ELSEVIER

Chemometrics and Intelligent Laboratory Systems 24 (1994) 145–167

Chemometrics and intelligent laboratory systems

Multiwavelength microscopic image analysis of a piece of painted chinaware: classification and regression ☆

Paul Geladi [a,*], Jurgen Swerts [b], Fredrik Lindgren [a]

[a] *Research Group for Chemometrics, Department of Chemistry, University of Umeå, S-901 87 Umeå, Sweden*
[b] *Micro and Trace Analysis Centre, Department of Chemistry, University of Antwerp, B-2610 Wilrijk, Belgium*

Received 2 August 1993; accepted 4 April 1994

Abstract

A multivariate microscopic study of a piece of painted china porcelain is undertaken by image analysis. The object has archeological and artistic value and therefore a detailed study may be worthwhile. The use of multivariate image analysis intends the exploration of the artefact (artistic object) in more spectral and spatial detail than by just visual inspection. The study is used as a means of introducing and further exploring the different aspects of multivariate image analysis. The goal of the paper is twofold: (1) showing how the multivariate image is constructed and analyzed and (2) using some of the obtained results to introduce and further expand some of the techniques of multivariate image analysis. The example image has a size of $6 \times 512 \times 512$. Classifications by feature space segmentation and by regression are shown to be useful and objective methods of acquiring insight in measurement errors and in the artistic detail of the painting. Some new concepts for multivariate analysis are introduced. One of them is related to the comparison of regression models.

1. Introduction

In chemometrics, and in multivariate data analysis in general, there has traditionally been an emphasis on visualizing results in plots and graphs. Score plots, loading plots, biplots, scree plots give the scientist something that mere numbers are unable to express. Visualization is an important aid in interpreting the data, even with data sets of small or intermediate size e.g. from five to a few hundred objects or variables. When large data sets are concerned, numerical interpretation becomes totally impossible. Visualization is then the only remaining tool. Some special techniques are needed for visualizing large amounts of data points.

In multivariate image analysis, huge amounts of data are handled. An image of reasonable resolution would need about 500×500 pixels. This is 250 000 objects. This image is often measured for different variables. Extreme cases exist in remote sensing, where airborne detectors can create images of e.g. 2000×2000 pixels in over 100 wavelength bands. More modest image sizes are created in microscopy, electron and ion microscopy and in the medical imaging techniques

☆

* Corresponding author.

SSDI 0169-7439(94)00035-H

for patient diagnosis such as magnetic resonance imaging and X-ray tomography. Here, images of size 256 × 256 or 512 × 512 having 5–15 variables are standard, but the evolution is always towards better spatial and spectral resolution.

Multivariate image analysis has been presented earlier [1–8] and only the most important basic aspects are repeated here. The object of study is a piece of painted and glazed chinaware, studied with low enlargement (pixel size 14 μm × 14 μm) in a stereoscope with CCD video camera digitalization. The painting is quite intricate and shows a lot of detail, combined with wear and erosion effects. The resulting multivariate image when using six visual wavelength bands has the size $6 \times 512 \times 512$. It is used for studying the spectral and spatial content of the intricate pattern of the painting but mainly as a means of introducing some new aspects of multivariate image analysis.

A multivariate image of size $I \times J$ pixels and $k = 1,\ldots,K$ variables is represented by a 3-way array $\underline{\mathbf{X}}$. After pretreatments such as mean-centering, linear or nonlinear rescaling, principal component analysis can be carried out as follows:

$$\underline{\mathbf{X}} = \sum_{a=1}^{A} \mathbf{T}_a * \mathbf{p}_a + \underline{\mathbf{E}} \tag{1a}$$

$$x_{ijk} = \sum_{a=1}^{A} t_{ija} * p_{ak} + e_{ijk} \tag{1b}$$

where $\mathbf{T}_a$ is a score matrix (size $I \times J$), $\boldsymbol{p}_a$ is a loading vector (size K), $\underline{\mathbf{E}}$ is the residual (size $K \times I \times J$), $*$ is a 3-way operation explained in Ref. [8], $a = 1,\ldots,A$ is the number of components.

Eq. 1b shows the same operation in arithmetical mode with x_{ijk} an element of $\underline{\mathbf{X}}$ etc. More about this can be found in Refs. [1,3,6–8].

The score matrices $\mathbf{T}_a$ have the same size as images and can be shown as images on a screen. The score vectors $\boldsymbol{p}_a$ can be used for making loading plots indicating the contribution of each variable to the PCA model. The score images are used separately or in false colour composites. For viewing the pixels in feature space, score plots are made and clusters, outliers and gradients are detected and analyzed in these. The final result of this is a segmentation in feature space with the corresponding segmentation in image space. This segmentation allows the detection of true classes, illumination and reflection errors and surface problems. Some new segmentation ideas are introduced and variable reduction is investigated.

The example is also used for explaining regression on multivariate images, in this case discriminant regression by biased regression methods. Principal component regression (PCR), partial least squares regression (PLS) and ridge regression (RR) are used. Some ideas about quality of regression models and about comparing regression models are illustrated.

Image analysis can only be explained by visualization of the results in images and graphs. Text and numbers alone cannot convey the most important details. Some numerical results are given in tables for completeness. These numbers would normally be less important in an analysis.

2. Experimental

2.1. The object

The object of study is a piece of chinaware (porcelain). It is about 100 years old. An account on the nature and age determination of these materials is given in Ref. [9]. The piece is painted and glazed and has undergone some wear and corrosion. Colour, glaze and painting are often used with these objects to determine quality and authenticity. The experts feel that more objective methods than just visual inspection are needed. The only pretreatment of the sample under study consisted of cleaning with acetone to remove residues of glue and fingerprints. A picture of the chinaware piece is shown in Fig. 1. The piece is from a saucer and is therefore not flat. The surface is slightly curved, and this may give reflection problems. Attempts to eliminate this reflection problem resulted in too weak illumination (noisy images), but in the future better solutions may be found.

The piece under study is assumed to belong the 'Famille Verte' class. The following chemical

substances were used for applying the colors: red = Fe_2O_3; green = copper oxides; turquoise = copper oxides with sometimes Al compounds; black = Mn, Fe and Cu oxides mixed; gold = Au. The colors were applied as enamel and baked in at 800°C. The porcelain itself was formed at 1200–1300°C. The colors were sometimes applied in different layers to increase depth illusion and intensity of the colors. The piece had been lying in a sewer for a long time when it was found. This accounts for some wear and fading of colors.

