Automated quantification of defective maize kernels by means of multivariate image analysis

Giorgia Orlandi, Rosalba Calvini, Giorgia Foca, Alessandro Ulrici*

Department of Life Sciences and Interdepartmental Research Centre BIOGEST-SITEIA, University of Modena and Reggio Emilia, Padiglione Besta, Via Amendola 2, 42122, Reggio Emilia, Italy

Corresponding author: alessandro.ulrici@unimore.it

Abstract

This article describes the development of a fast and inexpensive method based on digital image analysis for the automated quantification of the percentage of defective maize (%DM). Defective kernels tend to foster high levels of mycotoxins like Deoxynivalenol (DON), which represents a risk for the health of humans and of farm animals. In this work, 332 RGB images of 83 mixtures containing different amounts of defective maize kernels were acquired using a digital camera. The mixtures were also analysed with a commercial ELISA test kit to determine their concentration of DON, that resulted highly correlated with the amount of defective kernels. Each image was then converted into a signal, named *colourgram*, which codifies its colour-related information content. The colourgrams were firstly explored using Principal Component Analysis. Then, calibration models of the %DM values were developed using Partial Least Squares (PLS) and interval-PLS. The best interval-PLS model allowed to predict the %DM values of external test set samples with a root mean square error value equal to 2.6%. Based on the output of this model it was also possible to highlight the defective-maize areas within the images, confirming the significance of the proposed approach.

Keywords:
Maize; Defect detection; Mycotoxins; Multivariate Image Analysis; Multivariate calibration

1. Introduction

The great importance of maize (*Zea mays L.*) is due to its primary role for multiple uses, including human food, livestock feed, biofuels and bioplastics (FAO, 2006). A current issue of high relevance related to the consumption of maize as food or feed is its possible contamination with mycotoxins. Indeed, maize mycotoxins can be directly found in the human food or, as animal feed, they are ingested by animals and then pass to humans through the food chain. Due to their high toxicity, mycotoxins represent a major risk for human health. Their ingestion can lead to a wide range of effects, including deterioration of liver or kidney functions, skin necrosis, immunological disturbances, neurotoxicity and carcinogenicity (Steyn, 1995; Sweeney and Dobson, 1998; Edite Bezerra da Rocha et al., 2014).

Mycotoxins are secondary metabolites naturally produced by some filamentous fungi, which frequently develop in maize. The most common mycotoxins in maize are produced by fungi belonging to the genera of *Fusarium*, *Aspergillus* and *Penicillium* (Hossain and Goto, 2014). These microorganisms mainly develop in the field or at the post-harvest stage, when storage conditions are inadequate. The types and levels of contamination strongly depend on the contaminant fungi species, on the harvesting year and on environmental conditions such as temperature and humidity (Suleiman et al., 2013). One of the most common mycotoxins found in maize is deoxynivalenol (DON), also known as vomitoxin due to its strong emetic effects. DON is primarily produced by *Fusarium graminearum* and *Fusarium culmorum*, and is one of the most common mycotoxins found in maize (Kushiro, 2008; Sobrova et al., 2010; Edite Bezerra da Rocha et al., 2014). Because of the health hazards to humans and animals, the European Parliament has set a limit of 1750 ppb in the unprocessed maize used in foodstuff (Commission Regulation (EC) No 1126/2007), while in the animal feed materials the recommendations generally suggest to not exceed 8 ppm of DON (Commission Directive 2003/100/EC).

In order to ensure food safety, proper techniques to estimate the concentration of mycotoxins in maize have been developed, which are mainly based on chromatographic methods and on

54 immunoassays (Maragos and Busman, 2010). These methods allow to gain high sensitivity and
55 specificity, but present some drawbacks, mainly due to the relatively long times required for the
56 analysis, to the costs and to the limited amount of analysed sample, which implies the risk of a
57 poorly representative sampling of large maize batches. These aspects are particularly crucial during
58 the transfer phase of the maize crops to the warehouse, when it is necessary to evaluate in very short
59 times large amounts of product conferred by farmers, in order to fix the price and the final
60 destination of each batch, or to reject it.
61 In this context, the availability of proper systems to perform a fast analysis of representative maize
62 quantities might constitute a very useful tool, at least for a preliminary assessment in view of more
63 refined analyses of the accepted batches by traditional wet chemistry methods. To this aim, digital
64 image processing is suitable for screening heterogeneous food or feed matrices like maize, to detect
65 local defects connected to fungal and toxin contaminations (Udomkun et al., 2017). Some authors
66 have recently proposed the use of near infrared hyperspectral imaging (NIR-HSI) to detect maize
67 kernels infected with fungi, and to estimate the degree of infection with a fast and accurate system
68 (Del Fiore et al., 2010; Singh et al., 2012; Williams et al., 2012). Notwithstanding the great
69 advantages offered by NIR-HSI, the efforts needed to efficiently extract useful information from the
70 huge amount of hyperspectral data and the relatively high cost of hyperspectral cameras are still
71 limiting factors for its widespread application in maize monitoring (Ferrari et al., 2013; Ulrici et al.,
72 2013; Calvini et al., 2016).
73 For these reasons, much cheaper instrumentations based on the use of common digital cameras
74 constitute an interesting alternative for the implementation of fast and non-destructive methods to
75 monitor maize defects. In fact, although the lack of visible defects does not ensure the complete
76 absence of mycotoxins, the presence of stained, dark or rotten maize kernels is generally correlated
77 with the presence of fungal infections. In other words, the higher the amount of defective kernels,
78 the higher the possibility of significant mycotoxins contamination.

In this context, the use of Multivariate Image Analysis (MIA) offers a wide range of effective tools to properly detect and quantify visible defects through RGB imaging. Essentially MIA consists in the development and application of various chemometric strategies for the analysis of multivariate images, consisting of a given number of picture elements (pixels), each one characterized by a series of spectral variables, or channels (Esbensen and Geladi, 1989; Geladi and Grahn, 1996; Prats-Montalbán et al., 2011; Duchesne et al., 2012; Reis, 2014). Many approaches have been proposed to characterise food samples based on MIA applied to RGB images, by using information in the original RGB colour space, in the latent variable space (e.g., using PCA), or in other colour spaces like Hue, Saturation, Intensity (HSI) (Yu et al., 2003; Pereira et al., 2009; Pierini et al., 2016).

In particular, many research works have been reported in the literature, where morphological, textural and colour features extracted from RGB images were used to develop automated systems for monitoring damaged and non-damaged kernels (Ruan et al., 1998; Choudhary et al., 2008). Valeinte-Gonzàlez et al. (2014) developed an effective approach based on the combined use of computer vision and Principal Component Analysis (PCA) to identify the damaged regions of single maize kernels. However, in the perspective of an industrial application, the determination of the degree of defectiveness of a maize batch based on the investigation of single kernels would be too demanding in terms of time and computational effort.

In this context, this study was aimed at developing an automated system for a preliminary assessment of DON contamination, based on the simultaneous analysis of a dataset of RGB images of mixtures containing different percentages of defective maize (%DM). The correlation between the %DM values and the concentration of DON, estimated by means of a commercial ELISA test kit, was also investigated. Each image was converted into a one-dimensional signal, named *colourgram*, which codifies its colour-related information content (Antonelli et al., 2004; Lo Fiego et al., 2007; Foca et al., 2011; Ulrici et al., 2012). In turn, the colourgrams were used to develop calibration models to predict %DM, using Partial Least Squares (PLS) and the feature selection algorithm interval-Partial Least Squares (iPLS). Moreover, the reconstruction of the maize images

105 considering the colour-related features selected by iPLS allowed to visualize the defective kernels
106 and thus to evaluate in a critical manner the choices made automatically by the algorithm.
107
108 **2. Materials and methods**
109 *2.1 Maize samples*
110 In the present study, two different types of maize kernels were considered: dry maize (13 %
111 moisture) and wet maize (24 % moisture). For both the maize types, based on their visual aspect the
112 kernels were manually separated into defective (stained, dark or rotten) and non-defective (uniform
113 yellow pericarp) kernels (Nguyen V.H., 2013). After the separation between defective and non-
114 defective kernels, the maize samples were sealed in plastic bags and stored in the dark at 4 °C for a
115 maximum of two days before analysis.
116
117 *2.2 Image acquisition*
118 The RGB images were acquired using a Panasonic DMC-TZ25 digital camera, using a 24 mm focal
119 length (in 35 mm equiv.), 1/125 s shutter speed, ISO-100 and f/3.5. Before the acquisition sessions,
120 white balance was set to a constant value by pointing the camera towards a white paper sheet, under
121 the same lighting conditions used to capture sample images. The images, with 24-bit colour depth
122 and spatial resolution equal to 4000 × 3000 pixels (corresponding to an image area approximately
123 equal to 310 × 235 mm) were stored in JPEG format, with an average file size equal to 4.87 MB.
124 In order to have constant and homogeneous lighting conditions, the camera was mounted on a
125 carton box (Figure 1) whose inner surface was covered with white paper sheets. The lighting system
126 consisted in a strip of white light-emitting diode lamps (SMD 3528 LED 5V USB, colour
127 temperature 6500 K) assembled on a metallic support and directed upwards at a 90 degree angle
128 with the carton box wall. In this manner, the sample was only illuminated by diffused light to avoid
129 the presence of undesired shadows or reflection effects. A white conveyor belt was used as
130 background of the images. Furthermore, a colour reference was included in the image scene to

