all pixels in the PC 1 score image that were less than zero were set to 1. The "imfill" function of the Matlab image processing toolbox was used to fill any non-background regions that remained after thresholding.

Dataset 2: Packaging materials

NIR hyperspectral images of five different packaging materials: high density polyethylene (HDPE), polystyrene (PS), polypropylene (PP), cardboard and polyethylene-terephthalate (PET). Single images of each material were obtained as well as a mixed image including all materials. In the mixed image, the materials were cut into different shapes to enable their identification, as described in the caption of Figure 1. The materials were placed on a white tile for imaging. The image background was removed by thresholding. Again, due to the varying spectral characteristics of the samples, it was necessary to find different strategies for background removal in each sample type. For the cardboard, HDPE and PS samples a band ratio image of 1321–1489 nm with a threshold of 1.04 was sufficient for background removal, while for PET samples, a ratio image of 1594–1650 nm with a threshold of 1.04 provided good background removal. For the PP samples, the PC 1 scores image was thresholded with a value of 0.1.

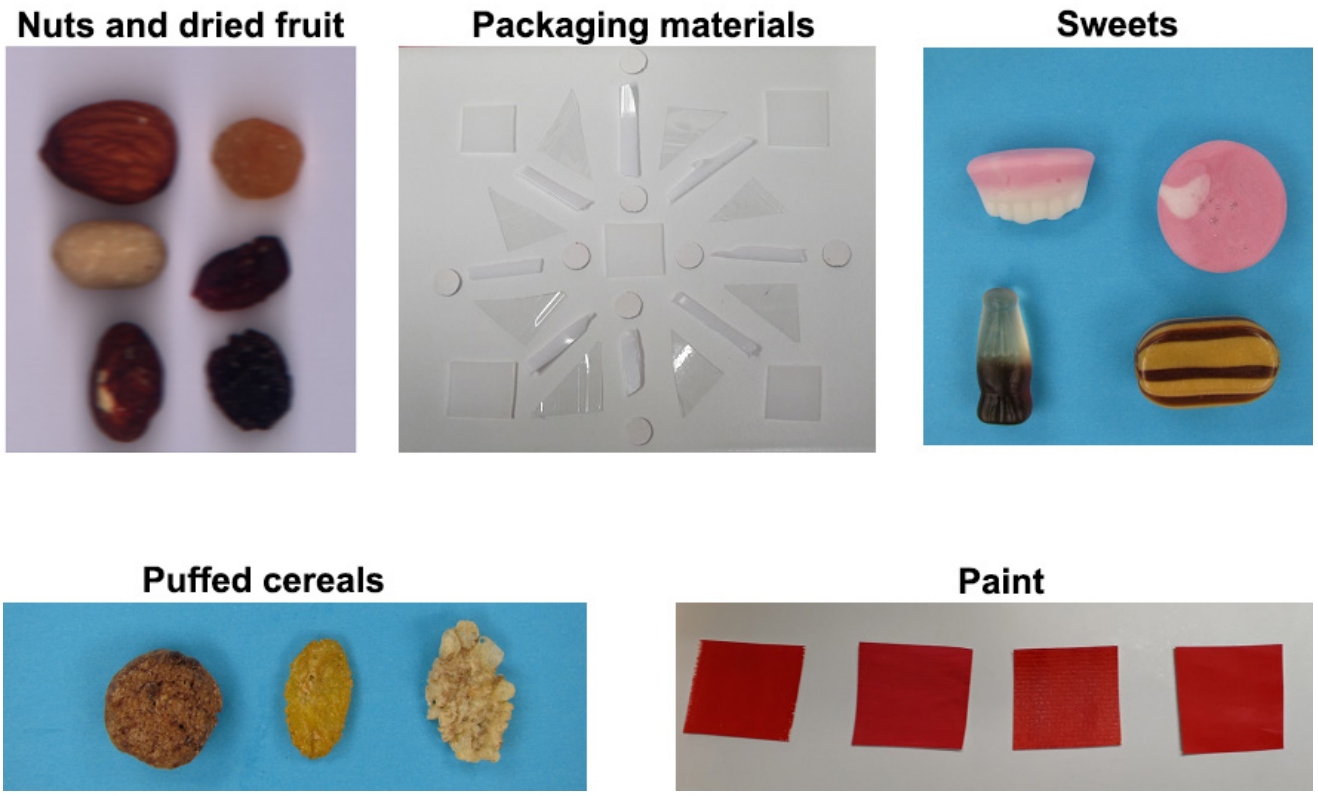

Figure 1. Colour images of representing all samples used in the study, from top left: Nuts and dried fruits used for model calibration and validation [going from top-left to bottom: almond; peanut; hazelnut, going from top-right to bottom: sultana, cranberry, raisin); Packaging materials used for validation [cardboard (circles), polypropylene (PP, thick rectangles), high density polyethylene (HDPE, squares), polystyrene (PS, thin rectangles), polyethylene terephthalate (PET, triangles)]; Sweet samples used for calibration (from top-left to bottom-right: teeth, RFM, tub and mint); Puffed cereal samples used for model calibration (from left to right: wheat, corn and rice); Paint paper samples used for calibration (from left to right paint types are oil, acrylic, water colour and ink).

Dataset 3: Sweets

This dataset consists of NIR hyperspectral images of four different sweets. The shape, colour and nutritional compositions were different among the four selected products: raspberry flavour mushroom in pink and white colour with a mushroom shape; mint humbugs in brown and golden stripe with an ellipse shape; teeth and lips in pink and white colour with a teeth-like appearance, and tub in brown with a cola bottle shape. These sweets, made and purchased from Tesco Ireland Ltd, were labelled as RFM, Mint, Teeth and Tub, respectively. Two images for each sweet type were obtained with one used for calibration and the other for prediction purposes. A mixed image involving all the sweets was also acquired. For background removal, PCA was first applied on the standard normal variate (SNV) pre-processed spectra. A mask for each sample was obtained by thresholding the PC1 score image. The threshold value was manually selected based on the histogram of the PC1 score image.

Dataset 4: Puffed cereal

This dataset consists of NIR hyperspectral images of three types of puffed cereals: honey nut corn-flakes, crunchy cookie cereal and crisp flakes of rice made and purchased from Tesco in Ireland. The sample types were labelled as Corn, Wheat and Rice, respectively, according to their main components. Two images for each cereal type were obtained with one used for calibration and the other for prediction purposes. A mixed image involving all the cereals was also acquired. Background removal was carried out in a similar way as described for the sweets sample.

Dataset 5: Paint

This dataset consists of NIR hyperspectral images of four types of red pigments applied on watercolour paper: oil (Rembrandt oil colour Series 3, 377 permanent Red medium), acrylic (Primary red magenta, Prismo, 256), Watercolour (Van Gogh, G371 Red) and ink (Pebeo, colorex watercolour ink; 50% mixture of 60 magenta and 59 primary yellow). A 2.5 × 2.5 cm square sample of each pigment applied on watercolour paper was used for calibration. An independent prediction image was composed of a mosaic of six to eight samples of each pigment on water colour paper cut in different shapes to facilitate their identification. The reflectance histogram at 1468 nm was used to remove the image background. The threshold value for each image was selected by the Otsu method,[13] applied via the "graythresh" function of the Matlab image processing toolbox.

Results and discussion

Variographic analysis for each dataset

Variogram plots are computed and shown in the Supplementary Material for each dataset. Due to different spatial patterns, the presented semi-variogram plot differs from sample to sample. The variogram of the Mint of Dataset 3, shows a clear example of how the spatial variability of the images is registered by the variogram plot. The directional variogram North–South follows a sinewave-type periodic oscillation. This is related to the horizontal stripe pattern of the mean hyperspectral image. The amplitude of the North–South variogram curve for the Mint corresponds to the average vertical distance between same colour stripes on the mean hyperspectral image of the Mint (15 pixels). This is because the lowest semi-variance corresponds to pairs of pixels for which the values are most correlated. In the case of the Mint, the pixels with the most correlated values would be pixels from the same colour stripes. This means that low semi-variances appear periodically at the same distance as the stripes repeat. Regarding the East–West spatial pattern of the image, the main spatial variation in this direction is due to the shape of the sweet. The shadows and the pattern of the lines at the East and West edges of the sweet produce a higher frequency of low intensity values on these regions. The width of the sweet in the image is 60 pixels. Accordingly, the East–West semi-variance starts decreasing after a lag distance of 40 pixels and reaches a local minimum at 60 pixels. This means that pairs of pixels located at 60 pixels of distance of each other (i.e. pairs of pixels at opposite ends of the sweet in the East–West direction) tend to have correlated values. This correlation can be explained by the shadows and similar spatial patterns found on the East and West edges of the sweet. This effect of the shape of the samples on the variogram can be observed for every curved sample and it is especially evident in the most regularly curved samples, such as the Peanut in Dataset 1 (Figure S2) and the PS spoon in Dataset 2 (Figure S3).