Multivariate microscopy

A small area of the piece of chinaware was used for creating a multivariate image. A Wild Apozoom stereoscope with a Dage-MTI CCD72 mounted camera was used for collecting the images in reflection mode. The digitization of the video signal was done with Kontron IBAS 2.0 hardware and software. Images of size 512 × 512 (14 μm × 14 μm pixel size) were stored on disk for further analysis. All images were collected by averaging 16 times to reduce random noise. Illumination was with a Zeiss Superlux 300 Xenon lamp with optical fibers organized in a ring around the stereoscope objective. The light was filtered with interference filters between two fiber bundles in order to get selected wavelength bands. This setup creates a volume of homogeneous monochromatic illumination around the sample. The wavelength bands used are shown in Table 1. Fig. 2 gives a schematic overview of the experimental setup. The resulting image is of size 6 × 512 × 512. All image visualization and collection in this paper uses the intensities as integers in the range 0–255. All calculations on the image data are done in double precision. The image data are available via anonymous FTP. Contact geladi@biovax.umdc.umu.se for more details.

Calculations and visualization

Calculations were carried out on a Sun Unix server, alternatively on a 486 personal computer. Programs used for univariate analysis were Erdas 7.5 and 8.0 for Unix [10] and Erdas 7.4 for DOS [11]. Special in-house programs for multivariate image analysis and multivariate image regression written in C and in Fortran for the Erdas Toolkit were used. Some regression calculations were also carried out in Splus for Unix [12]. Colour visualization was on a multisync video screen connected to a Revolution Number Nine PC hardware. Photography was from a Microvitec multisync video screen on 200 ISO slide film. Most graphics were done on an Apple Macintosh computer using MacDraw and GraphMu [13]. The regression programs were published earlier, except for the ridge regression found in the Appendix.

3. Univariate statistical analysis

The raw images are shown in Fig. 3a. A univariate analysis gives an idea of the properties of the wavelength bands in the multivariate image. Global statistics and histograms were calculated. Table 1 gives some statistical properties of the images. A false colour composite of the wavelength bands at 450 nm (blue), 540 nm (green) and 630 nm (red) is used as an RGB colour image. It can be seen in Fig. 3b. The histograms are shown in Fig. 4. They are clearly multimodal, making global statistics rather superfluous. Univariate statistics on large heterogeneous data sets such as images and especially multivariate images is in most cases quite useless. It may however serve a purpose in checking the integrity of the data after network transfer or transformations. Univariate statistics are excellent tests for errors when image files are moved and transformed. They also serve in the detection of physical errors such as extreme noise and illumination problems, but normally these errors would be detected and removed under the data collection stage and not during data analysis.

4. Multivariate image analysis

Multivariate image analysis was introduced earlier and the references [1–8] give ample explanation of the most important details. Fig. 5 gives a schematic overview of the methods used in going from a multivariate image to feature space segmentation. A first step in a multivariate image

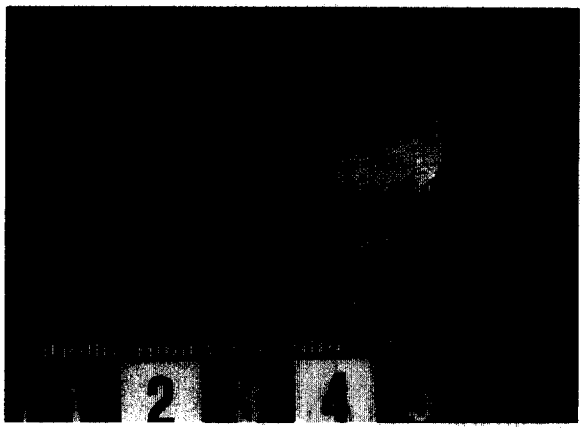


Fig. 1. A picture of the piece of chinaware under study. A scale is included for reference. The scale is in centimeters, with subdivision in millimeters.

analysis is to find out what the univariate analysis explains about the images. As shown in Table 1, the image at 680 nm is darker than the other ones. This has to do with the limitations of the light source and with limited CCD camera sensitivity. Reweighting of a dark image may be a way

Table 1
Wavelength bands and statistical analysis

Band No. or Variable No.	λ (nm)	Mean	Standard deviation	Comment
1	460	79.8	44.8	Blue
2	500	80.0	39.4	Blue-green
3	540	92.9	42.0	Green
4	580	113.9	51.2	Yellow
5	630	89.3	42.2	Red
6	680	30.0	14.2	Dark red

of alleviating the problem, but with the risk that noise is amplified too. An important aspect in all multivariate analysis is correlation between variables. A correlation table of the six wavelength bands is given in Table 2. The table shows that the wavelengths are highly correlated with their closest neighbours and less (but still very well) with the more remote ones, as can be expected of spectral data. The correlations in Table 2 are global. By taking subsets such as specific geometrically defined regions, other values may be obtained.

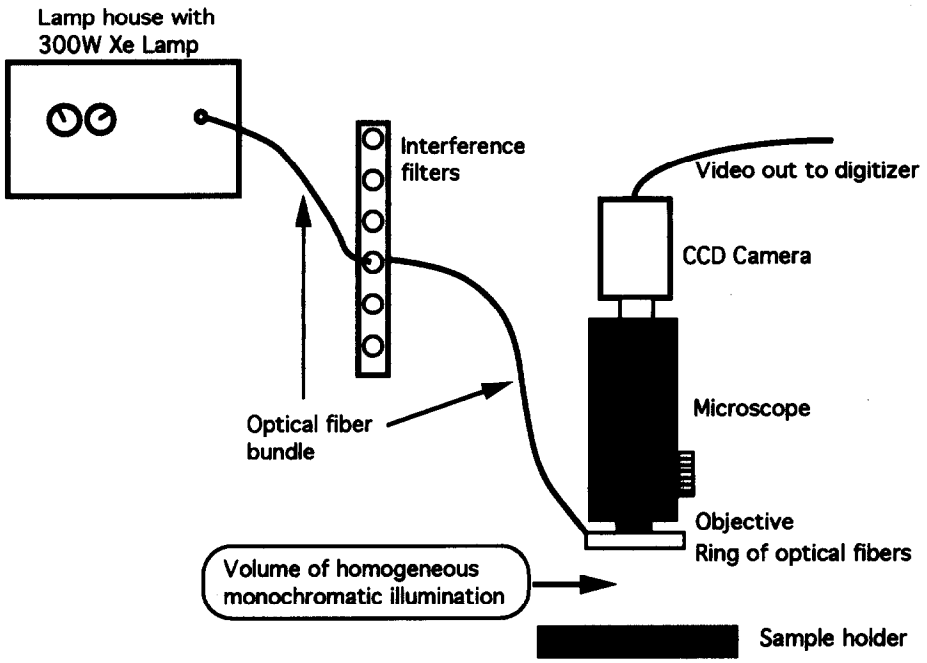


Fig. 2. A schematic overview of the experimental setup for monochromatic illumination and video registration of images.