131 correct possible variations in the lighting conditions. The colour reference, reported in Figure 2,
132 consisted in a white paper sheet including a series of eight squares (size 1 cm^2) of different colours,
133 i.e., white, black, the primary additive colours (red, green, blue) and the primary subtractive colours
134 (cyan, magenta, yellow), that were obtained using a laser printer (HP LaserJet Pro 200 color MFP
135 M276n).
136 The acquisition of the RGB images was performed in two subsequent steps. A schematic
137 representation of the procedure followed for the preparation of the imaged samples is reported in
138 Figure 3.
139 In the first acquisition step, the amount of defective maize kernels was varied considering 13
140 different levels, corresponding to the following percentages by weight (w/w): 0%, 5%, 10%, 20%,
141 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95% and 100%. In particular, for dry maize two mixtures
142 were prepared for each considered level, that were split in the two mixture groups D1a and D1b.
143 For wet maize, due to a smaller amount of available defective maize, only one mixture group (W1)
144 was considered. Therefore, in the first step 26 samples of dry maize and 13 samples of wet maize
145 were considered, for a total of 39 different samples. Each mixture consisted in a total amount of
146 maize kernels equal to 150 g.
147 For each sample two images were acquired, shuffling the kernels before each acquisition. The same
148 image acquisition procedure was repeated in a different day in order to check the day-to-day
149 variability. All the samples were acquired in random order to minimise, as much as possible, the
150 effects of uncontrollable factors. Therefore, on the whole, 156 images (= 39 samples × 2 repeated
151 acquisitions × 2 measurement sessions) were obtained in this first step.
152 Simultaneously, for each mixture the concentration of DON was determined using a commercially
153 available ELISA kit (see Section 2.3). The results of this analysis showed that mixtures with an
154 amount of defective kernels equal to or greater than 10% presented high concentrations of DON.
155 For this reason, a second acquisition step was planned with the aim of collecting additional images

with an amount of defective kernels in the range between 0 and 10%, in order to improve the performance of the calibration models in proximity to low %DM values.

Therefore, in the second acquisition step, additional mixtures were prepared, considering 11 levels corresponding to percentages by weight (w/w) of defective maize kernels ranging from 0% to 10%, with steps of 1%. In this case, for both dry and wet maize types two mixtures for each considered level were prepared, that were split in the D2a and D2b mixture groups for dry maize, and in the W2a and W2b mixture groups for wet maize. Therefore, in the second step a total of 44 samples were obtained, each one containing 150 g of kernels.

The experimental procedure followed for image acquisition was the same for the previous step, leading to 176 additional images (= 44 samples × 2 repeated acquisitions × 2 measurement sessions). The digital images acquired during the two acquisition steps were merged together to obtain a final dataset composed of 332 images.

2.3 Determination of deoxynivalenol

In order to verify the correlation between the %DM values and the concentration of DON, ELISA test was performed on the maize samples using AgraQuant® DON 0.25/5.0 Assay (Romer Labs Inc., USA).

Before the first acquisition step, the concentration of DON was determined on the defective and non-defective kernels of wet and dry maize. In particular, for dry maize 10 aliquots of non-defective kernels and 10 aliquots of defective kernels were randomly collected, while for wet maize 5 aliquots were collected for each group of kernels. Each aliquot consisted in 20 g of maize kernels, which were ground and subjected to ELISA test following the standard procedure provided by the manufacturer.

Afterwards, simultaneously with image acquisition, the ELISA test was performed also on the mixtures imaged during both the acquisition steps. In this case, the entire amount of 150 g of kernels of each sample was grinded and analysed with AgraQuant kit. The quantification range of

182 the AgraQuant DON assay is between 0.25 and 5.0 ppm; therefore, samples containing DON levels
183 higher than 5 ppm were diluted with deionized water in order to fall within the quantification range.
184 The dilution was performed up to a maximum quantification value equal to 10 ppm.
185
186 *2.4 Image analysis*
187 *2.4.1 Standardization and conversion to colourgrams*
188 The key steps followed for the elaboration of the RGB images are summarized in Figure 2. Firstly,
189 from each original image the two areas corresponding to the reference (coloured squares) and to the
190 sample (maize kernels) were automatically selected and stored as separate images. The size of the
191 obtained images was equal to 305 × 1714 pixels and to 2653 × 3733 pixels for reference images and
192 for the sample images, respectively. Then, in order to minimize the effect of uncontrollable factors
193 such as drifts in the acquisition system or variations of the illumination conditions, each sample
194 image, S_i, was standardized using the corresponding reference image, R_i. To this aim, each
195 reference image, was compared with the reference of the first captured image, that was defined as
196 the master reference image, M_R. In particular, for each channel c (equal to R, G or B), the difference
197 between the mean value of all the pixels of R_i and the corresponding mean value of all the pixels of
198 M_R was computed as follows:
199 $$\Delta_i(c) = \overline{R_i}(c) - \overline{M_R}(c) \quad (1)$$
200 Then, this difference was used to calculate the corrected sample image, CS_i:
201 $$CS_i(c) = S_i(c) - \Delta_i(c) \quad (2).$$
202 After image standardization, the sample images were converted into the corresponding
203 colourgrams. Essentially, colourgrams are one-dimensional signals obtained by merging in
204 sequence the frequency distribution curves of a series of colour-related parameters extracted from
205 each RGB image, together with the loading vectors and the eigenvalues of PCA models calculated
206 on the RGB data. In this manner, datasets of RGB images are converted into matrices of signals,

each one acting like a fingerprint of the corresponding image and codifying its colour-related information content, while the spatial resolution is lost. The colourgrams matrix can be further analysed by means of suitable multivariate analysis techniques, allowing to evaluate all the acquired images together, i.e., to consider the colour-related information of the dataset as a whole.

For the conversion of images to colourgrams, the three-dimensional data array corresponding to each RGB image with size {2653 pixel rows × 3733 pixel columns × 3 R, G and B channels} was firstly unfolded into a two-dimensional matrix with size {9903649 rows (total number of pixels) × 3 columns (corresponding to the R, G and B channels)}.

Then, this matrix was expanded by adding a series of columns, corresponding to parameters calculated for each pixel starting from the R, G and B values:

- Lightness (L), defined as the sum of the three channel values;
- the relative colours (rR, rG and rB), defined as the ratio between each channel and L;
- Hue (H), Saturation (S) and Intensity (I), obtained by converting the RGB data into the HSI colour space;
- the nine score vectors obtained by calculating three PCA models on the raw, mean-centered and autoscaled RGB data (three principal components for each PCA model).

Then, for each one of the 19 columns of the resulting data matrix, the corresponding 256 points-long frequency distribution curve was calculated. The 19 frequency distribution curves were joined in sequence to form a unique vector, at the end of which the loading vectors and the eigenvalues of the three PCA models were also added.

The so obtained signal with length equal to 4900 points (= 256 × 19 + 36) is the colourgram, which retains the colour-related information of the corresponding image. For a more detailed description of the algorithm used to create the colourgrams, the reader is referred to Antonelli et al. (2004).

In this work, the 332 digital images were converted into the corresponding colourgrams, thus obtaining a colourgrams matrix with size {332 rows × 4900 columns}.

2.4.2 Exploratory data analysis and calibration of the colourgrams matrix

In order to obtain an overview of the dataset structure and to identify possible outlier images, a first evaluation of the colourgrams matrix was made by PCA. Both mean-centering and autoscaling were considered as column preprocessing methods to calculate the PCA models, and the number of PCs was selected according to the analysis of the corresponding scree plots.

Subsequently, Partial Least Squares (PLS) regression was applied to the colourgrams matrix in order to calculate calibration models able to predict the %DM values of the imaged samples. To this aim, the 332 colourgrams were split into:

- a training set composed of 180 signals, corresponding to 45 samples with %DM values equal to 0, 2, 4, 6, 8, 10, 30, 50, 70, 90 and 100;
- a test set composed of 152 signals corresponding to 38 samples, with %DM values equal to 1, 3, 5, 7, 9, 20, 40, 60, 80 and 95.