Comparison of sampling times for each method

The sampling times required for each method and dataset are shown in Table 1. In general, the KS selection method was the slowest, requiring up to 4 min, depending on the sample set. The Stratified sampling method was faster (0.3–0.6 s), while the grid and variogram approaches were faster again and took a similar amount of time for selection (0.05–0.15 s). The random method was by far the fastest overall (0.001–0.003 s). The time difference between different datasets is much higher in KS than in the other methods. Datasets in the table are ordered according to the number of pixels in the calibration dataset. Packaging materials are by far the largest samples, followed by nuts. Both datasets have more than 20,000 pixels as an average for each class on the calibration dataset, while the paint, sweets and cereal dataset have below 10,000 average number of pixels for each class. As observed in Table 1, except for the stratified method, sampling times were higher for samples with higher number of pixels on the calibration dataset. KS sampling times were by far the most affected by the average size of the calibration samples. This could be explained by the fact that the other methods are based on separating the calibration dataset in groups and then randomly selecting a certain number of pixels from each group; while the KS method directly selects pixels for calibration out of the whole calibration dataset. Regarding the stratified sampling method, as it is an iterative method where the clustering needs to converge according to the spectra of the pixels, the sampling times could be more related to the spectral differences between classes than to the size of the samples, which would explain why the sampling times for this method were not ordered according to the size of the calibration dataset.

Table 1. Average sampling times over classes of calibration samples for each method and dataset (seconds).

Dataset (*N* pixels)	Rand	KS	Strat	Grid	Var
Packaging (27,529)	0.003	254.886	0.410	0.133	0.152
Nuts (21,187)	0.003	110.319	0.371	0.127	0.142
Paint (8544)	0.001	2.949	0.336	0.065	0.045
Sweets (4287)	0.001	2.009	0.370	0.049	0.042
Cereal (3738)	0.001	1.242	0.555	0.049	0.038

Classification model performance for each dataset and sampling method, in terms of G-mean and %CC calculated on test set and prediction image, are shown in Table 2 (with the highest classification metric in each column printed in red). Compared to the model built with all the pixels from each class image (see prediction maps in Figure S7), the best sampling method has presented better or similar performance. Overall, the %CC for the spectral test set was similar to that of the Image dataset, with the exception of the "Paint" dataset, as discussed further below. With respect to the "Nuts and dried fruits" dataset, the random selection method resulted in the best model performance in terms of the %CC of the spectral test set, while the Variogram method performed best in terms of the test image. However, the random, Grid and Variogram selection methods performed similarly to each other when applied to the prediction image, with %CC ranging from 92% to 92.61% and G-mean ranging from 0.92 to 0.93. On the other hand, for the "Packaging" dataset, while the Grid and Variogram methods performed best in the spectral test set, the KS and Stratified sampling methods performed best when applied to the prediction image, and were very similar to the Grid and Variogram selection methods in terms of %CC and G-mean. Considering the "Cereal" dataset, the KS selection method performed best overall, resulting in substantially higher G-mean and %CC in both the spectral test set and prediction image, whereas for the "Sweet" dataset, the Grid method resulted in the best model performance. Finally, for the "Paint" dataset, the Grid and Variogram methods performed best on the spectral test set (although all selection methods produced a similar %CC of >99%), while the KS and random selection methods performed best on the prediction image. Moreover, the %CC on the "Paint" prediction image (ranging from 75.49% to 77.17%) was substantially lower than that for the spectral test set (ranging from 99.12% to 99.29%). Observing the variations in model performance over the different datasets studied, it is clear that no single method of spectral selection is optimal. However, it is interesting to compare the selection methods in more detail, in order to gain insights as to when a given selection method may be more appropriate to use. With this in mind, the following sections discuss the results for each dataset individually.

Performance indicators for Dataset 1: Nuts and dried fruit

The number and percentage of selected pixels (averaged over the six classes), grid size and sampling method is shown in Table 3. As the number of grids increased from 5×5 to 10×10 the average number of pixels selected from each class increased from 31–28 to 98–104, corresponding to 0.2% to 0.7% of pixels per class. The number and corresponding percentage of pixels selected using the Variogram method was consistently higher than each of the other methods. This is because the "ceil" function of Matlab to round to the next higher integer was used to determine the number of spectra selected from each rectangular grid.

Although the mean model performance indicators presented in Table 2 and discussed in the previous section give an initial basis for comparison of the different sampling methods tested, they do not provide any information on the stability of performance metrics over repeated random sampling. In order to display this information, boxplots showing variation in selection method performance indicators are shown in Figure 2. The top row shows the variation in %CC of the 100 bootstraps of the selection method as calculated on the spectral validation set. For the random selection method, the %CC was quite stable, remaining at around 91% as the number of grids increased. However, the variation in %CC decreased as the number of grids increased, indicating higher model stability. By comparison, for any given number of grids there was no variation in the KS, since the KS selection

Table 2. Classification model performance for each dataset and sampling method, in terms of G-mean and % correct classification (%CC) calculated on test set and prediction image. The highest classification metric in each column is printed in red.

Method	Nuts and dried fruit			Packaging materials			Cereal grains			Sweet			Paint		
	%CC-1	%CC-2	G	%CC-1	%CC-2	G	%CC-1	%CC-2	G	%CC-1	%CC-2	G	%CC-1	%CC-2	G
Full	–	92.32	0.92	–	99.14	0.99	–	88.16	0.88	–	93.32	0.94	–	77.47	0.76
Rand	91.40	92.00	0.92	98.32	97.52	0.97	86.43	86.17	0.86	96.35	92.29	0.93	99.26	77.07	0.76
KS	88.95	91.02	0.90	98.74	99.21	0.98	90.38	91.63	0.92	96.55	90.82	0.90	99.12	77.17	0.76
Strat	91.28	90.87	0.90	98.82	99.03	0.99	86.68	88.37	0.88	96.30	92.53	0.93	99.26	75.49	0.74
Grid	90.16	92.21	0.92	99.36	98.55	0.98	86.12	86.57	0.87	97.06	93.19	0.94	99.29	76.00	0.75
Var	91.26	92.61	0.93	99.03	98.34	0.98	86.02	86.23	0.86	96.71	92.80	0.93	99.29	76.25	0.75

%CC-1: %CC for test set; %CC-2: %CC for prediction image; G: G-mean for prediction image; Full indicates the model built using all the pixels from each class image.

method has no random component. However, the %CC for the KS selection increased with the number of grids from 86% (5×5) up to a stable value of 89% from 7×7 grids. Similar to the random sampling method, the stratified sampling method resulted in a stable %CC around 91% at all numbers of grids, with a smaller variation in %CC at higher numbers of grids. The regular grid method exhibited unusual behaviour with respect to %CC, being higher for odd numbers of grids than for even numbers of grids. The overall highest %CC (92%) and lowest variation in %CC was found for 9×9 grid selection. As for the variogram selection method, the %CC increased with the number of selected spectra, reaching a maximum value around 91% for the 10×10 grids. Although the random selection method resulted in the best model performance in terms of the %CC of the spectral validation set (Table 2), the Variogram method produced more stable model performance metrics.

The bottom row of Figure 2 shows the absolute deviation of the RMS of the selected pixels from the global RMS of the image from which the pixels were sampled. The lower this value is, the more representative the selected spectra are of the original data from which they were sampled. Considering at first the absolute deviation of RMS for the spectra selected using the KS method, it is clearly higher than that for any of the other selection methods, indicating the KS-selected spectra are least representative. In addition, the absolute deviation of RMS decreases as the number of sampled spectra increases, which could be expected: a higher number of spectra should be more representative. The lower representativeness can also be related to the relatively poorer classification model performance observed for the KS-selected spectra. Likewise, the more representative spectra as selected by the Stratified sampling and Variogram sampling methods (with a lower absolute deviation in RMS) resulted in better performing classification models. Moreover, the Variogram method consistently selected fewer spectra.

Prediction maps, calculated from the model closest to the mean performance, are shown in the top row of Figure 3. Corresponding to the %CC results in Table 2, the Random, Grid and Variogram methods resulted in the lowest number of misclassified pixels. Three classes, i.e. sultana, cranberry and peanut, were all classified well, regardless of the sampling method used. The raisin class experienced some misclassified pixels, primarily at the edges; however, the KS sampling method was less prone

to such misclassifications. The almond and hazelnut samples were more difficult to classify, with misclassified pixels occurring both at the edge and within each sample. Compared with the other selection methods, KS experienced a higher percentage of misclassified pixels within the hazelnut sample.

From examination of the pixels selected by each method, as shown in the lower panels of Figure 3, it is

Table 3. Number of selected pixels per method for Dataset 1 (Nuts).