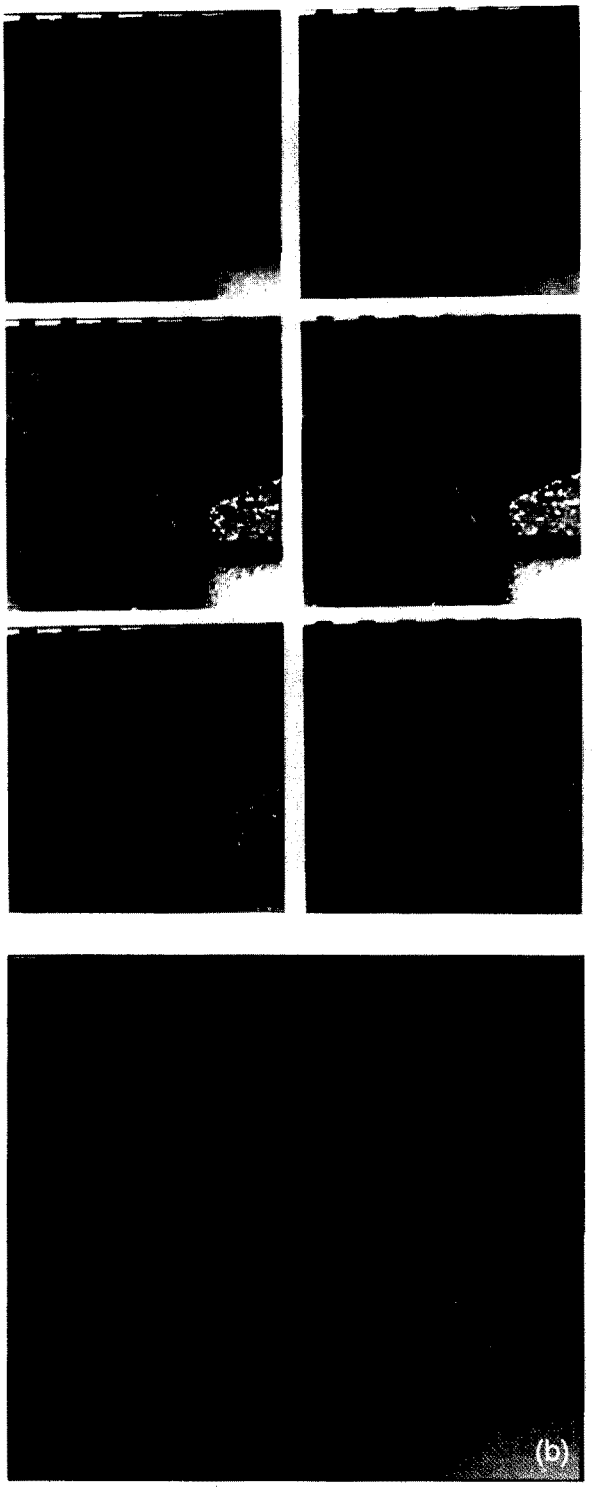


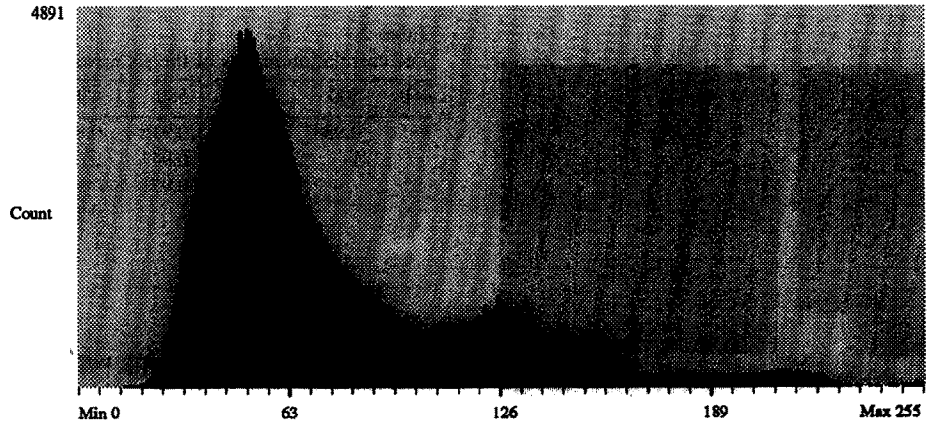
Table 2
The correlation matrix of the six wavelength bands

460	500	540	580	630	680	nm
1	0.988	0.960	0.915	0.861	0.826	460
	1	0.988	0.943	0.885	0.851	500
		1	0.971	0.922	0.892	540
			1	0.981	0.956	580
				1	0.990	630
					1	680

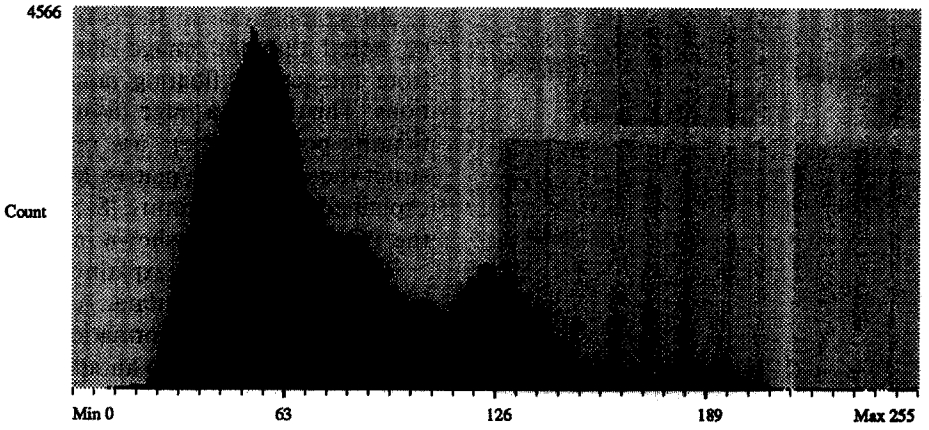
The best way of dealing with highly correlated variables is to reduce the dimensionality by principal component analysis (PCA). PCA was carried out on the multivariate image after pretreatment. This pretreatment consisted of mean-centering and rescaling each variable to variance 1, also comparable to the z-transform. It should be noted that the image data are transformed from integers to floating point by these operations. This is not a huge disadvantage, since the floating point numbers are only needed to construct the correlation matrix. More about this was explained in the literature [8]. The first results of the PCA analysis are shown in Table 3.