Also for the PLS models, both mean-centering and autoscaling were considered as column preprocessing methods. The performance of the PLS models was evaluated by means of the Root Mean Square Error (RMSE) and of the coefficient of determination (R^2) statistics, calculated on the calibration set (RMSEC, R^2_{Cal}), in cross-validation (RMSECV, R^2_{CV}) and in prediction of the test set (RMSEP, R^2_{Pred}). The optimal number of Latent Variables (LVs) was chosen by minimizing the value of RMSECV. In particular, a custom cross-validation method was used, subdividing the samples in 4 deletion groups (D1a+D2a; D1b+D2b; W1+W2a; W2b, *see* Figure 3).

Generally the information contained in the colourgram is partially redundant, since the whole signal is calculated without choosing a priori some relevant variables on the basis of the specific problem at hand. The evaluation of the image dataset by PCA and PLS considering the whole colourgram can be therefore helpful to perform a global assessment of the sources of colour variability of the analysed samples. However, in order to better focus on the quantification of the %DM values, and to increase predictive performance and robustness of the calibration models, it is necessary to retain only the useful (defect-related) colour features by means of proper variable selection algorithms.

To this purpose, a wide choice of methods is available in the literature, such as interval Partial Least Squares (Norgaard et al., 2000; Foca et al., 2016), genetic algorithms (Leardi, 2000) and sparse methods (Rasmussen and Bro, 2012; Calvini et al., 2015). Furthermore, feature selection can be applied in conjunction with transform methods able to compress the useful information pieces into a limited number of relevant variables, such as the wavelet transform (Antonelli et al., 2004; Foca et al., 2011; Pereira et al., 2011; Ulrici et al., 2012).

In particular, in the present work the simple but effective interval Partial Least Squares (iPLS) method (Norgaard et al., 2000; Ferrari et al., 2013) has been applied to the colourgrams matrix. Briefly, the iPLS algorithm starts by subdividing the whole signal in intervals of equal length, defined by the user. In the forward iPLS search strategy that has been used in this work, firstly local PLS models are calculated on each interval, to select the one leading to the minimum value of RMSECV. Then, local PLS models are calculated considering all the combinations of the selected interval together with each one of the other intervals, and the best two-intervals combination is selected again on the basis of the lowest RMSECV value. If the single-interval RMSECV value is lower than the two-intervals RMSECV value, only the first interval is selected. Otherwise, this iterative procedure is repeated by increasing each time the number of considered intervals, until no further decrease of the RMSECV value is achieved.

In this work forward iPLS was used considering six different interval size values (256, 128, 64, 32, 16 and 8 variables), and using both mean-centering and autoscaling as signal preprocessing methods. Finally, the best overall iPLS model was again selected on the basis of the lowest RMSECV value.

2.4.3 Image reconstruction using selected features

In addition to the parameters that are commonly used to evaluate the performance of the calibration models, the relevance of the best iPLS model results was also assessed by means of a specific algorithm that allows to represent the colourgram selected variables into the original image domain.

285 Firstly, the image reconstruction algorithm converts the indexes of the colourgram variables
286 selected by iPLS into the corresponding colour property values. For example, if one of the selected
287 regions is in the range from 300 to 319 colourgram units, this region corresponds to the green
288 channel values from 43 to 62. Then, the original image is segmented according to the selected
289 range: only the pixels with green values from 43 to 62 are kept, while the remaining ones are set
290 equal to 0 for all the R, G and B channels.
291 For each colourgram selected region, the resulting reconstructed image is displayed. In this manner,
292 although no spatial information about the original image is retained in the colourgram, it is however
293 possible to localize the image areas corresponding to the features of interest.
294 The algorithms used for image correction, conversion in colourgrams and image reconstruction of
295 the selected features were written in MATLAB language (ver. 7.12, The Mathworks Inc., USA),
296 while PCA, PLS and iPLS models were calculated using PLS_Toolbox (ver 7.5, Eigenvector
297 Research Inc., USA).
298
299 **3. Results and discussion**
300 *3.1 ELISA analysis*
301 As described in Section 2.3, the ELISA test was firstly performed on the groups of manually
302 selected defective and non-defective maize kernels. Concerning the defective maize, the results of
303 ELISA test showed that for both dry and wet types the concentration of DON was greater than the
304 maximum quantification value, equal to 10 ppm. As regards the non-defective maize, the average
305 concentrations of DON estimated by ELISA were equal to 2.00 ppm and to 1.66 ppm for dry and
306 wet types, respectively.
307 Afterwards, the ELISA test was also performed on the mixtures of defective and non-defective
308 maize kernels. In particular, the results of the analysis performed during the first acquisition step
309 showed that mixtures with %DM values greater than or equal to 10 presented concentration values
310 of DON exceeding 10 ppm.

311 This observation was confirmed by the ELISA test performed during the second acquisition step,
312 where concentrations of DON below the 10 ppm threshold were observed for the mixtures with
313 %DM values between 0 and 6, as reported in Figure 4. In this range, the concentration of DON was
314 directly proportional to the %DM value.
315
316 *3.2 Exploratory data analysis of the colourgram dataset*
317 The colourgram dataset was initially investigated by means of PCA considering both autoscaling
318 and mean-centering as signal preprocessing methods. The PCA model calculated on autoscaled
319 colourgrams was found to have an optimal dimensionality equal to 3 PCs, accounting for about
320 65% of the total variance. The score plot of the first two PCs is reported in Figure 5a, where the
321 samples are coloured according to the %DM values, and in Figure 5b, where the samples are
322 coloured according to the maize type (dry and wet). Figure 5a shows that the samples are
323 distributed along PC1 (40 % explained variance) according to the amount of defective kernels,
324 while Figure 5b highlights that the two maize types are separated along PC2. Also PC3 (not shown)
325 accounts for the separation between dry and wet maize. The separation between the two maize types
326 observed along PC2 and PC3 can be ascribed to a generally more reddish colour of the non-
327 defective wet maize with respect to the non-defective dry maize.
328 However, it must be underlined that the colourgrams variability due to difference between dry and
329 wet maize is orthogonal to that ascribable to %DM. For this reason, the moisture content of maize
330 should have a relatively limited influence on the development of calibration models for the
331 prediction of the %DM values.
332 The same colourgram dataset was also investigated by means of PCA using mean-centering as
333 signal preprocessing method. Two PCs were selected, accounting for about 70% of the total
334 variance. Also in this case, the samples are distributed along PC1 according to the %DM values
335 (Figure 5c), while the separation between dry and wet maize is not clearly visible (Figure 5d). The
336 arch-shaped distribution of the samples in these PC1-PC2 score plots is due to the fact that PC2

337 essentially accounts for the sample heterogeneity: homogeneous samples (i.e., samples that are
338 either almost completely non-defective or almost completely defective) are located at negative
339 values of PC2, while heterogeneous samples (i.e., mixtures with significant percentages of both
340 defective and non-defective maize kernels) have positive values of PC2.
341
342
343 *3.3 PLS calibration models*
344 The results of the PLS calibration models calculated on the whole colourgrams using both
345 autoscaling and mean-centering as signal preprocessing methods are reported in the first two rows
346 of Table 1. Both the PLS calibration models led to satisfactory results, with R^2 values always
347 greater than 0.975. The best PLS calibration model, chosen on the basis of the lowest RMSECV
348 value, was obtained using autoscaling and led to a RMSEP value equal to 3.1%.
349 Figure 6a shows the plot of the %DM values predicted with the best PLS model versus the
350 experimental %DM values. It is possible to observe that, at low %DM values (from 0 to 10), the
351 samples of dry maize are generally underestimated (samples mainly under the bisector), while the
352 samples of wet maize are generally overestimated (samples mainly over the bisector).
353
354 *3.4 iPLS calibration models*
355 On the whole, 12 different iPLS calibration models were calculated, considering six different
356 interval size values both for the mean-centered and for the autoscaled colourgrams. The last two
357 rows of Table 1 report the results of the iPLS models showing the lowest RMSECV values for each
358 preprocessing method. The two models led to almost identical results, both in terms of performance
359 and as for the number of LVs. However, the iPLS model calculated on the autoscaled colourgrams
360 can be considered as the best one since it is more parsimonious, including only 64 colourgram
361 variables.