	Rand	KS	Strat	Grid	Var
Grid 5×5 *N* pixels (% pixels)	31 (0.2)	31 (0.2)	32 (0.2)	31 (0.2)	38 (0.3)
Grid 6×6 *N* pixels (% pixels)	39 (0.3)	39 (0.3)	40 (0.3)	39 (0.3)	46 (0.3)
Grid 7×7 *N* pixels (% pixels)	52 (0.4)	52 (0.4)	53 (0.4)	52 (0.4)	60 (0.4)
Grid 8×8 *N* pixels (% pixels)	65 (0.5)	65 (0.5)	66 (0.5)	65 (0.5)	72 (0.5)
Grid 9×9 *N* pixels (% pixels)	82 (0.6)	82 (0.6)	83 (0.6)	82 (0.6)	89 (0.6)
Grid 10×10 *N* pixels (% pixels)	98 (0.7)	98 (0.7)	99 (0.7)	98 (0.7)	104 (0.7)

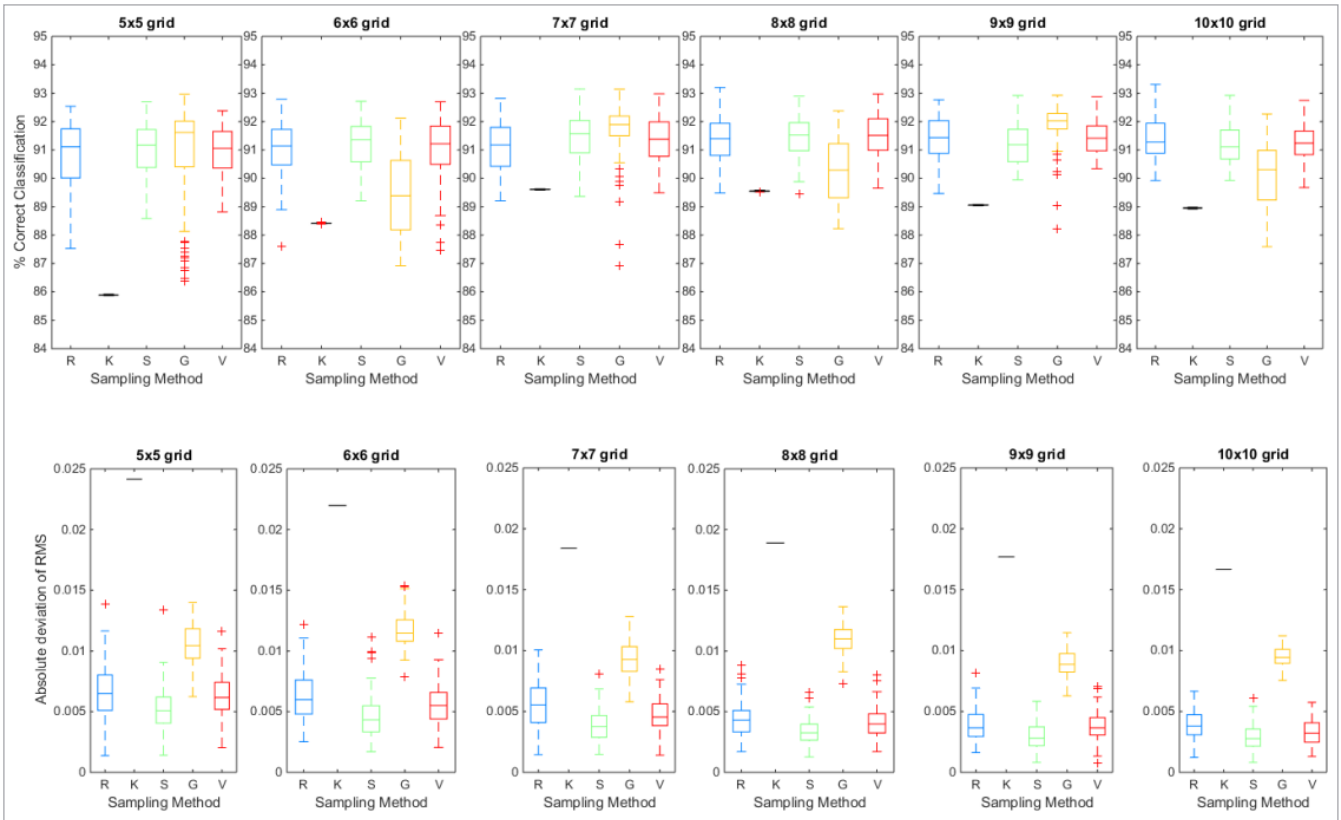

Figure 2. Boxplots showing variation in selection method performance indicators grouped by grid size for the "Nuts and dried fruits" dataset. The top row shows the % Correct classification of the 100 bootstraps of the selection method as calculated on the spectral test set, while the bottom row shows the absolute deviation of the RMS of the selected pixels from the global RMS of the image from which the pixels were sampled. The sampling method is shown on the x-axis in each subplot, where R = random, K = Kennard–Stone, S = Stratified Sampling, G = grid sampling, V = variogram sampling.

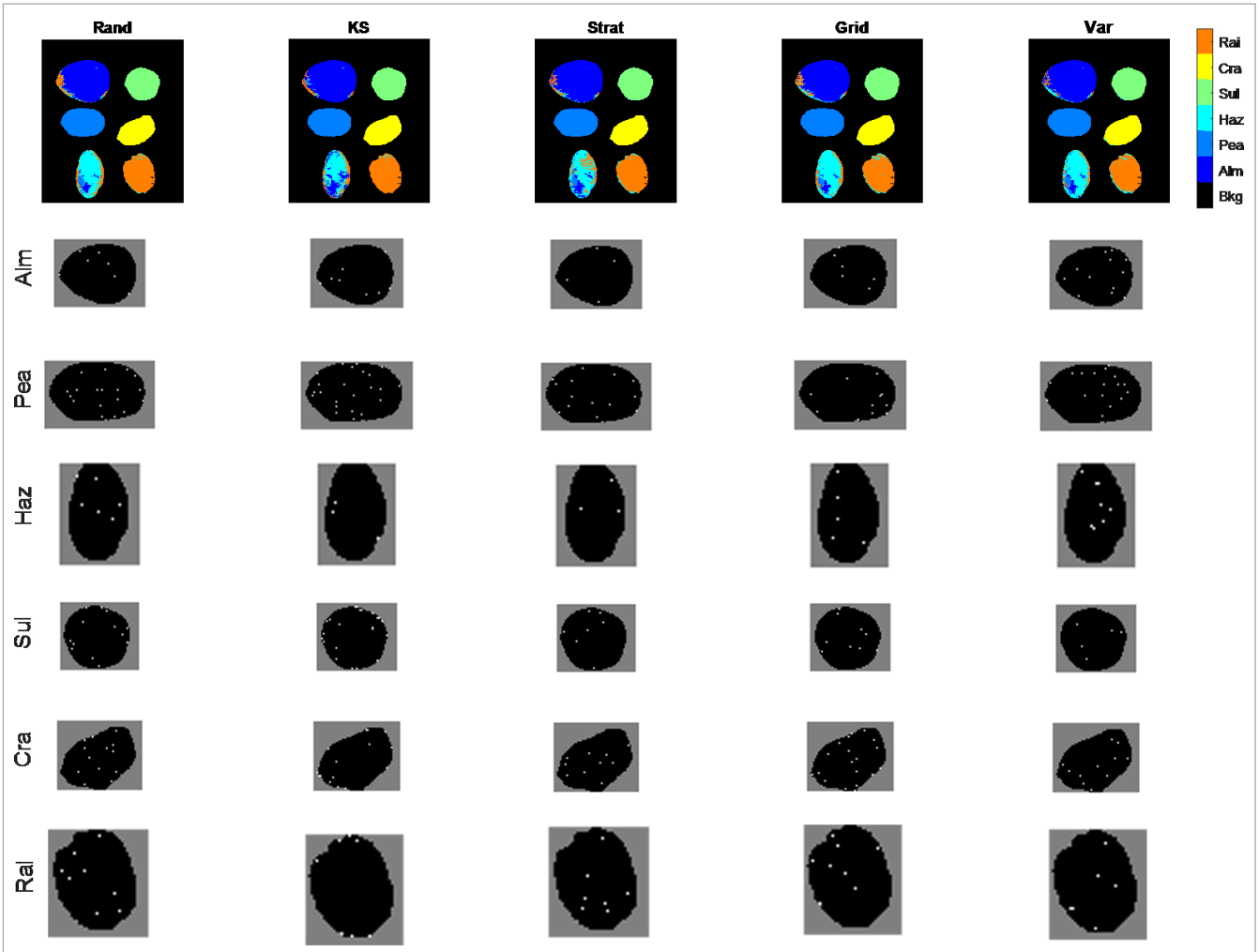

Figure 3. Prediction maps (top row) for each sampling method applied to the "Nuts and dried fruits" dataset. The PLS model closest to the mean model performance of 100 runs was selected for comparison of the different selection methods. The pixels selected from each class image and for each selection method are shown below the prediction maps.

clear that the KS method consistently selected edge pixels, as these are the pixels that produce the most different spectra, while the Random, Grid and Variogram methods provide a more regular distribution of selected pixels over the sample surface. This explains why the absolute deviation in RMS is higher for the KS-selected spectra and why the edge pixels of the raisin sample are better predicted, while the central regions of hazelnut are worse predicted, by the model built using KS-selected spectra.