Three components explain 99.7% of the total sum of squares of the data. A fourth component may contain some information. Components 5 and 6 are very small. As an additional visual clue to the table, a modified scree plot is given in Fig. 6. This plot is different from the usual scree plot because $\lambda^{1/2}$ is used instead of λ. The score images 1–4 are shown in Fig. 7. Score images 5 and 6 contain little visual information and are not shown here. The scores are given as integers rescaled to fill the range 0–255. For practical reasons, this is often the best way of presenting image scores, but the true size double precision

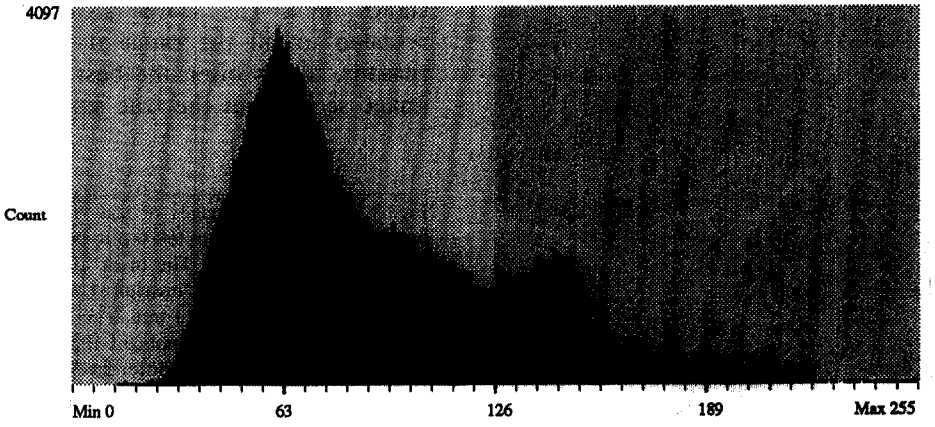
Fig. 3. (a) The six images of size 512×512 in wavelength bands of 460, 540, 630 nm left top to bottom and 500, 580, 680 nm right top to bottom. The black and white pattern in the top of the images is the measuring scale. The distance between two black stripes is 1 mm. The total image is about 7.5 mm×7.5 mm. It can be noticed that the images are very correlated, but that small differences exist. (b) The colour image is a composite of the bands around 460 nm (blue), 540 nm (green) and 630 nm (red). It is a false colour composite. A real colour TV camera would give 'true' colours.



460 nm



500 nm



540 nm

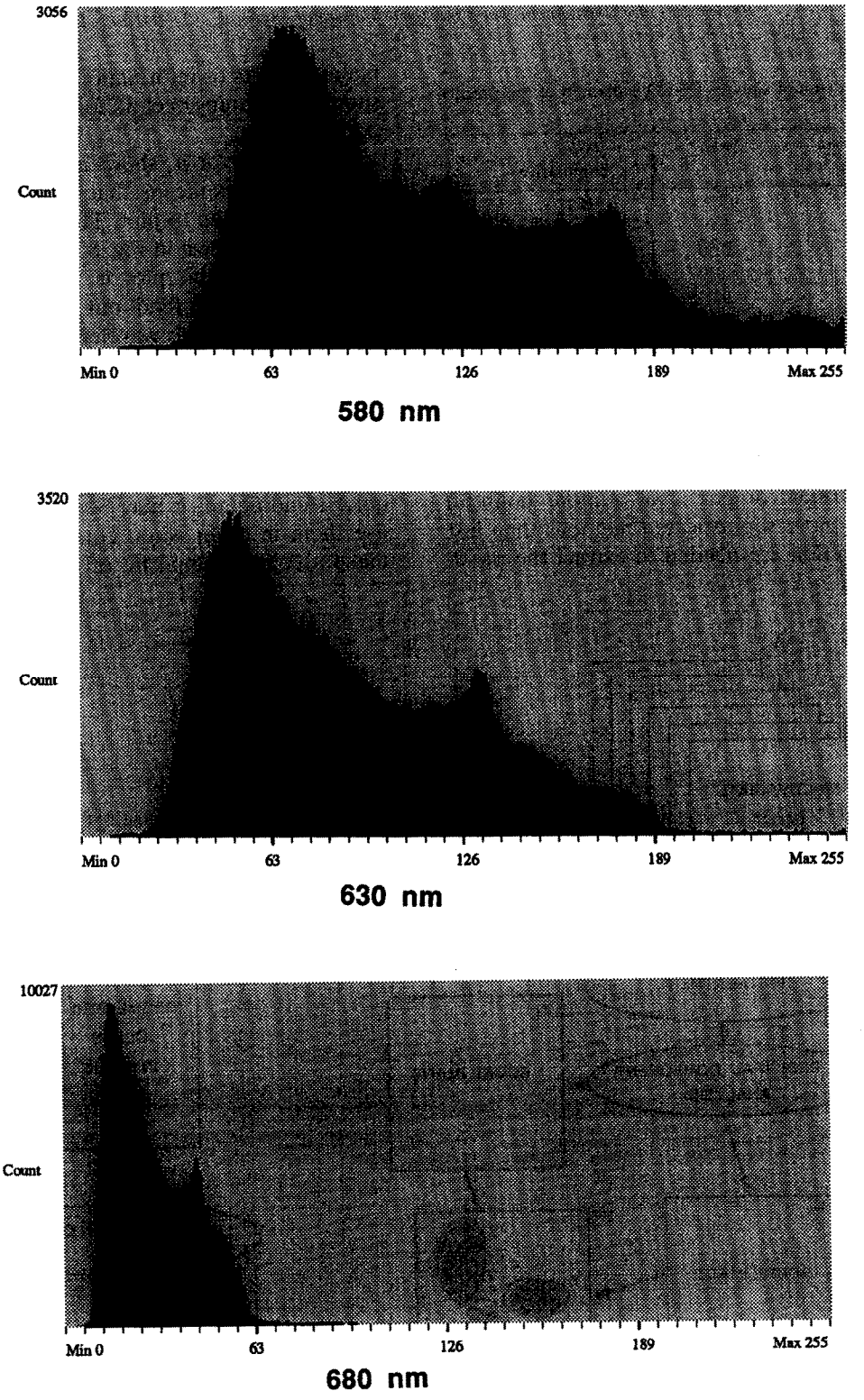


Fig. 4. The intensity histograms for the six untreated wavelength bands (images). The histograms are clearly multimodal. The histogram at 680 nm contains only dark pixels. The original image at 680 nm was at the limit of illumination and camera sensitivity.

Table 3
Percentage of sum of squares (%SS) explained of the preprocessed array

Component No.	%SS	%SS cumulative
1	94.1	94.1
2	5.0	99.1
3	0.59	99.7
4	0.22	99.9
5	0.035	99.98
6	0.022	100

values can always be made available. It is clear that the score images for components 1, 2 and 3 have good visual content. For the fourth component, the visual content is of a more doubtful quality and more sophisticated methods than just visual inspection are needed to extract the possible useful bits of information from them. Table 4 shows the loading vectors for the first three components.