362 Compared with the corresponding PLS model calculated on the whole colourgram, the best iPLS
363 model allowed to obtain only a slight reduction of the RMSEP value, from 3.1% to 3.0%. However,
364 compared to Figure 6a, Figure 6b shows that variable selection allowed to drastically reduce the
365 effect of the different maize types (dry/wet) on the estimate of the lowest %DM values (from 0 to
366 10).
367 Moreover, a detailed visual inspection of the images of the analysed samples revealed that the
368 actual composition of the mixtures was slightly different from the supposed one, i.e., that the
369 experimental %DM values were affected by a small experimental error. For example, some samples
370 with a supposed %DM value equal to 0 (non-defective samples) actually showed the presence of
371 some defective kernels (some of which are highlighted with black circles in Figure 7a). Similarly,
372 some samples with a supposed %DM value equal to 100% (defective samples) showed the presence
373 of some maize kernels without defects (black circles in Figure 7b), at least on the kernel side that
374 was imaged.
375 Therefore, the RMSEP values of the PLS and iPLS models were at least partly affected by the
376 presence of experimental error in the reference measurement values. Moreover, the presence of few
377 defective kernels within the non-defective maize could explain the relatively high average value of
378 DON found with the ELISA test in non-defective dry maize (2.00 ppm), which is slightly above the
379 limit of 1.75 ppm set for unprocessed maize used in foodstuff by the European Parliament
380 (Commission Regulation (EC) No 1126/2007).
381 Finally, it has to be underlined that the calibration models were calculated considering the four
382 images acquired for each mixture as separate objects, in order to evaluate the reproducibility of the
383 %DM estimated values. Therefore, the RMSE values reported in Table 1 include also the within-
384 sample variability and represent an overestimate of the values that would be obtained in a real
385 application of the models. Indeed, for an industrial application the estimate of the %DM values
386 would be obtained as the average of the values predicted from multiple images acquired in sequence
387 on different aliquots of the same sample. The application of the same approach to the test set

388 objects, considering the average of the four %DM values predicted for each sample, led to a
389 RMSEP value equal to 2.6% for the best iPLS model. Referring to the data reported in Figure 4, this
390 RMSEP value corresponds approximately to a concentration of DON ranging between 3 and 5 ppm.
391 Although not comparable to the error of the reference analytical methods, this result suggests
392 however the possibility to use RGB image analysis for a quick preliminary estimate of the degree of
393 maize contamination by DON, allowing for example the separate storage of the maize batches
394 depending on their %DM values, and/or to immediately reject those batches whose %DM value is
395 excessively high.
396
397 *3.5 Reconstruction of the selected features*
398 In order to obtain more direct information about variable selection made by the iPLS algorithm,
399 samples with different %DM values were randomly selected, and the corresponding images were
400 used for the reconstruction and the visualization of the features selected by the best iPLS model.
401 First of all, the plot of the regression vector of the best iPLS model was analysed in order to identify
402 the regions with regression coefficient values greater than zero, corresponding to variables that are
403 directly proportional to the %DM values. As shown in Figure 8, these regions belong to the
404 distribution curves of green, relative green, intensity and – with a smaller contribution – lightness.
405 As an example, Figure 9 reports the reconstruction of the selected features of green (Fig. 9b),
406 relative green (Fig. 9c) and intensity (Fig. 9d) of a portion of an image containing 50% of defective
407 maize kernels, together with the original RGB image (Fig. 9a). The reconstruction of the image
408 considering the selected features demonstrated that the colour-related parameters automatically
409 selected by the algorithm were actually related to the presence of defective maize kernels.
410 Furthermore, the portion of the white background pixels that were reconstructed was negligible,
411 therefore it did not interfere with the identification of defects in maize kernels.
412
413

4. Conclusions

In the present paper, an approach based on multivariate analysis of RGB images for the determination of the percentage of defective maize (%DM) has been presented, that could be used as a fast pre-screening of large maize batches for a preliminary estimate of the degree of maize contamination by DON. In fact, the analyses performed on the investigated samples with a commercially available ELISA test kit demonstrated that the %DM values and the concentration of DON are highly correlated with each other.

Through the automated selection of the colour features related to the presence of defective maize kernels, it was possible obtain a satisfactory prediction of the average %DM values of the test set samples (RMSEP = 2.6%). Interestingly, the best calibration model was scarcely affected by the marked colour differences between the two considered maize types (wet and dry). The robustness towards this source of variability can be reasonably ascribed to the fact that the colour features selected by the best calibration model were essentially related to the maize defective areas, as it was confirmed by the inspection of the reconstructed images.

These promising results demonstrated the possibility to develop a fast, cheap and non-destructive automated system for a preliminary screening of maize quality based on the presence of defective kernels. Indeed, based on the outcome of this research work, an industrial prototype is currently under development, which allows to automatically analyse 3 kg of maize in less than 1 min. Compared with the commercially available ELISA test kits, which usually require 20 minutes to analyse 20 g maize samples, this system should allow to speed up the transfer phase from harvesting to the warehouse and to further increase the quality and safety level of the final product.

In view of industrial implementation, based on the experience gained in the present work, further improvements will be made. Firstly, defective and non-defective maize samples used for model calibration will be submitted to multiple selection steps, thus minimizing the contribution of human error in the definition of the reference mixture samples. Moreover, a wider dataset of images including a larger number of batches and of maize varieties will be acquired, in order to better

estimate the effect of these sources of variability on the prediction error. Further improvements can be also reasonably gained by implementing a more refined image standardization procedure, in order to adjust the possible variations of the colour dynamic ranges. In addition to the first estimation of the degree of maize contamination by DON, the predicted %DM could also constitute an objective tool to quickly evaluate the maize batches, allowing, e.g., to define three quality categories that could be stored separately from each other: i) batches that could be potentially used in foodstuff, after proper evaluation by reference analytical methods; ii) batches that could be only used in animal feed materials, after proper evaluation by reference analytical methods; iii) batches that could not be used as food or feed.

Future developments could lead to automated systems for real-time evaluation of the maize quality of whole batches, also enabling the removal of defective kernels. Moreover, in an industrial application, the image reconstruction could also help to inspect outlier samples, to detect foreign particles or to highlight instrumental faults.

Acknowledgements

The authors wish to thank Freeray S.r.l. and Fornasier Tiziano & C. S.a.s. for providing technical and financial support.

458 **References**

459 Antonelli A., Cocchi M., Fava P., Foca G., Franchini G.C., Manzini D., Ulrici A. (2004). Automated
460 evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification
461 algorithm. *Analytica Chimica Acta*, 515, 3-13.

462 Calvini R., Foca G., Ulrici A. (2016). Data dimensionality reduction and data fusion for fast characterization
463 of green coffee samples using hyperspectral sensors. *Analytical and Bioanalytical Chemistry*, 408 (26),
464 7351-7366.

465 Calvini R., Ulrici A., Amigo J.M. (2015). Practical comparison of sparse methods for classification of
466 Arabica and Robusta coffee species using near infrared hyperspectral imagining. *Chemometrics and*
467 *Intelligent Laboratory Systems,* 146, 503–511.

468 Choudhary R., Paliwal J., Jayas D.S. (2008). Classification of cereal grains using wavelet, morphological,
469 colour, and textural features of non-touching kernel images. *Biosystems Engineering*, 99, 330-337.

470 Commission Directive 2003/100/EC of 31 October 2003 amending Annex I to Directive 2002/32/EC of the
471 European Parliament and of the Council on undesirable substances in animal feed. *Official Journal of the*
472 *European Union*, 285, 33-37.

473 Commission Regulation (EC) No 1126/2007 of 28 September 2007 amending Regulation (EC) No
474 1881/2006 setting maximum levels for certain contaminants in foodstuffs as regards *Fusarium* toxins in
475 maize and maize products. *Official Journal of the European Union*, 255, 14-17.

476 Del Fiore A., Reverberi M., Ricelli A., Pinzari F., Serranti S., Fabbri A.A., Bonifazi G., Fanelli C. (2010).
477 Earlt detection of toxigenic fungi on maize by hyperspectral imagining analysis. *International Journal of*
478 *Food Microbiology*, 144, 64-71.

479 Duchesne C., Liu J.J., MacGregor J.F. (2012). Multivariate image analysis in the process industries: A
480 review. *Chemometrics and Intelligent Laboratory Systems*, 117,116-128.

481 Edite Bezerra da Rocha M., Freire F.D.C.O., Erlan Feitosa Maia F., Izabel Florindo Guedes M., Rondina D.
482 (2014). Mycotoxins and their effects on human and animal health. *Food Control*, 36 (1), 159-165

Esbensen K., Geladi P. (1989). Strategy of Multivariate Image Analysis (MIA). *Chemometrics and Intelligent Laboratory Systems*, 7, 67-86.

FAO (2006). Maize: international market profile. Grains team food and agriculture organization of the United Nations economic and social department trade and markets division. Available at [http://www.ibrarian.net/navon/paper/Maize__International_Market_Profile.pdf?paperid=16236950], last accessed June 19, 2017.

Ferrari C., Foca G., Ulrici A. (2013). Handling large datasets of hyperspectral images: reducing data size without loss of useful information, *Analytica Chimica Acta*, 802, 29-39.

Foca G., Masino F., Antonelli A., Ulrici A. (2011). Prediction of compositional and sensory characteristics using RGB digital images and multivariate calibration techniques. *Analytica Chimica Acta*, 706, 238-245.

Foca G., Ferrari C., Ulrici A., Ielo M.C., Minelli G., Lo Fiego D.P. (2016). Iodine Value and Fatty Acids Determination on Pig Fat Samples by FT-NIR Spectroscopy: Benefits of Variable Selection in the Perspective of Industrial Applications, *Food Analytical Methods*, 9 (10), 2791-2806.