Performance indicators for Dataset 2: Plastics

The average number and percentage of selected pixels per class for the plastics dataset is shown in Table 4. Again, the number of pixels selected increases with the grid size, while, again, a slightly higher number of pixels was chosen using the variogram method. As the grid size increased from 5×5 pixels to 10×10 pixels, the average number of pixels selected per class increased from 34–41 to 105–110.

The variation in spectral selection performance indicators over the 100 bootstraps, as shown in Figure 4, indicates that all selection strategies resulted in excellent model performance (>99 % CC), with the variation in % CC decreasing as the number of pixels selected increased. Similarly, the absolute deviation in RMS decreased as the number of pixels increased, and the trends in KS were similar to the previously discussed dataset. For all other selection methods, the number of spectra selected had a greater impact than the selection method. For the maximum number of selected spectra (10×10 grid in Figure 4), the Grid and Variogram

sampling methods resulted in the best overall model performance, lowest variation in model performance, and lowest absolute deviation in RMS. However, inspection of the prediction maps in Figure 5 indicates that the stratified sampling method resulted in the lowest number of misclassified pixels overall, agreeing with the highest G-mean in Table 2. The model built using KS-selected spectra had fewer edge pixels misclassi-

Table 4. Number of selected pixels and sampling times per method for Dataset 2 (plastics).

	Rand	KS	Strat	Grid	Var
Grid5 × 5 *N* pixels (% pixels)	34 (0.2)	34 (0.2)	35 (0.2)	34 (0.2)	41 (0.2)
Grid6 × 6 *N* pixels (% pixels)	40 (0.2)	40 (0.2)	41 (0.2)	40 (0.2)	46 (0.3)
Grid7 × 7 *N* pixels (% pixels)	59 (0.3)	59 (0.3)	61 (0.4)	59 (0.3)	66 (0.4)
Grid8 × 8 *N* pixels (% pixels)	70 (0.4)	70 (0.4)	71 (0.4)	70 (0.4)	76 (0.4)
Grid9 × 9 *N* pixels (% pixels)	93 (0.5)	93 (0.5)	95 (0.5)	93 (0.5)	98 (0.6)
Grid10 × 10 *N* pixels (% pixels)	105 (0.6)	105 (0.6)	106 (0.6)	105 (0.6)	110 (0.6)

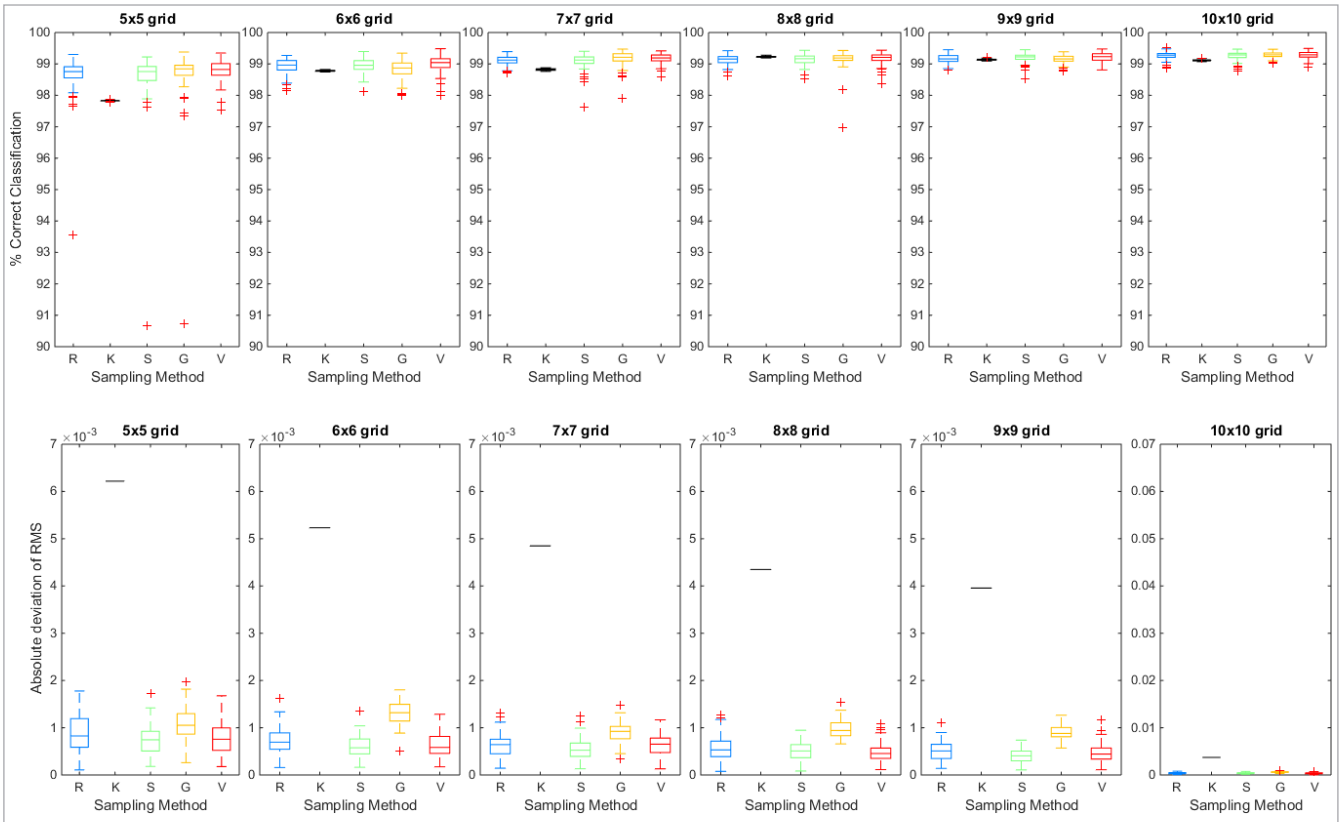

Figure 4. Boxplots showing variation in selection method performance indicators grouped by grid size for the "Plastics" dataset. The top row shows the % Correct classification of the 100 bootstraps of the selection method as calculated on the spectral test set, while the bottom row shows the absolute deviation of the RMS of the selected pixels from the global RMS of the image from which the pixels were sampled. The sampling method is shown on the x-axis in each subplot, where R = random, K = Kennard–Stone, S = Stratified Sampling, G = grid sampling, V = variogram sampling.

fied, while the models built using the Grid or Variogram approaches had fewer within object pixels misclassified. Again, this can be related back to the tendency of the KS method to select edge pixels (see lower panels of Figure 5).

Performance indicators for Dataset 3: Sweets

When comparing the number of pixels selected by each method applied to the "Sweets" dataset, as shown in Table 5, it is again clear that the number increased with gridsize, ranging from 1.3 % to 4.9 % of the total number