Loading vector p_1 shows almost constant loading values. This means that this loading will not be too interesting to plot. The loading plot of p_3 against p_2 is shown in Fig. 8. The most important aspect seen in this plot is that the wavelength bands are well spread out and appear in sequence, according to wavelength. This sequential behaviour is to be expected from spectral data where each wavelength is correlated with its neighbours. All wavelengths seem to contribute reasonably well to the three first components.

Although the score images themselves reveal quite some detail, it may be more interesting to use them in other ways. One possible way is to make a colour composite of the first three com-

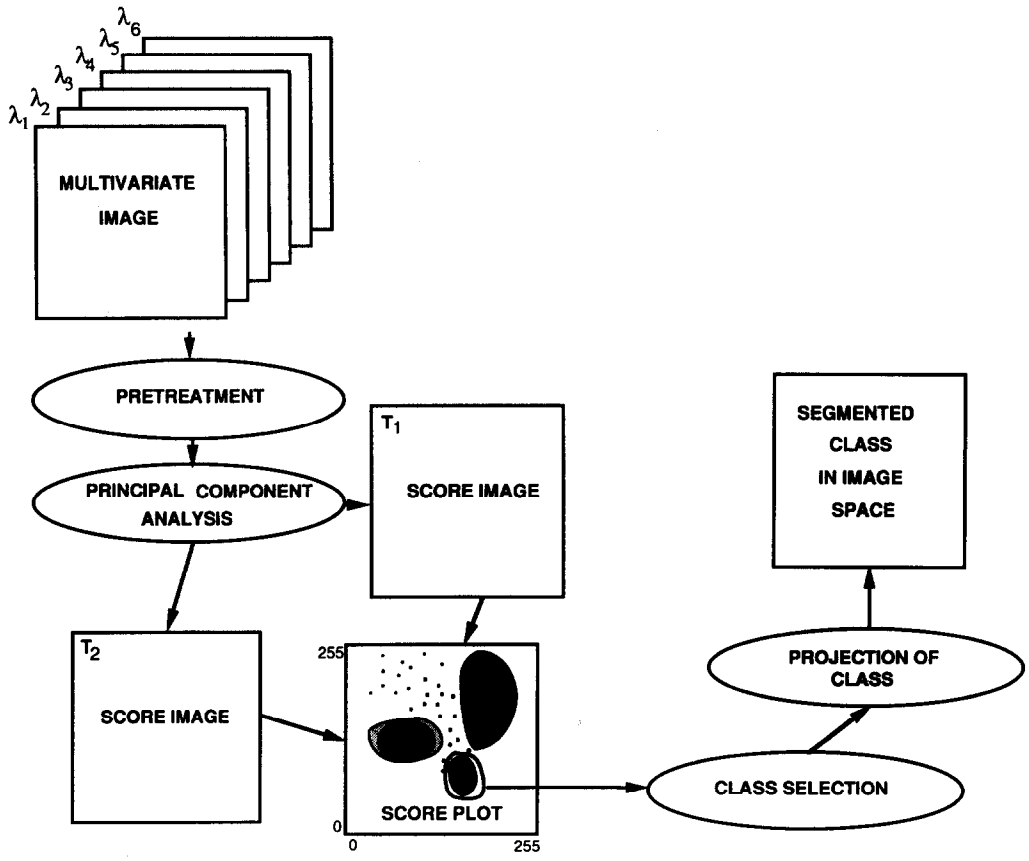


Fig. 5. An overview of the operations of multivariate image analysis leading to classification by feature space segmentation.

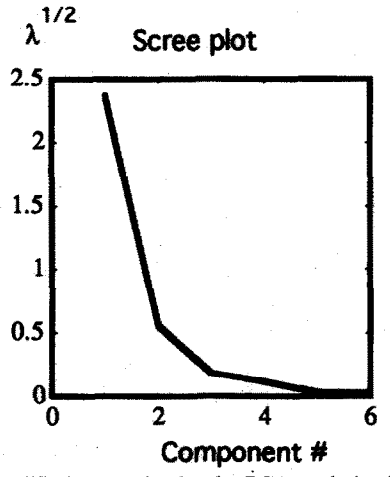


Fig. 6. A modified scree plot for the PCA analysis after mean centering and variance scaling. It is clear that three components explain almost 100% of the sum of squares.

ponents in the red, green and blue colour planes of the monitor. An example of this is shown in Fig. 9. Since the first three components explain very much of the data structure, their colour composite gives a good subjective view of the details in the painting, although not in the colors

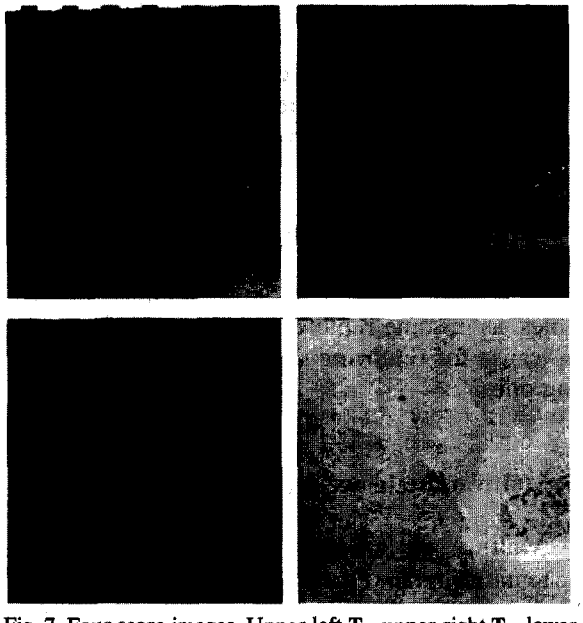


Fig. 7. Four score images. Upper left T_1, upper right T_2, lower left T_3 and lower right T_4. The scores were transformed to fill the intensity range 0–255 optimally.

Table 4
The first three loading vectors of the pretreated 6×512×512 image

Band No.	p_1	p_2	p_3
1	0.401	−0.499	0.623
2	0.409	−0.421	−0.047
3	0.415	−0.220	−0.560
4	0.417	0.152	−0.409
5	0.408	0.439	0.079
6	0.399	0.556	0.350

that the human eye expects from prior knowledge of the object (compared to Fig. 3b).

A more objective way of finding information in the score images is by constructing their scatter plots, called score plots. These are shown in Fig. 10 (left side). The score plots show pixel clusters, with colour indicating density. Quite a number of dense clusters, gradients and outliers can be observed. These can be used for explaining almost all the details in the multivariate image. This technique is called multivariate image segmentation.