Geladi P., Grahn H. (1996). *Multivariate Image Analysis*. John Wiley & Sons, Chichester, UK.

Hossain M.Z., Goto T. (2014). Near-and mid-infrared spectroscopy as efficient tools for detection of fungal mycotoxin contamination in agricultural commodities. *World Mycotoxin Journal*, 7 (4), 507-515.

Kushiro M. (2008). Effects of milling and cooking processes on the Deoxynivalenol content in wheat. *International Journal of Molecular Sciences*, 9, 2127-2145.

Leardi R. (2000). Application of genetic algorithm–PLS for feature selection in spectral data sets. *J. Chemom.*, 14 (5-6), 643-655.

Lo Fiego D.P., Comellini M., Ielo M.C., Ulrici A., Volpelli L.A., Tassone F. (2007). Preliminary investigation of the use of digital image analysis for raw ham evaluation. *Italian Journal of Animal Science*, 6 (1), 693-695.

Maragos C.M. and Busman M. (2010). Rapid and advanced tools for mycotoxin analysis: a review. *Food Additives and Contaminants,* Vol. 27, No. 5, 668-700.

508 Nguyen V.H. (2013) Technical Specifications for maize, V13.1, World Food Programme. Available at
509 [http://documents.wfp.org/stellent/groups/public/documents/manual_guide_proced/wfp261422.pdf], last
510 accessed August 31, 2017.

511 Norgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval partial
512 least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared
513 spectroscopy. *Applied Spectroscopy*, 54 (3), 413-419.

514 Pereira A.C., Reis M.S., Saraiva P.M. (2009). Quality control of food products using image analysis and
515 multivariate statistical tools. *Industrial and Engineering Chemistry Research*, 48 (2), 988-998.

516 Pereira A.C., Reis M.S., Saraiva P.M., Marques J.C. (2011). Development of a fast and reliable method for
517 long- and short-term wine age prediction. *Talanta*, 86, 293-304.

518 Pierini G.D., Fernandes D.D.S., Diniz P.H.G.D., de Araújo M.C.U., Di Nezio M.S., Centurión M.E. (2016).
519 A digital image-based traceability tool of the geographical origins of Argentine propolis. *Microchemical*
520 *Journal*, 128, 62-67.

521 Prats-Montalbán J.M., de Juan A., Ferrer A. (2011). Multivariate image analysis: A review with applications.
522 *Chemometrics and Intelligent Laboratory Systems*, 107, 1-23.

523 Rasmussen M.A., Bro R. (2012). A tutorial on the Lasso approach to sparse modeling. *Chemometrics and*
524 *Intelligent Laboratory Systems*, 119, 21–31.

525 Reis, M. S. (2014). Multivariate image analysis. In D. Granato (Ed.), *Mathematical and Statistical Methods*
526 *in Food Science and Technology*, John Wiley & Sons, Ltd, Chichester, UK.

527 Ruan R., Ning S., Song A., Ning A., Jones R., Chen P. (1998). Estimation of Fusarium scab in wheat using
528 machine vision and a neural network. *Cereal Chemistry*, 75 (4), 455-459.

529 Singh C.B., Jayas D.S., Paliwal J., White N.D.G. (2012). Fungal damage detection in wheat using short-
530 wave near-infrared hyperspectral and digital colour imaging. *International Journal of Food Properties*, 15
531 (1), 11-24.

532 Sobrova P., Adam V., Vasatkova A., Beklova M., Zeman L., Kizek R. (2010). Deoxynivalenol and its
533 toxicity. *Interdisciplinary Toxicology*, 3, 94-99.

534 Steyn P.S. (1995) Mycotoxins, general view, chemistry and structure. *Toxicology Letters*, 82, 843-851

535 Suleiman R., Rosentrater K.A., Bern C. (2013). Effects of deterioration parameters on storage of maize: a
536 review. *Journal of Natural Sciences Research*, Vol. 3, No. 9, 147-165.

537 Sweeney M.J. and Dobson A.D.W. (1998). Mycotoxin production by *Aspergillus*, *Fusarium* and *Penicillium*
538 species. *International Journal of Food Microbiology*, 43, 141-158.

539 Udomkun P., Wiredu A.N., Nagle M., Müller J., Vanlauwe B., Bandyopadhyay R. (2017). Innovative
540 technologies to manage aflatoxins in foods and feeds and the profitability of application – A review. *Food*
541 *Control*, 76, 127-138.

542 Ulrici A., Foca G., Ielo M.C., Volpelli L.A., Lo Fiego D.P. (2012). Automated identification and
543 visualization of food defects using RGB imaging: Application to the detection of red skin defect of raw
544 hams. *Innovative Food Science and Emerging Technologies*, 16, 417–426.

545 Ulrici A., Serranti S., Ferrari C., Cesare D., Foca G., Bonifazi G. (2013). Efficient chemometric strategies for
546 PET-PLA discrimination in re cycling plants using hyperspectral imagining. *Chemometrics and Intelligent*
547 *Laboratory Systems*, 122, 31-39.

548 Valiente-González J.M., Andreu-García G., Potter P., Rodas-Jordá Á. (2014). Automatic corn (*Zea mays*)
549 kernel inspection system using novelty detection based on principal component analysis. *Biosystems*
550 *Engineering*, 117, 94-103.

551 Williams P.J., Geladi P., Britz T.J., Manley M. (2012). Investigation of fungal development in maize kernels
552 using NIR hyperspectral imaging and multivariate data analysis. *Journal of Cereal Science*, 55, 272-278.

553 Yu H., MacGregor J.F., Haarsma G., Bourg W. (2003), Digital imaging for on-line monitoring and control of
554 industrial snack food processes. *Industrial and Engineering Chemistry Research*, 42, 3036-3044.

555

556 **Captures to tables and figures**

557 **Table 1** – Results of the PLS models and of the two best iPLS models.

558 **Figure 1 –** Experimental setup used for acquisition of the sample images.

559 **Figure 2 –** Key steps followed for the elaboration of the RGB images.

560 **Figure 3 –** Diagram representing the different mixture samples considered in the two acquisition
561 steps.

562 **Figure 4** – Concentration of DON measured by ELISA test vs. percentage of defective maize
563 kernels (%DM).

564 **Figure 5 –**PC1 vs PC2 score plots of the PCA models calculated on autoscaled (a and b) and mean
565 centered (c and d) colourgrams. In (a) and (c) the samples are coloured according to the
566 concentration of defective kernels, in (b) and (d) the samples are coloured according to maize type.

567 **Figure 6** – Results of the best PLS (a) and iPLS (b) calibration models calculated on the autoscaled
568 colourgram variables: test set predicted %DM vs. experimental %DM values.

569 **Figure 7** – RGB images of samples with experimental %DM values equal to 0 (a) and 100 (b).

570 **Figure 8** – Regression coefficients of the best iPLS model.

571 **Figure 9 –** Original RGB image (a) of a sample with a %DM value equal to 50 and reconstructed
572 images considering the selected features of green (b), relative green (c) and intensity (d) values.

573

574

*Highlights (for review)

- RGB image analysis was used to quantify the percentage of defective maize (%DM)
- A positive correlation was observed between %DM and concentration of Deoxynivalenol
- Conversion of images into signals (colourgrams) allowed to use large image datasets
- Feature selection applied to colourgrams led to satisfactory prediction of %DM
- Image reconstruction of the selected features allowed easy defects visualization

Table1

Table 1 – Results of the PLS models and of the two best iPLS models.

Calibration method	Pretreatment	iPLS interval size	Included variables	LVs	RMSEC	RMSECV	RMSEP	R^2_{Cal}	R^2_{CV}	R^2_{Pred}
PLS	Autoscaling	-	4900	4	2.0	3.3	3.1	0.997	0.990	0.991
PLS	Mean-centering	-	4900	6	2.2	5.3	3.4	0.996	0.975	0.988
iPLS	Autoscaling	8	64	6	1.8	2.0	3.0	0.997	0.997	0.992
iPLS	Mean-centering	16	640	6	1.7	2.0	3.1	0.997	0.996	0.991