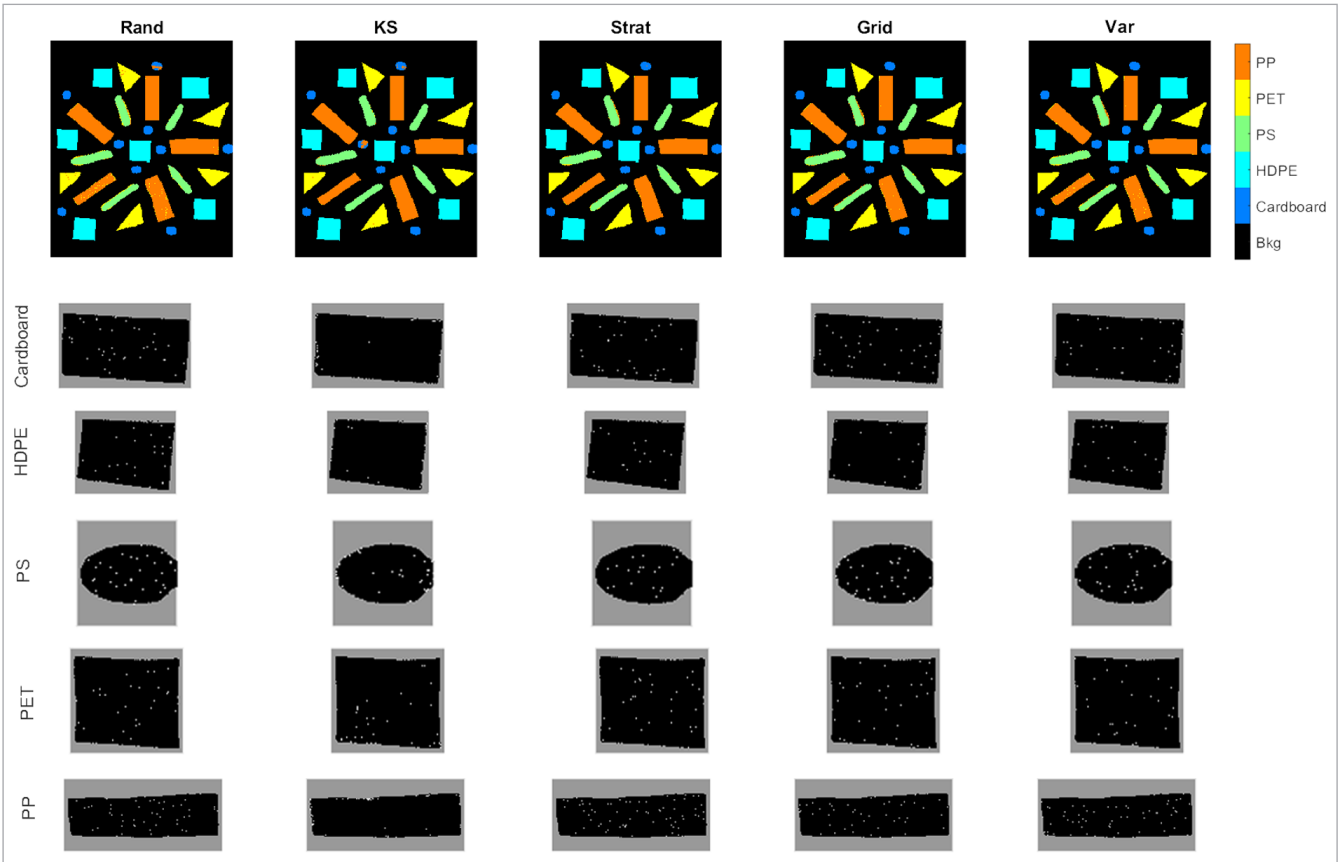

Figure 5. Prediction maps (top row) for each sampling method applied to the "Plastics" dataset. The PLS model closest to the mean model performance of 100 runs was selected for comparison of the different selection methods. The pixels selected from each class image and for each selection method are shown below the prediction maps.

Table 5. Number of selected pixels and sampling times per method for Dataset 3 (sweets).

	Rand	KS	Strat	Grid	Var
Grid 5×5 *N* pixels (% pixels)	31 (1.3)	31 (1.3)	32 (1.3)	31 (1.3)	44 (1.8)
Grid 6×6 *N* pixels (% pixels)	39 (1.6)	39 (1.6)	40 (1.7)	39 (1.6)	50 (2.1)
Grid 7×7 *N* pixels (% pixels)	56 (2.3)	56 (2.3)	57 (2.4)	56 (2.3)	70 (2.9)
Grid 8×8 *N* pixels (% pixels)	72 (3.0)	72 (3.0)	73 (3.0)	72 (3.0)	84 (3.5)
Grid 9×9 *N* pixels (% pixels)	89 (3.7)	89 (3.7)	90 (3.7)	89 (3.7)	98 (4.1)
Grid 10×10 *N* pixels (% pixels)	107 (4.4)	107 (4.4)	108 (4.5)	107 (4.4)	119 (4.9)

of pixels per class. In this case, similar to the previous dataset, the variogram method consistently selected a slightly higher number of pixels than the other methods, (44–119 vs 31–107), indicating the number of grids defined by the variogram method was higher than 5 × 5.

The variation in selection method performance indicators for this dataset (Figure 6) indicates again that variation and the number of outliers decreased as the number of spectra selected increased. In this case, the highest overall %CC and lowest variation in %CC over the 100 bootstraps was found for the Grid and Variogram methods. However, the absolute deviation in RMS was much higher for the Grid selection method than it was for the Variogram selection method, indicating that a more representative calibration set, in the sense of the variation from the mean as captured by the RMS statistic, may not always produce a better classification model. Observation of the prediction maps in Figure 7 again indicates that the Random, Grid and Variogram selection methods performed well, despite some misclassified edge pixels, while the KS and Stratified methods experienced a higher amount of misclassified pixels, especially in central regions of the "teeth" class. This can be understood by observation of the selected pixels from each dataset, as shown in the lower panels of Figure 7, indicating a better distribution of selected pixels over the sample surface for the Random, Grid and Variogram methods than for the spectral-based KS or Stratified methods.

Performance indicators for Dataset 4: Cereals

In terms of the number of pixels selected by each method (Table 6), similarly to the Plastics and Sweets datasets, the number of pixels selected by the Variogram method was higher than that selected by the other methods (34–104 vs 28–99). Varying the number of grids from 5 × 5 to 10 × 10 increased the percentage of selected spectra from just over 1% to just over 4%.

When considering the boxplots, showing the variation in selection performance metrics over the 100 bootstraps (Figure 8), KS was by far the best performing

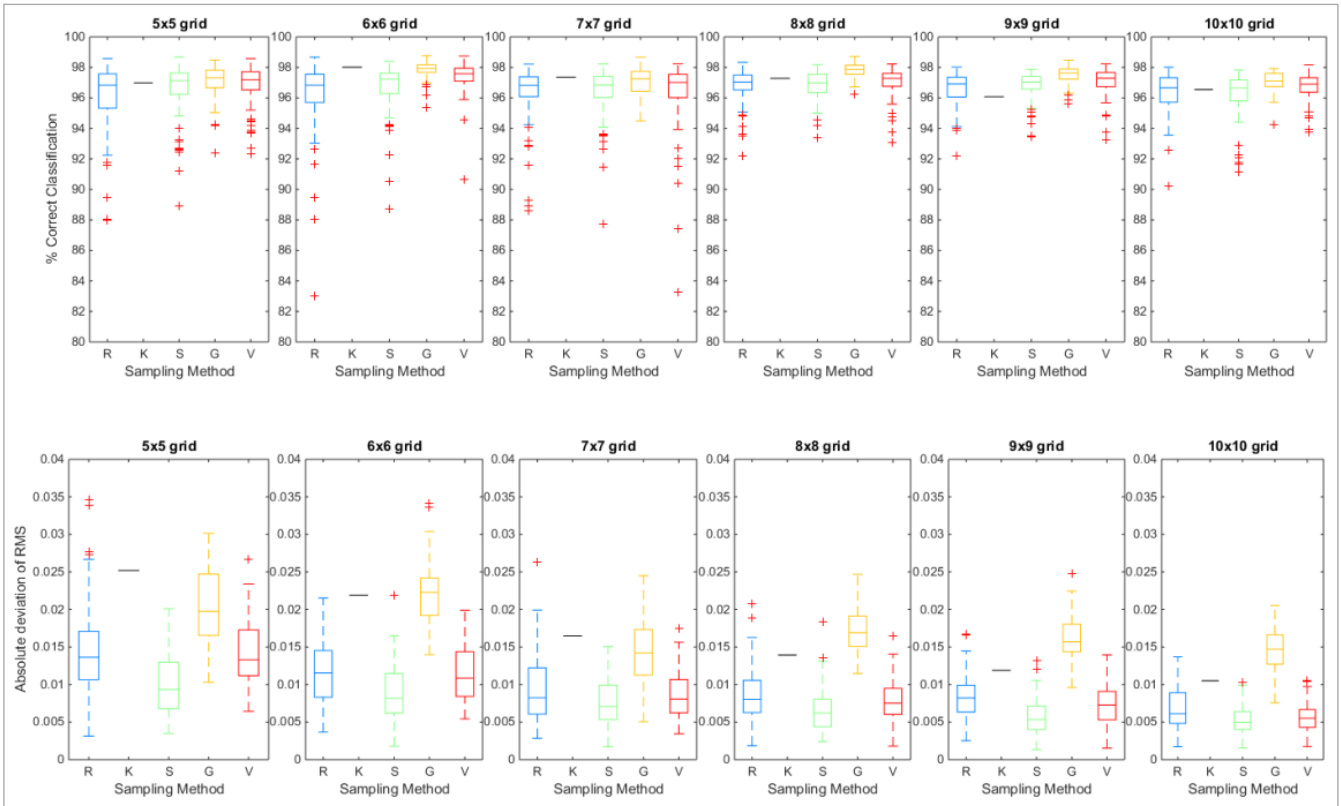

Figure 6. Boxplots showing variation in selection method performance indicators grouped by grid size for the "Sweets" dataset. The top row shows the %Correct classification of the 100 bootstraps of the selection method as calculated on the spectral test set, while the bottom row shows the absolute deviation of the RMS of the selected pixels from the global RMS of the image from which the pixels were sampled. The sampling method is shown on the x-axis in each subplot, where R = random, K = Kennard–Stone, S = Stratified Sampling, G = grid sampling, V = variogram sampling.

selection method, resulting in the highest %CC (>90% as compared to 86–88% for the other methods). Similar to the other datasets, the absolute deviation in RMS was lowest for the Stratified and Variogram methods, however, this did not result in a better classification model performance. When considering the spectra selected by each method (Figure 9, lower panels), it appears that the KS method selected more spectra