5. Variable reduction

It can sometimes be argued that the tri-variable colour image of an object shows all the detail

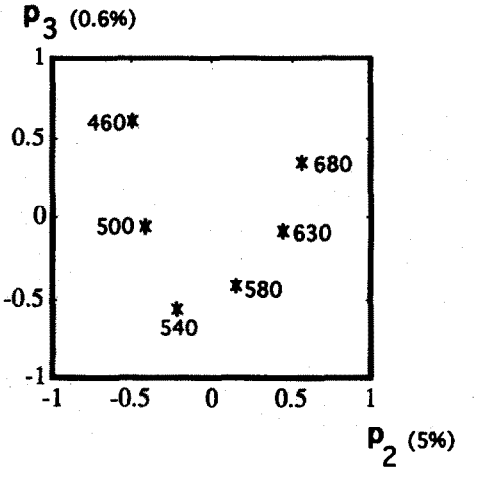


Fig. 8. The loading plot for components 2 and 3. The wavelength bands are circularly spread and in sequence.

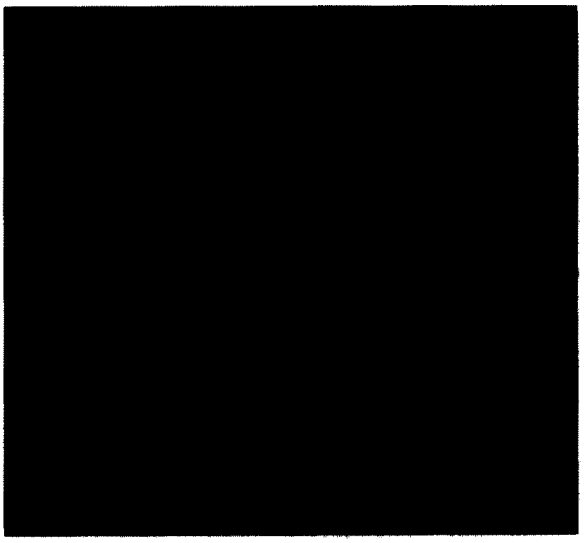


Fig. 9. A false colour composite image of the three score images T_1 = red, T_2 = green, T_3 = blue.

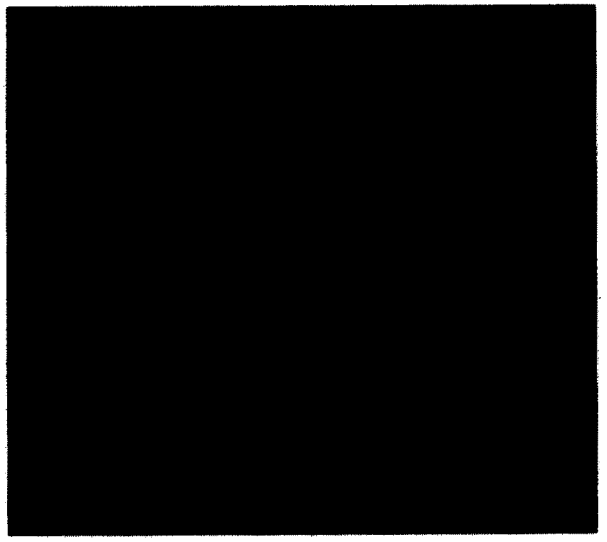


Fig. 11. The scatter plot of scores T_1 and T_4 indicates that for certain groups of pixels there is information to be found in component 4.

that is necessary. Therefore, variable reduction was tried by only using the red (630 nm), green (540 nm) and blue (460 nm) wavelength bands that constitute a colour image. The results of a PCA analysis after z-scaling, which is the same as mean-centering and variance scaling, are shown in Fig. 10 (right side). At first it may seem that both the score plots for the six-variable model and for the three-variable model show the same structure of classes, gradients and outliers, even though some small differences are noticeable. This is to be expected, since the painting is meant for the human eye that observes colours as tri-stimulus red-green-blue combinations. Fig. 11 shows the score plot of components 1 and 4 for the six-variable case. There is a clear indication that also the fourth component is important in separating classes and creating gradients. This makes it important to include more wavelength bands than just three. In general, when only few variables are available it is to be recommended to use them all. If there is an abundance of variables, as in some remote sensing cases, there may be cause for removing some of the least interesting ones.

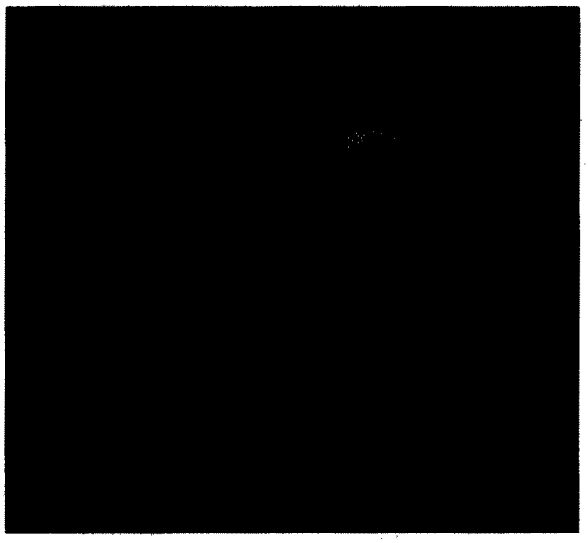


Fig. 10. The score plots. Upper left quadrant: T_2 against T_1. Lower left quadrant: T_3 against T_1. Right side: the same plots repeated for a simplified model with only three variables.

6. Multivariate image segmentation

Multivariate image segmentation is shown schematically in Fig. 5. It has been explained earlier [1–8] and there is no need to repeat it in detail. One special type of segmentation is highlighted here. It is shown in Fig. 12a. It is a special

case, where two classes of pixels, defined in two different score plots, are combined by intersection of the classes to make a very powerful segmentation device. In the case of Fig. 12a, the class 'bright red' has been segmented in the T_1–T_2 and T_3–T_4 score plots. Segmentation in only one of the score plots would have led to much more misclassification in this case. The technique presented here uses two tubes of irregular shape (see Fig. 12a for their cross-sections) in hyperspace