Figure1
Click here to download high resolution image

Figure2
Click here to download high resolution image

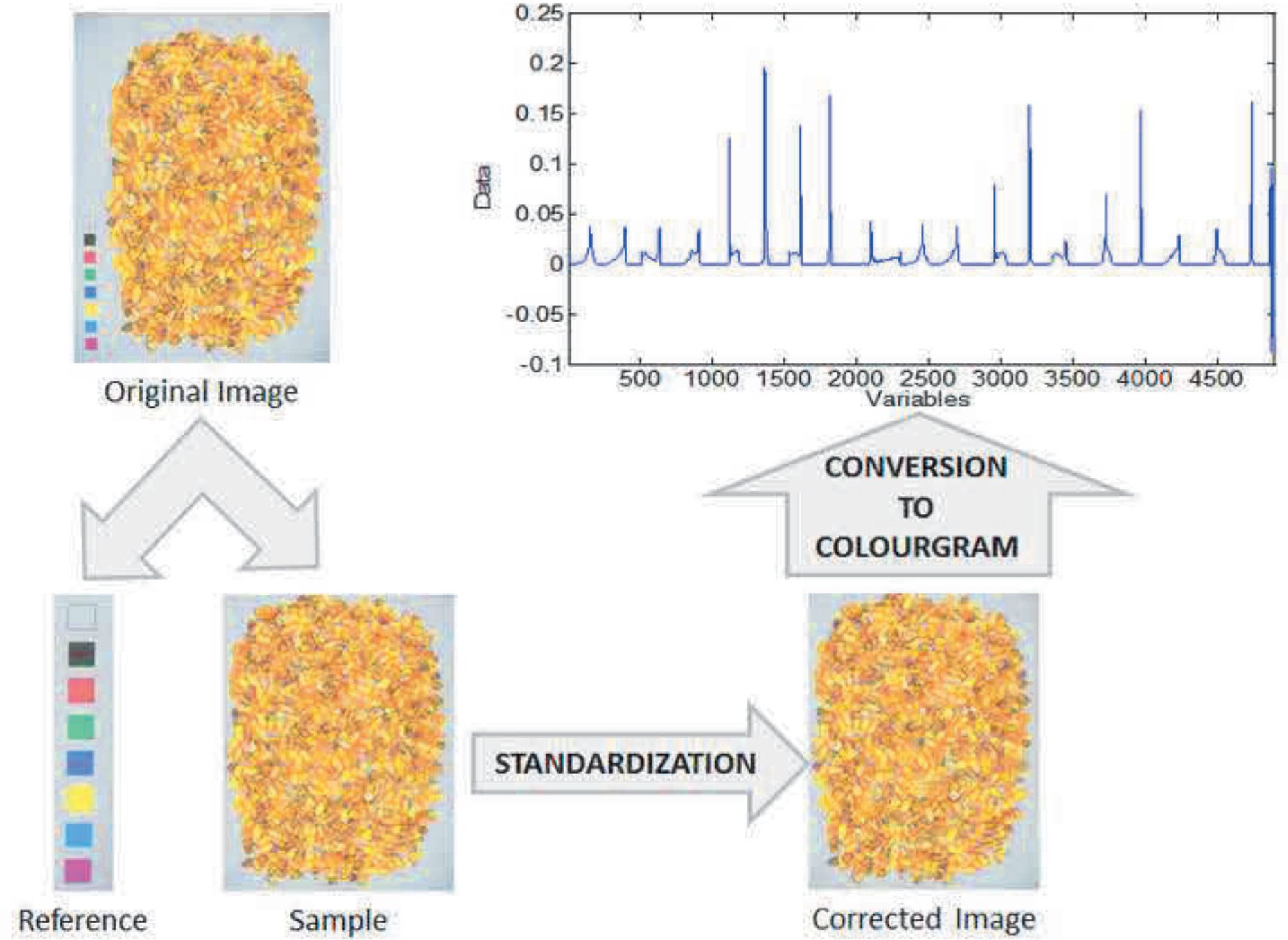

Figure3
Click here to download high resolution image

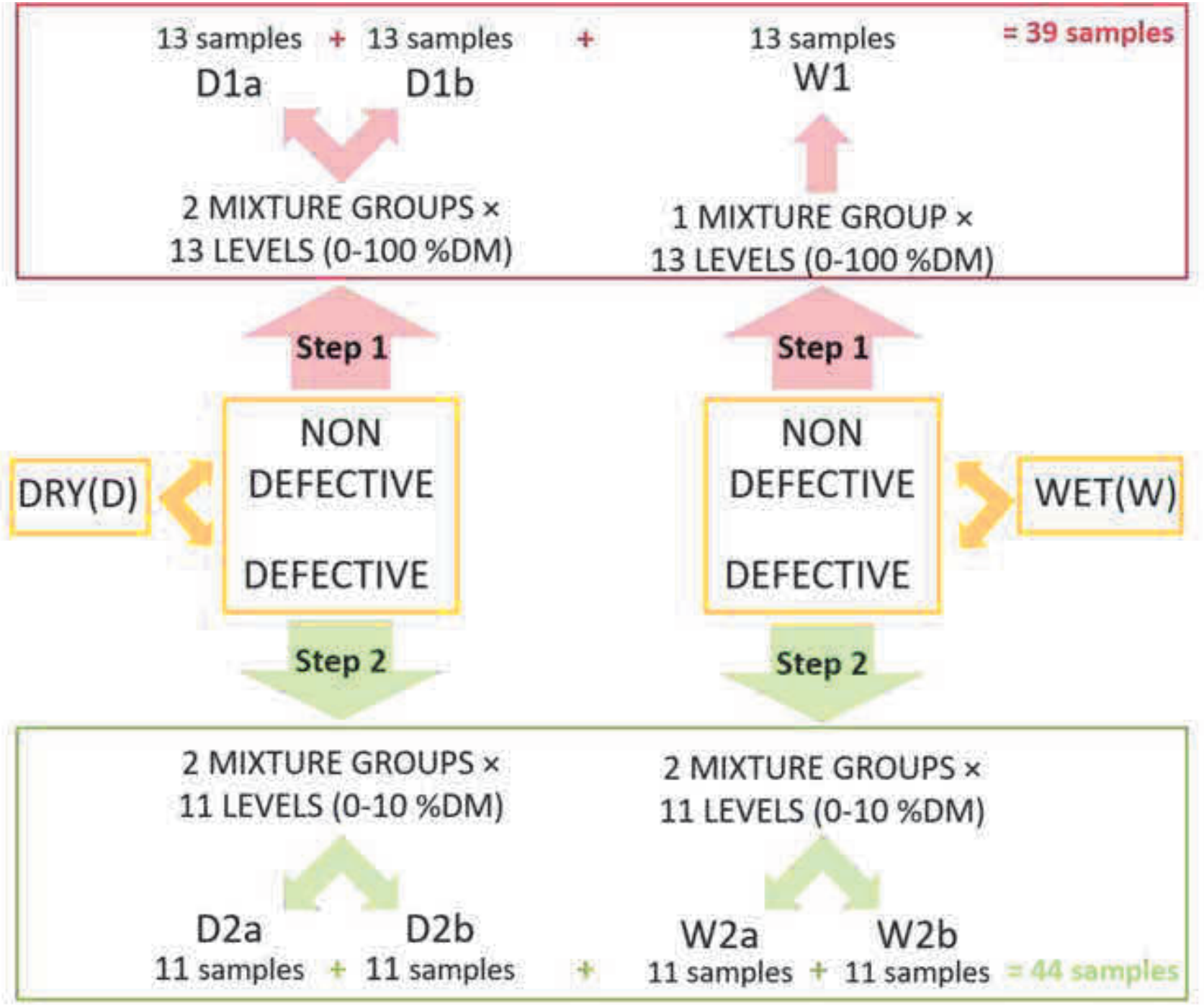

Figure4