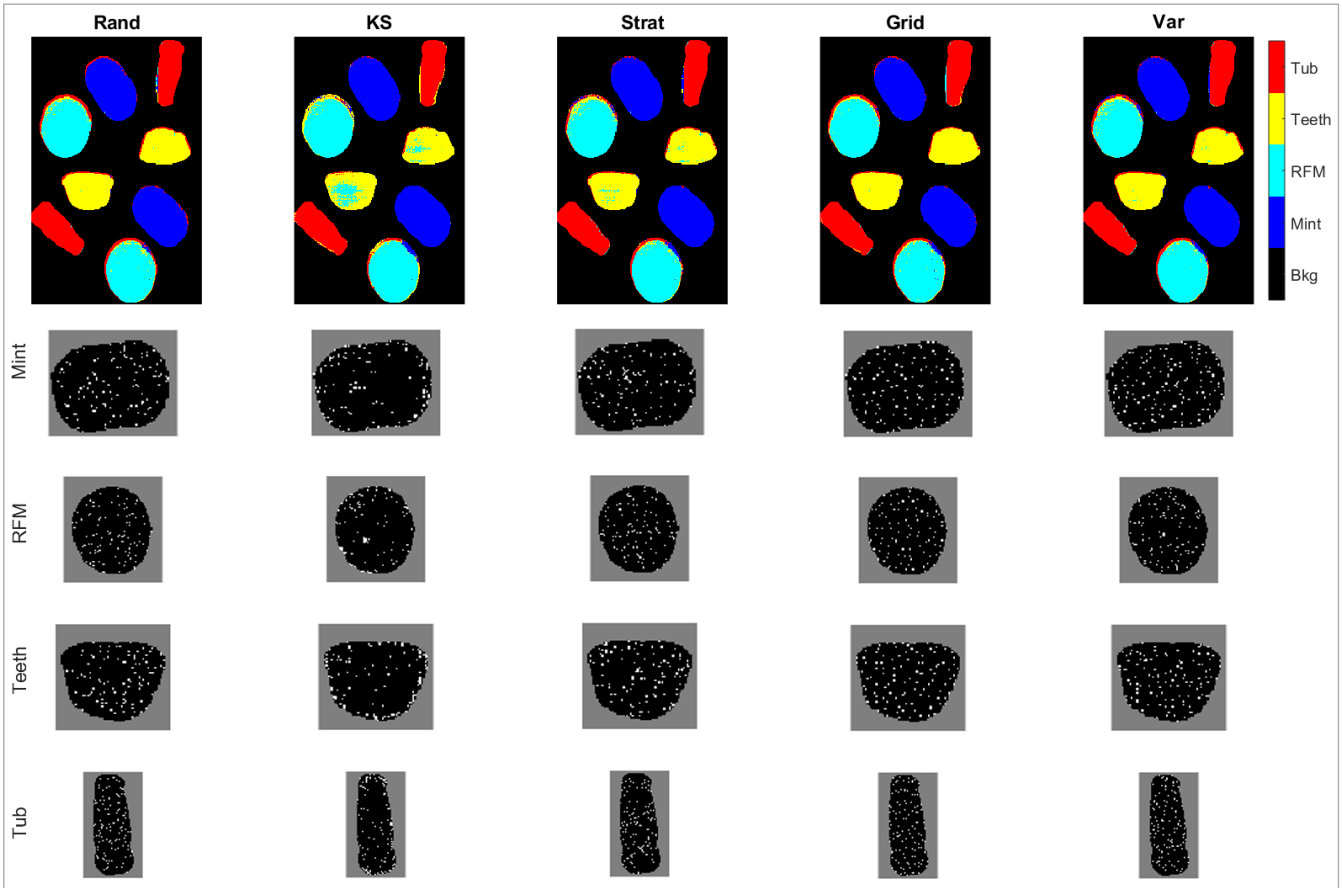

Figure 7. Prediction maps (top row) for each sampling method applied to the "Sweets" dataset. The PLS model closest to the mean model performance of 100 runs was selected for comparison of the different selection methods. The pixels selected from each class image and for each selection method are shown below the prediction maps.

Table 6. Number of selected pixels and sampling times per method for Dataset 4 (cereal).

	Rand	KS	Strat	Grid	Var
Grid 5×5 *N* pixels (% pixels)	28 (1.2)	28 (1.2)	30 (1.2)	29 (1.2)	34 (1.4)
Grid 6×6 *N* pixels (% pixels)	40 (1.7)	40 (1.7)	42 (1.7)	40 (1.7)	47 (2.0)
Grid 7×7 *N* pixels (% pixels)	55 (2.3)	55 (2.3)	56 (2.3)	55 (2.3)	61 (2.5)
Grid 8×8 *N* pixels (% pixels)	66 (2.7)	66 (2.7)	67 (2.8)	66 (2.7)	72 (3.0)
Grid 9×9 *N* pixels (% pixels)	88 (3.7)	88 (3.7)	90 (3.7)	88 (3.7)	95 (3.9)
Grid 10×10 *N* pixels (% pixels)	98 (4.1)	98 (4.1)	99 (4.1)	98 (4.1)	104 (4.3)

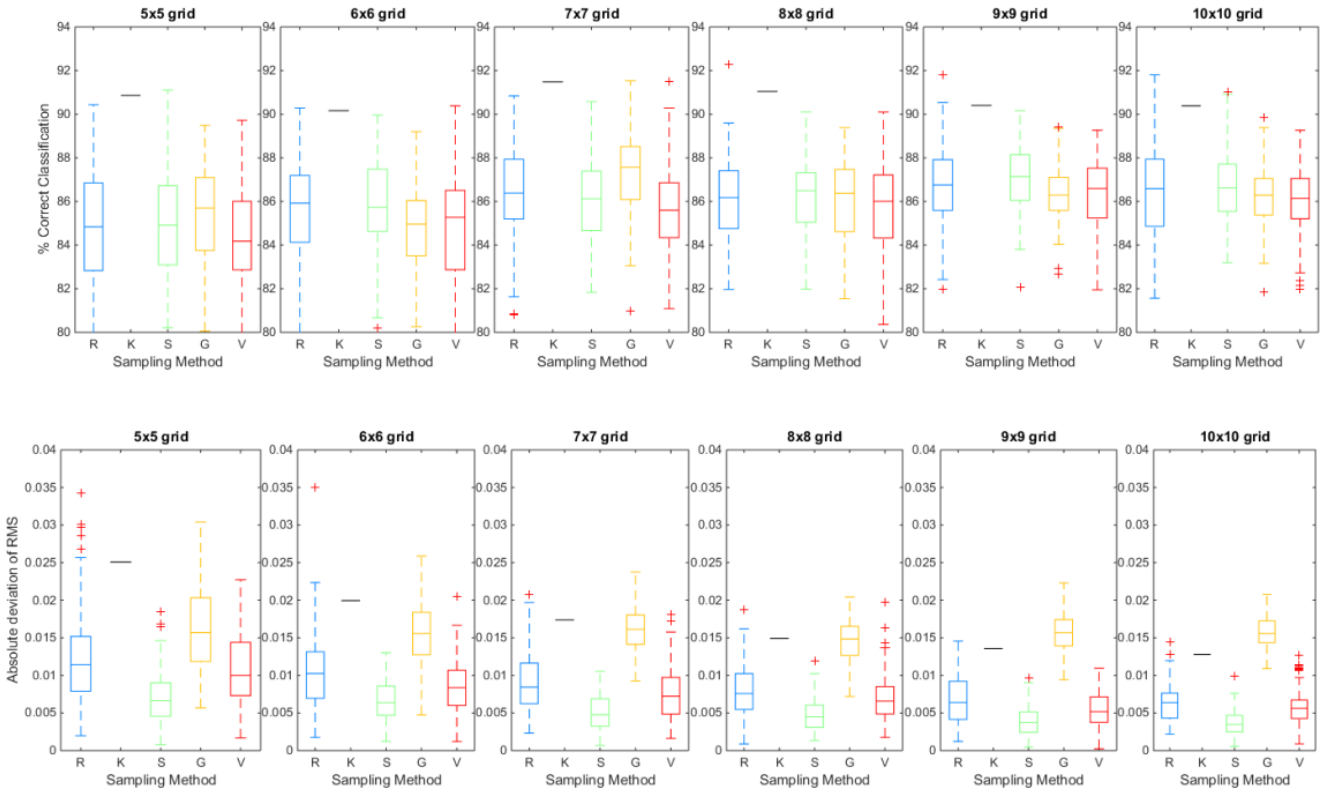

Figure 8. Boxplots showing variation in selection method performance indicators grouped by grid size for the "Cereals" dataset. The top row shows the % Correct classification of the 100 bootstraps of the selection method as calculated on the spectral test set, while the bottom row shows the absolute deviation of the RMS of the selected pixels from the global RMS of the image from which the pixels were sampled. The sampling method is shown on the x-axis in each subplot, where R = random, K = Kennard–Stone, S = Stratified Sampling, G = grid sampling, V = variogram sampling.