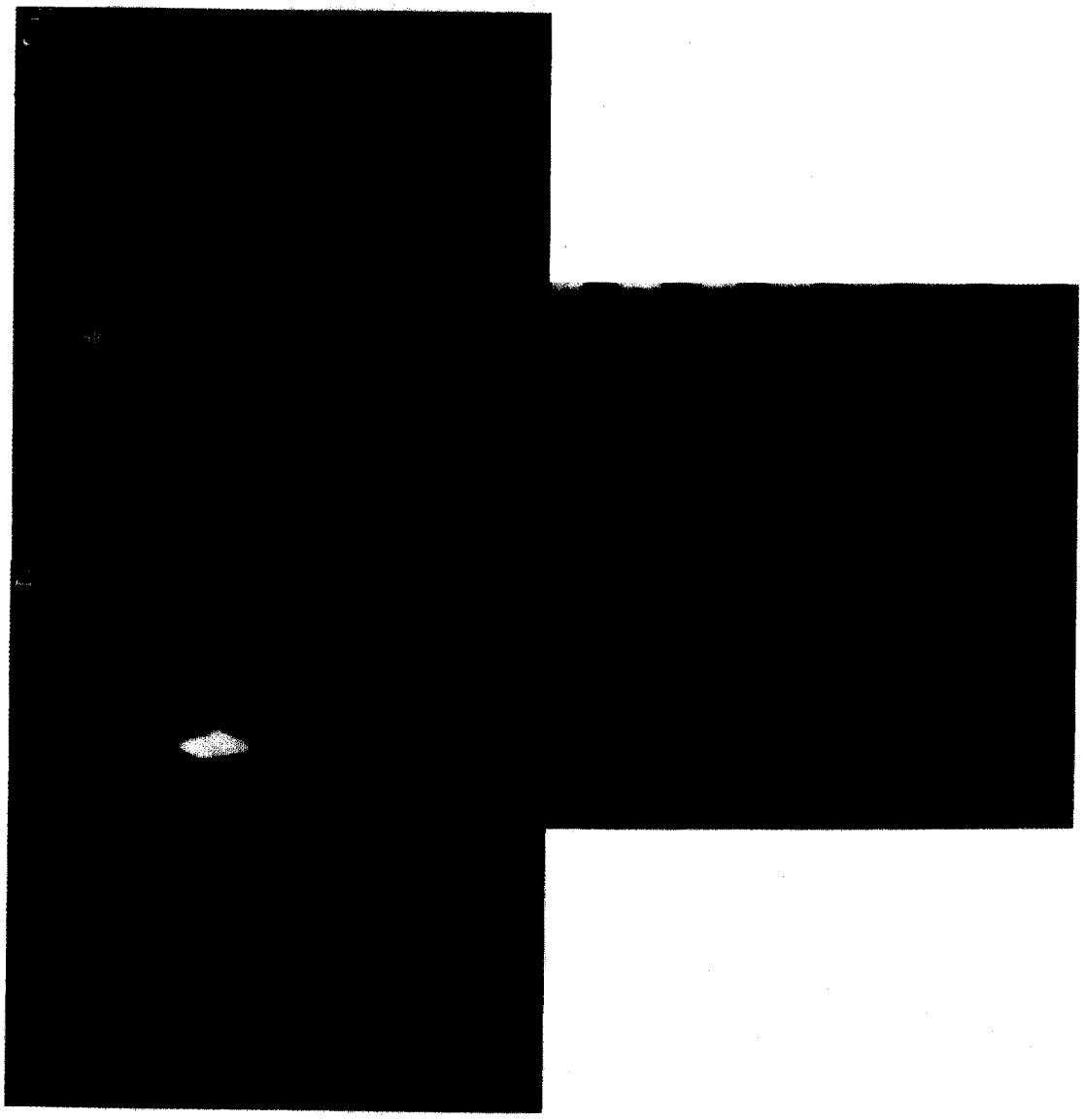


Fig. 12. (a) A class mask in the T_1–T_2 score plot and one in the T_3–T_4 score plot are combined by intersection to give a well-defined class of pixels having the characteristic 'bright red'. This class is shown as an overlay on top of the T_1 image (right side). (b)–(k) Classes defined in the score plots T_1–T_2 ((b)–(g)) and T_1–T_3 ((h)–(k)) are shown as binary images. (b) White porcelain background. (c) Extra reflection in some corners. (d) The class 'gold' as a spot with a linear stroke through it and on the edge on the leaf LL. (e) Leaf UR as a homogeneous class. (f) Leaf LL as a non-homogeneous class. (g) Light grey strokes in the middle of the studied area. (h) The class 'gold' in another segmentation. (i) The class 'red' as in (a) in a different segmentation. (j) Showing that leaf LL is not a homogeneous enamel. (k) Weak dark lines under leaf UR.