Click here to download high resolution image

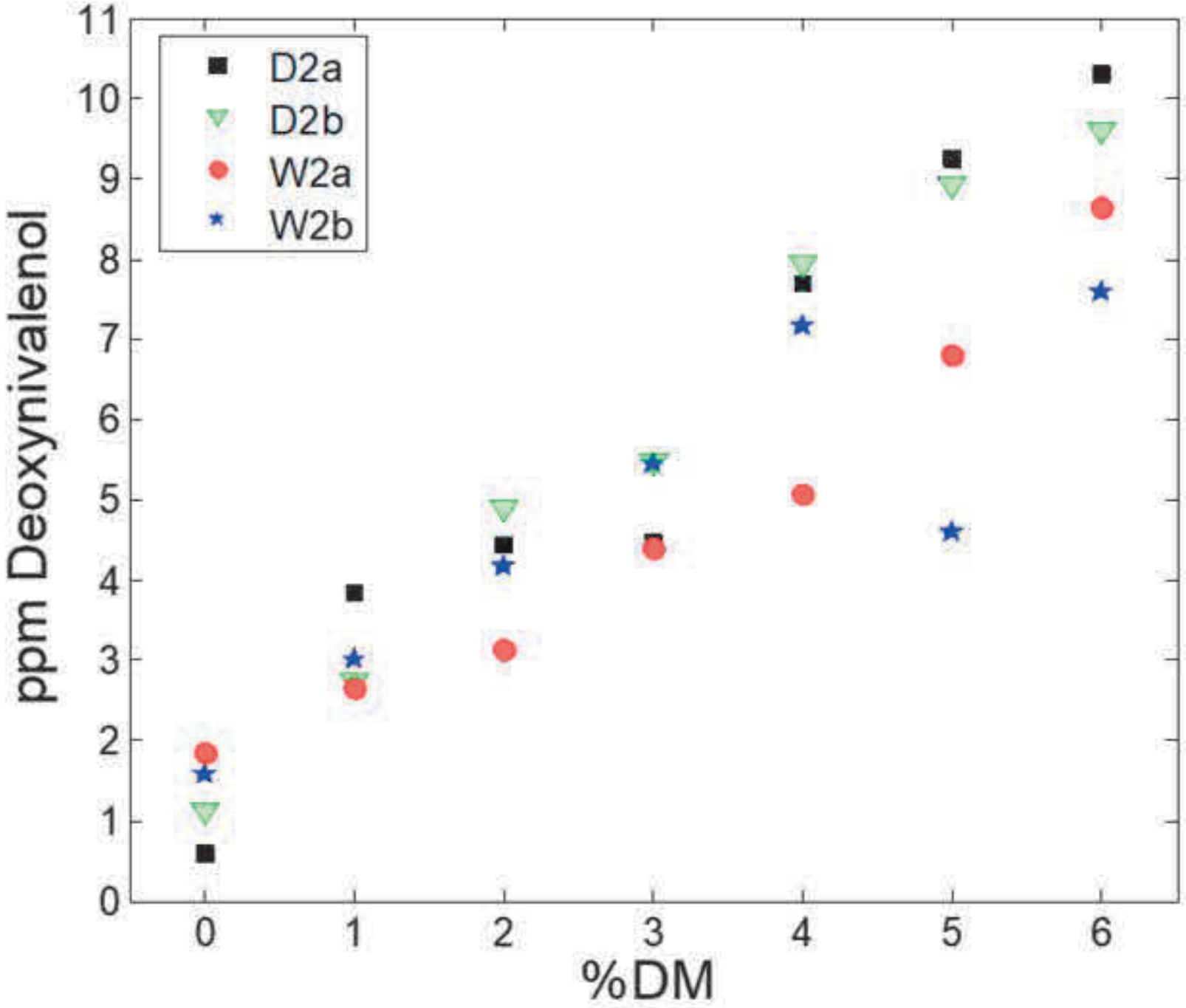

Figure5
Click here to download high resolution image

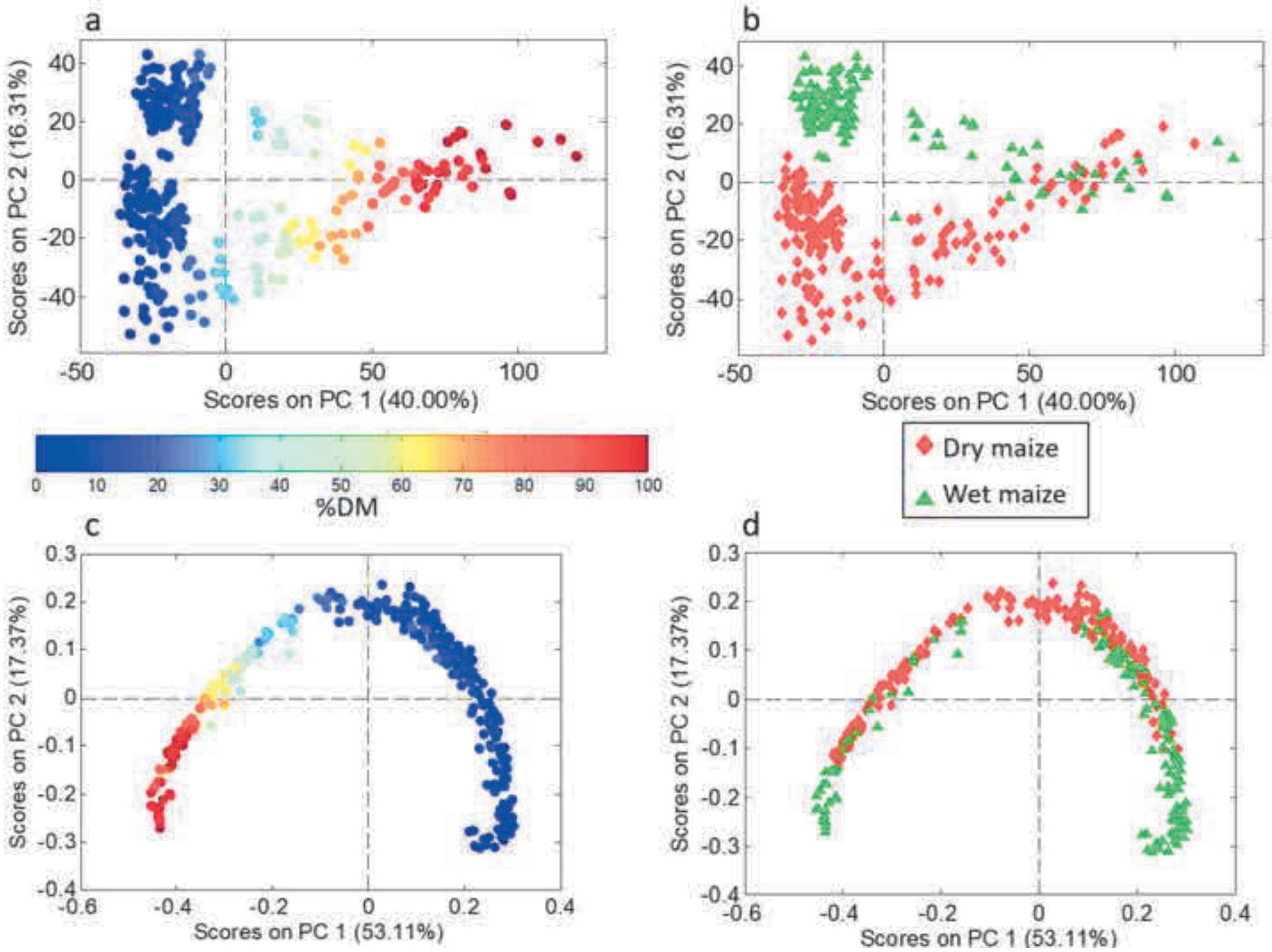

Figure6
Click here to download high resolution image

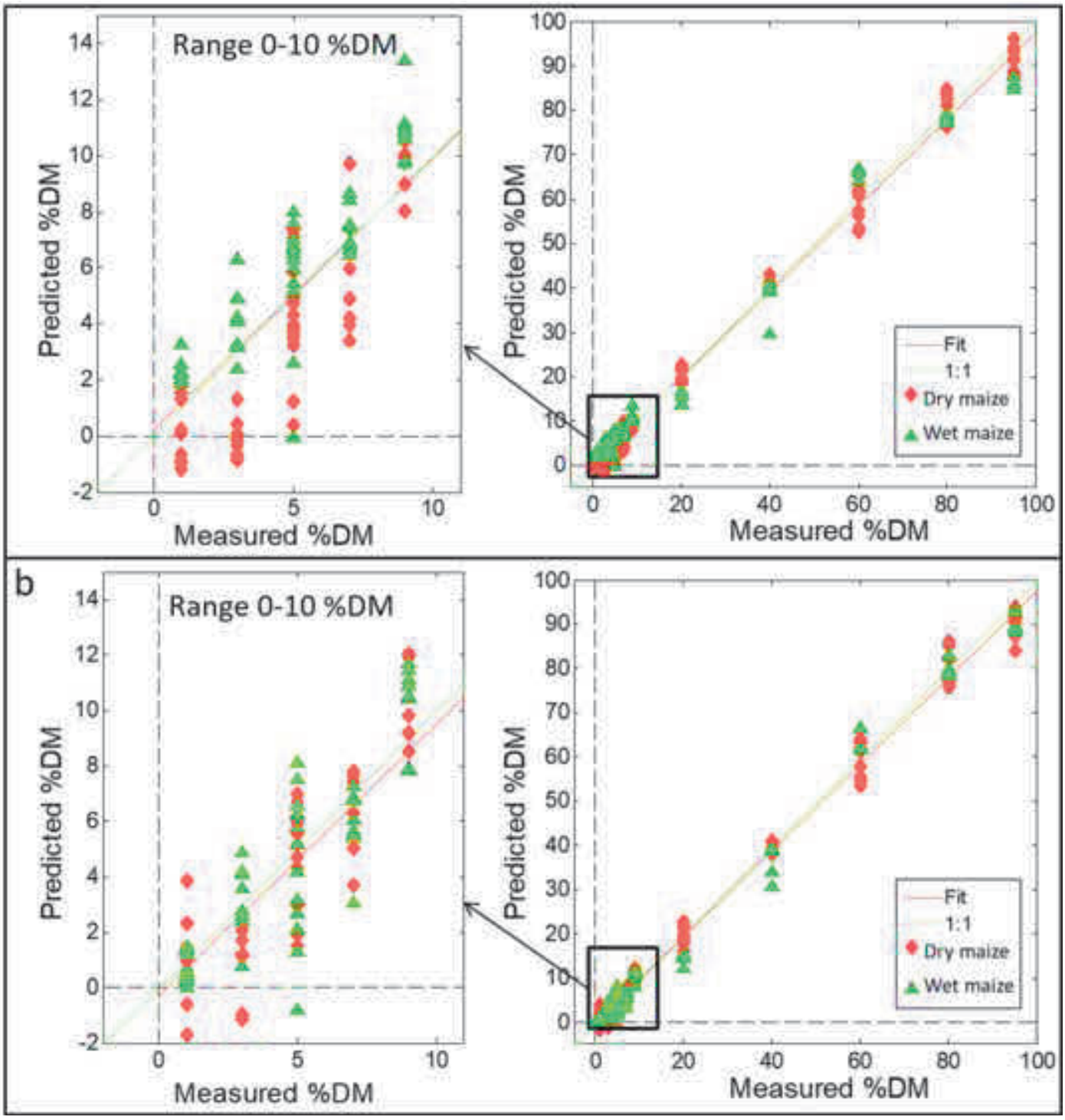