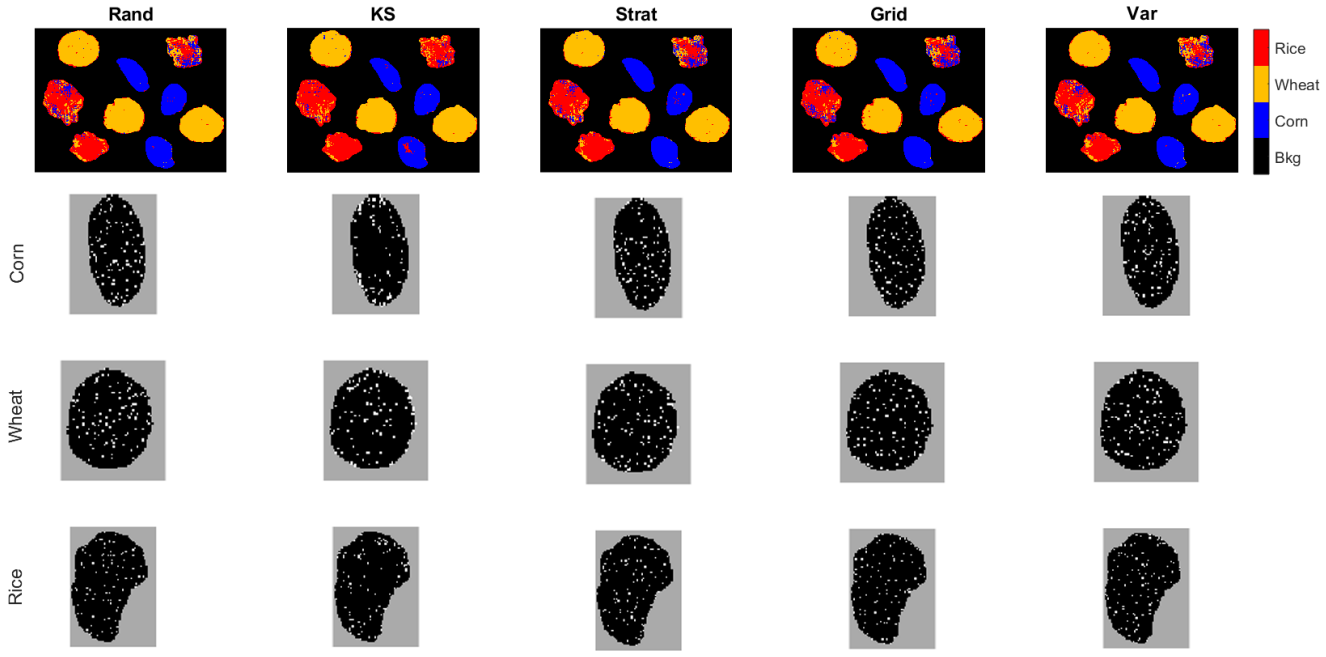

Figure 9. Prediction maps (top row) for each sampling method applied to the "Sweets" dataset. The PLS model closest to the mean model performance of 100 runs was selected for comparison of the different selection methods. The pixels selected from each class image and for each selection method are shown below the prediction maps.

within the classes than in previous instances. Perhaps this and the improved performance of the KS method can be understood when considering the morphology of the samples included in the cereal dataset, as compared to the others. While the other samples were relatively smooth surfaces, the cereals had rough edges within the samples, which provide sufficient spectral variability for selection by KS.

Table 7. Number of selected pixels and sampling times per method for Dataset 4 (paints).

	Rand	KS	Strat	Grid	Var
Grid 5×5 *N* pixels (% pixels)	36 (0.9)	36 (0.9)	37 (0.9)	36 (0.9)	42 (1.1)
Grid 6×6 *N* pixels (% pixels)	50 (1.3)	50 (1.3)	51 (1.3)	50 (1.3)	57 (1.4)
Grid 7×7 *N* pixels (% pixels)	66 (1.7)	66 (1.7)	67 (1.7)	66 (1.7)	72 (1.8)
Grid 8×8 *N* pixels (% pixels)	80 (2.0)	80 (2.0)	81 (2.1)	80 (2.0)	86 (2.2)
Grid 9×9 *N* pixels (% pixels)	108 (2.7)	108 (2.7)	109 (2.8)	108 (2.7)	115 (2.9)
Grid 10×10 *N* pixels (% pixels)	125 (3.2)	125 (3.2)	126 (3.2)	125 (3.2)	131 (3.3)

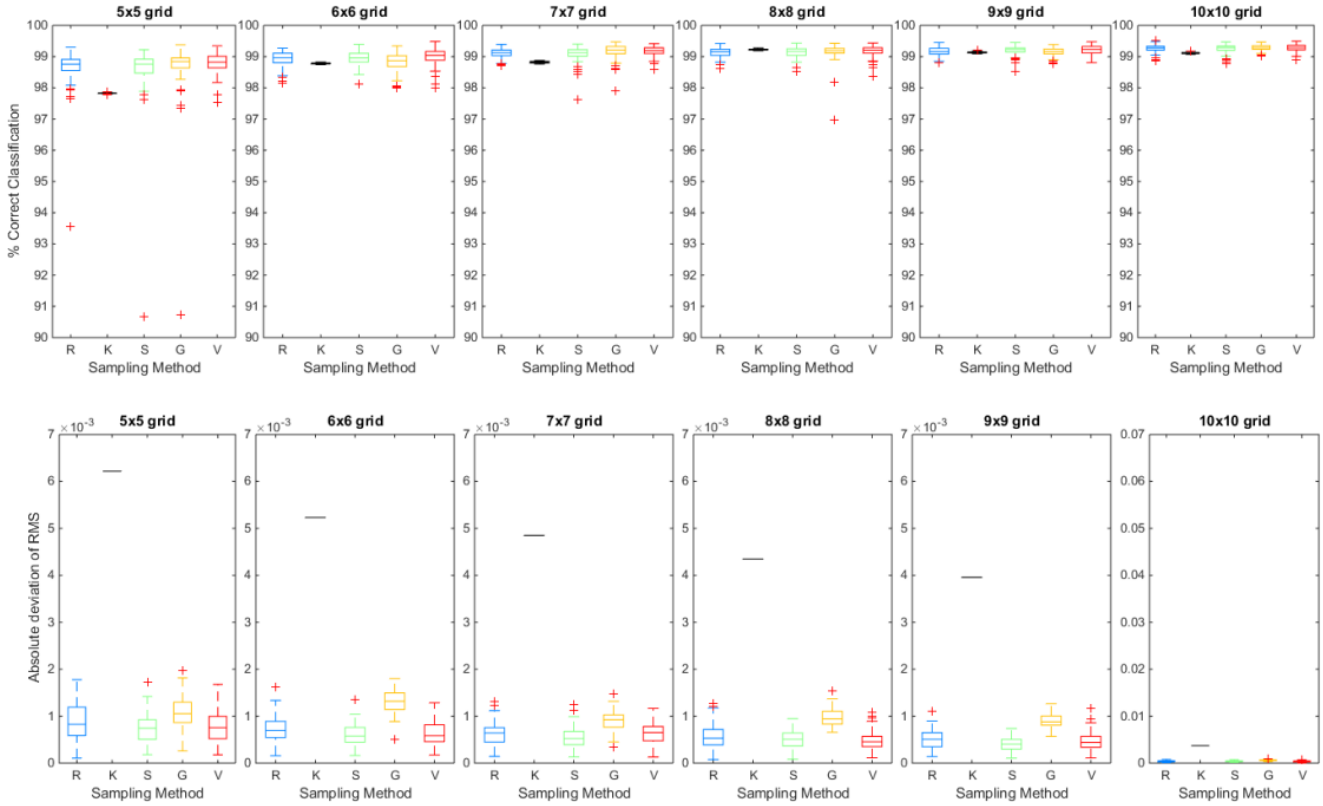

Figure 10. Boxplots showing variation in selection method performance indicators grouped by grid size for the "Paint" dataset. The top row shows the % Correct classification of the 100 bootstraps of the selection method as calculated on the spectral test set, while the bottom row shows the absolute deviation of the RMS of the selected pixels from the global RMS of the image from which the pixels were sampled. The sampling method is shown on the x-axis in each subplot, where R = random, K = Kennard–Stone, S = Stratified Sampling, G = grid sampling, V = variogram sampling.

Performance indicators for Dataset 5: Paints

Similarly to the plastics, sweets and cereals datasets, the selection of spectra from the paint samples resulted in a higher number of pixels selected using the Variogram method (42–131 vs 36–125), representing around 1 % (for 5 × 5 grids) and 3 % (for 10 × 10 grids) of the total amount of pixels in each class image (Table 7).

In this instance, the boxplots showing the variation in selection performance indicators (Figure 10) exhibit a relatively high number of outliers for the Random, Stratified and Grid methods when 1 % of pixels are selected (corresponding to 5 × 5 grid in Figure 10). However, the number of outliers decreased as the number of selected pixels increased, indicating greater model stability. At the highest number of selected pixels (approximately 4 %, corresponding to the 10 × 10 grid in Figure 10), the selection methods all produced similar models in terms of % CC, all with relatively low variation in % CC, indicating that these samples had little spatial variability in the spectra—i.e. they were relatively homogeneous compared to the "Nuts and dried fruits" or "Sweets" datasets discussed previously. Similarly, the absolute deviation in RMS decreased both in magnitude and variation as the number of selected pixels increased. Inspection of the prediction maps indicates substantial misclassification around the edge regions of the samples and confu-

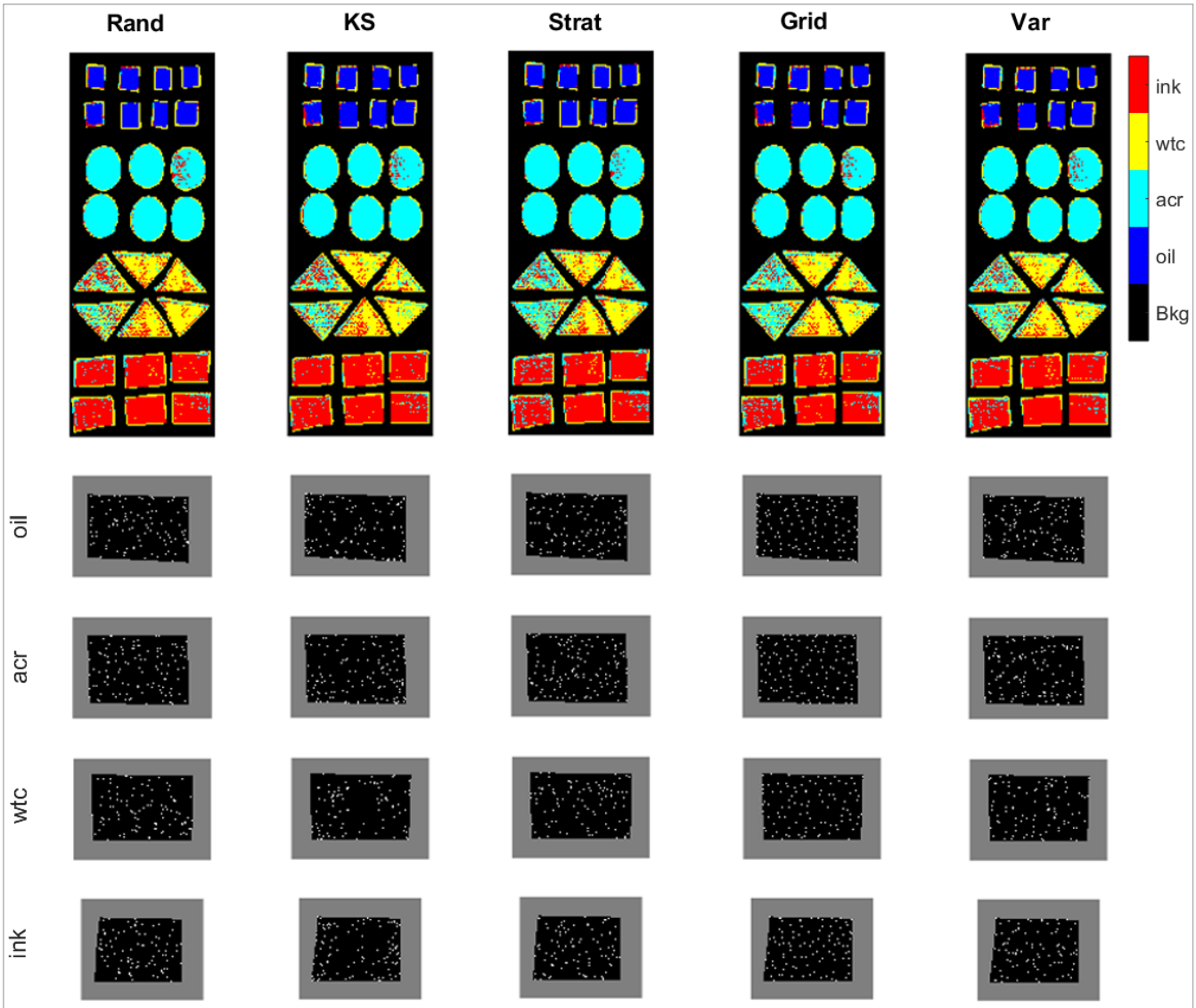

Figure 11. Prediction maps (top row) for each sampling method applied to the "Paint" dataset. The PLS model closest to the mean model performance of 100 runs was selected for comparison of the different selection methods. The pixels selected from each class image and for each selection method are shown below the prediction maps.

sion between the Acr, WTC and Ink classes. This explains the substantial decrease in %CC observed when the models were applied to the test image, as seen in Table 2.

Conclusions

In this study, several sampling methodologies were compared for building classification models from hyperspectral imaging data. In terms of time and model performance, random sampling emerged as the optimal strategy. However, the variation model performance indicators over 100 bootstraps in the random selection was generally higher than in any of the other sampling methods. The use of variographic analysis in the selection of a gridsize for image-plane-based spectral selection resulted in a more representative subset selection, while also reducing the variation in classification error over 100 bootstraps. Considering the significantly lower time required for the random selection, however, the improvements in model performance obtained using the Grid or Variogram methods are somewhat modest. The results show that the optimal selection method depends on the characteristics of the sample studied. For a homogeneous sample, random selection presented a good balance between classification error and time required for selection. However, for a smooth heterogeneous sample, the Grid or Variogram methods performed better. The Kennard–Stone selection method did not perform well in most cases, due to the selection of extreme edge pixels. However, in the case of a sample with rough edges in its interior (i.e. the cereal samples) the Kennard–Stone method performed very well. As hyperspectral images provide additional spatial information not available in conventional spectroscopic data, the use of variographic analysis as a preliminary step to understand sample morphology prior to the decision on a spectral sampling technique is recommended.

Acknowledgements

Funding for this research was provided by the European Research Council (ERC) under the starting grant programme ERC-2013-StG call—Proposal No. 335508—BioWater; and Science Foundation Ireland (SFI) under the investigators programme Proposal ID 15/IA/2984—HyperMicroMacro.

References

1. B.. Weinstock, J. Janni, L. Hagen and S. Wright, "Prediction of oil and oleic acid concentrations in individual corn (*Zea mays* L.) kernels using near-infrared reflectance hyperspectral imaging and multivariate analysis", *Appl. Spectrosc.* **60,** 9 (2006). https://doi.org/10.1366/000370206775382631
2. G. ElMasry, N. Wang, A. ElSayed and M. Ngadi, "Hyperspectral imaging for nondestructive determination of some quality attributes for strawberry", *J. Food Eng.* **81,** 98 (2007). https://doi.org/10.1016/j.jfoodeng.2006.10.016
3. J. Zhao, S. Vittayapadung, Q. Chen, S. Chaitep and R. Chuaviroj, "Nondestructive measurement of sugar content of apple using hyperspectral imaging technique", *Maejo Int. J. Sci. Technol.* **3,** 130 (2009).
4. M. Taghizadeh, A. Gowen, P. Ward and C.P. O'Donnell, "Use of hyperspectral imaging for evaluation of the shelf-life of fresh white button mushrooms (*Agaricus bisporus*) stored in different packaging films", *Innov. Food Sci. Emerg. Technol.* **11,** 423 (2010). https://doi.org/10.1016/j.ifset.2010.01.016
5. M. Daszykowski, B. Walczak and D.L. Massart, "Representative subset selection", *Anal. Chim. Acta* **468(1),** 91 (2002). https://doi.org/10.1016/S0003-2670(02)00651-7
6. R. Kennard and L. Stone, "Computer aided design of experiments", *Technometrics* **11,** 137 (1969). https://doi.org/10.1080/00401706.1969.10490666
7. R.D. Clark, "Optisim: an extended dissimilarity selection method for finding diverse representative subsets", *J. Chem. Inf. Comput. Sci.* **37,** 1181 (1997). doi: https://doi.org/10.1021/ci970282v
8. P.M. Gy, *Sampling for Analytical Purposes*. Translated by A.G. Royle (1998). ISBN: 978-0471979562
9. C. Nansen, A.J. Sidumo and S. Capareda, "Variogram analysis of hyperspectral data to characterize the impact of biotic and abiotic stress of maize plants and to estimate biofuel potential", *Appl. Spectrosc.* **64,** 627 (2010). https://doi.org/10.1366/000370210791414272
10. A.A. Gowen, F. Marini, C. Esquerre, C. O'Donnell, G. Downey and J. Burger, "Time series hyperspectral chemical imaging data: challenges, solutions and applications", *Anal. Chim. Acta* **705,** 272 (2011). https://doi.org/10.1016/j.aca.2011.06.031

11. M. Taghizadeh, A.A. Gowen and C.P. O'Donnell, "The potential of visible-near infrared hyperspectral imaging to discriminate between casing soil, enzymatic browning and undamaged tissue on mushroom (*Agaricus bisporus*) surfaces", *Comput. Electron. Agric.* **77,** 74 (2011). https://doi.org/10.1016/j.compag.2011.03.010

12. M.L. Martínez, A. Garrido-Varo, E. De Pedro and L. Sánchez, "Effect of sample heterogeneity on near infrared meat analysis: the use of the RMS statistic", *J. Near Infrared Spectrosc.* **6,** 313 (1998). https://doi.org/10.1255/jnirs.214

13. R.G. Brereton and G.R. Lloyd, "Partial least squares discriminant analysis for chemometrics and metabolomics: how scores, loadings, and weights differ according to two common algorithms", *J. Chemometr.* **32,** 1 (2018). https://doi.org/10.1002/cem.3028