Figure7
Click here to download high resolution image

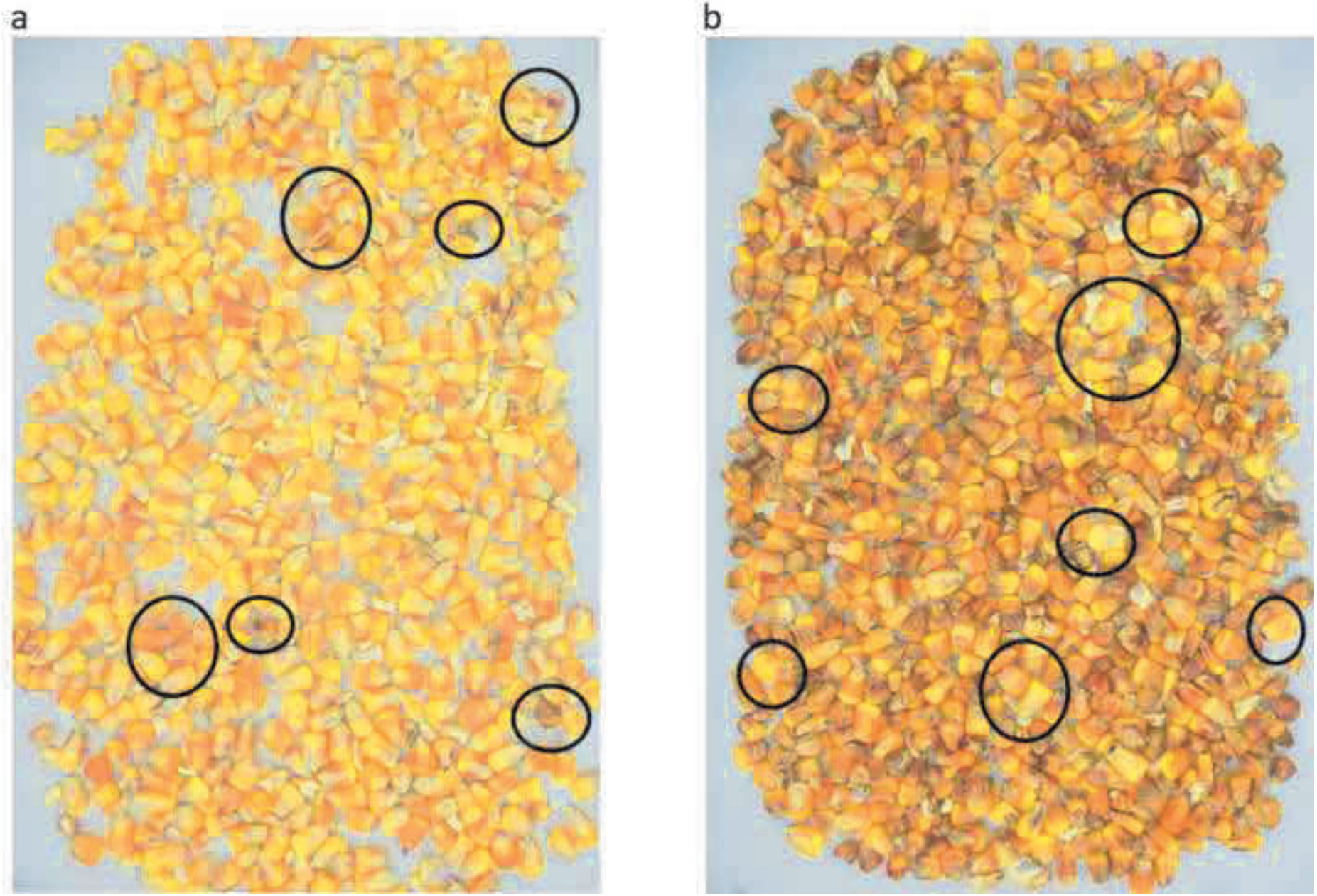

Figure8
Click here to download high resolution image

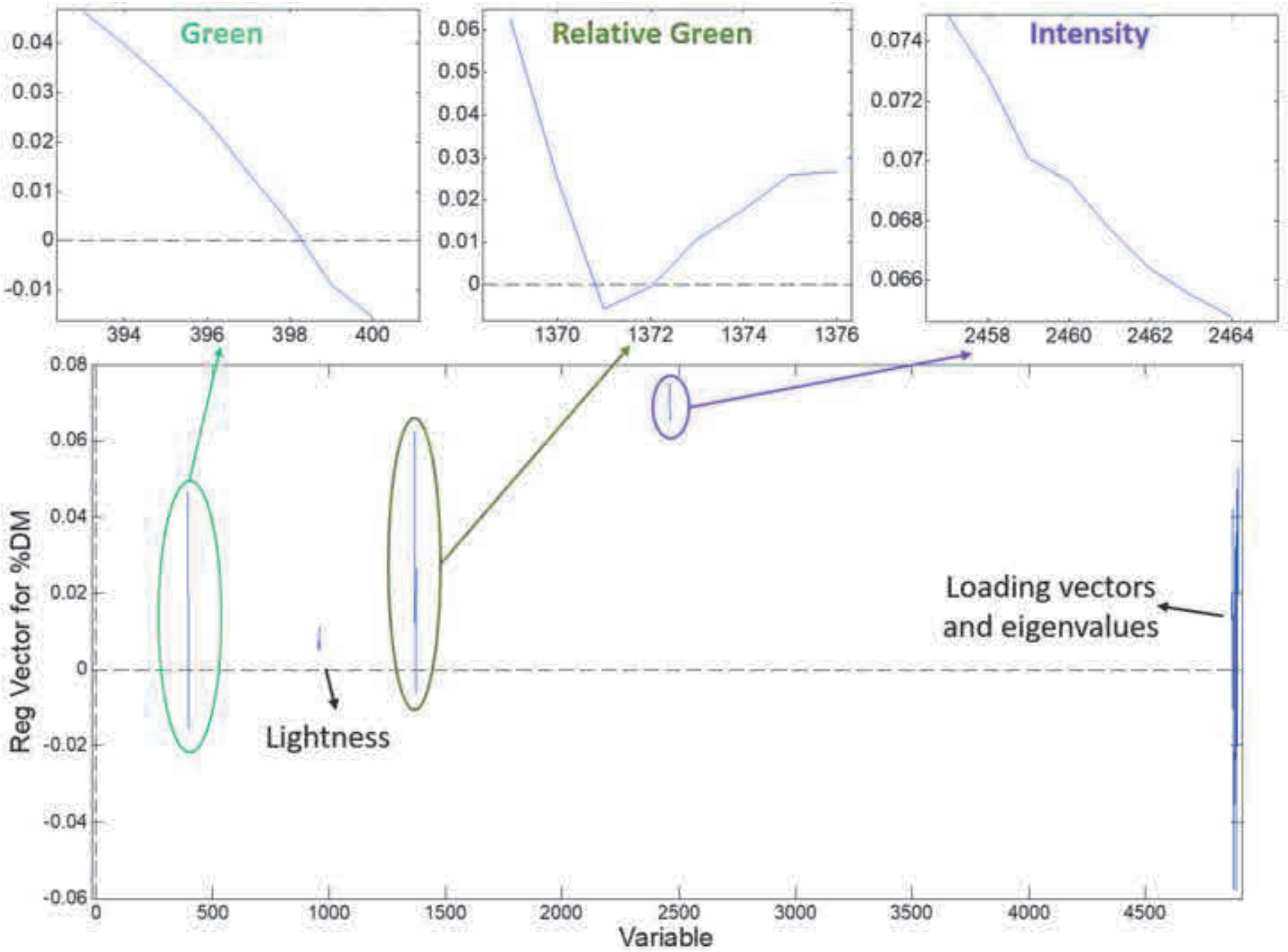

Figure9
Click here to download high resolution image