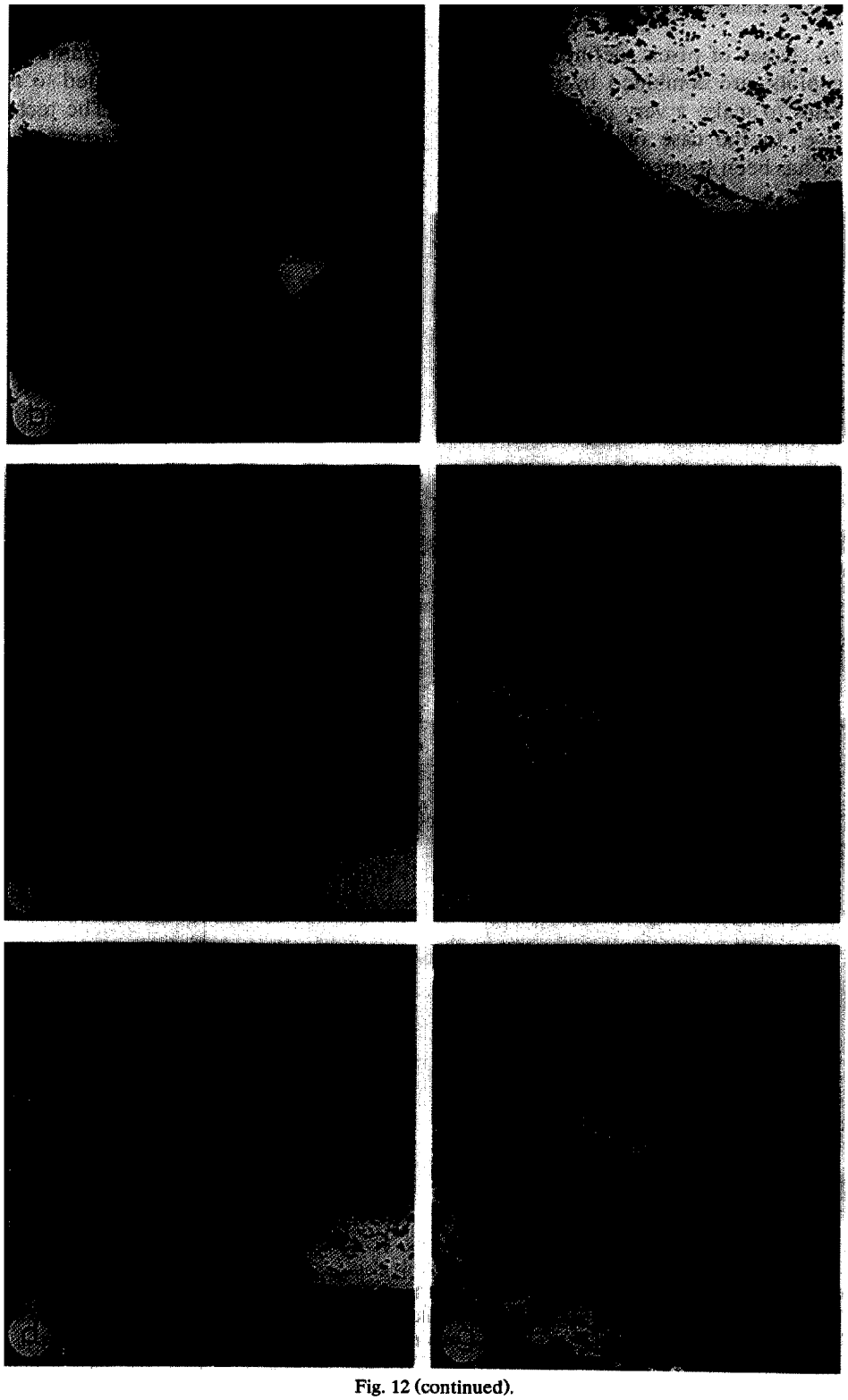


Fig. 12 (continued).

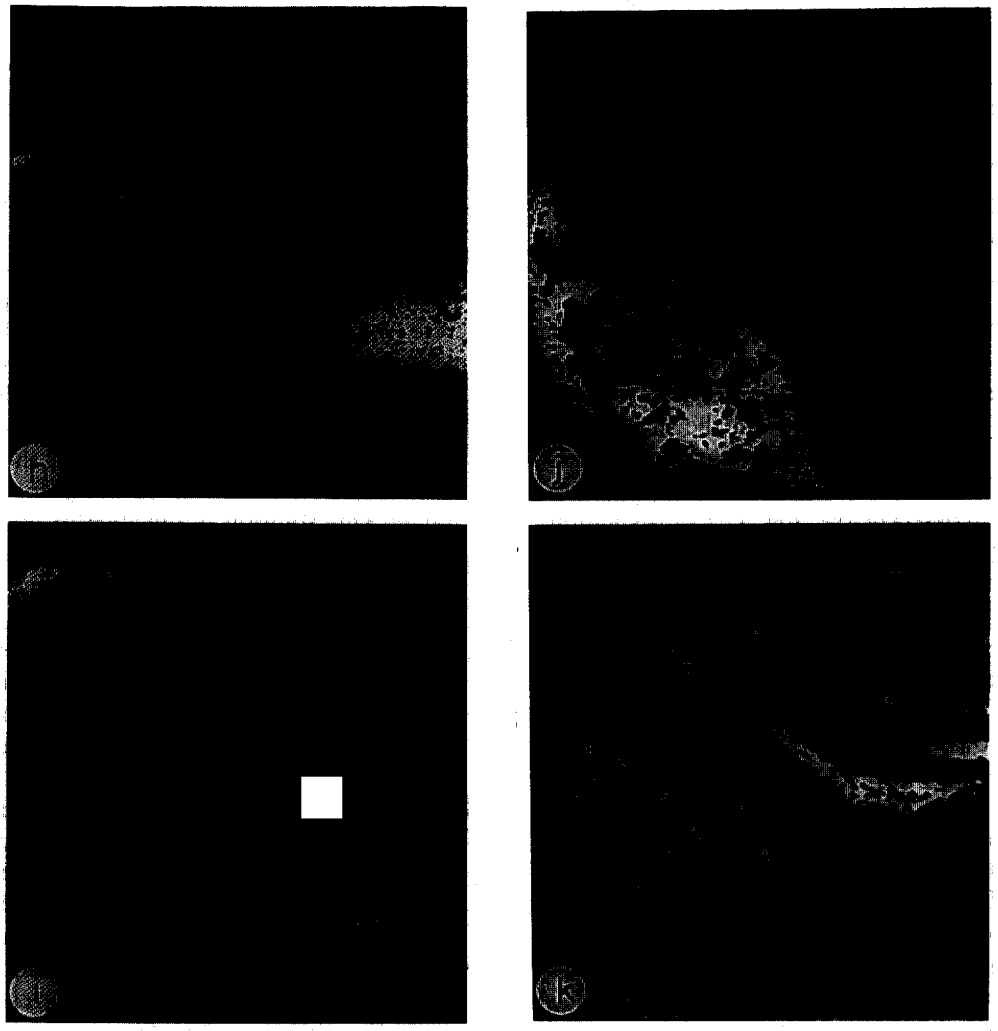


Fig. 12 (continued).

and makes a class of the intersection part of the two tubes. The mentioned shapes of the tubes are selected manually by visual interactive inspection of the score plots.

A complete and detailed segmentation of the image is possible. It leads to the detection of pure and mixed classes, and also illumination and reflection errors. The results of a first attempt at segmentation are shown in Figs. 12b–12k. These results are obtained by selecting classes in the T_1–T_2 and T_1–T_3 score plots and projecting them to image space as binary images. The effect of the millimeter scale highest up in the image was ignored. First the classes found in the T_1–T_2 score plot are treated. In Fig. 12b, the white porcelain background is shown. This background area is remarkably smaller than expected from inspection of image 3b. This means that the enamel is more abundant than would be noticed at first sight. In Fig. 12c, the extra bright reflection in the lower right corner of the object is segmented out. This was confirmed by visual inspection of the piece and the images. Fig. 12d shows the class of the gold. It is cut in half by a stroke in the lower part. Also some gold at the edge of a leaf in the lower left corner (from now

on called leaf LL) is shown. Fig. 12e shows the class formed by the leaf in the upper right corner (leaf UR). In this segmentation, it looks like a homogeneous class. Fig. 12f shows how the leaf LL is not homogeneous, but a complex mixture of different enamels. Fig. 12g shows the segmentation of some very weak (light grey) linear strokes. These strokes may have been black originally, but they may have faded by erosion. The strokes are not a pure class. A lot of misclassification is found. This is so because the classes overlap in the score plot. Even with overlap in the score plot, classes can still be separated visually by using the spatial information in the image space to find out about class membership and misclassification. The classes found in the $\mathbf{T}_1$–$\mathbf{T}_3$ score plot are discussed below. Fig. 12h shows another segmentation of the class 'gold' that is somewhat different from that in Fig. 12d. This is an indication that it may be possible to combine the classes by intersection and get a better result. Fig. 12i shows a segmentation of the class 'red'. This one can be compared to the segmentation in Fig. 12a. Fig. 12j shows a segmentation of the leaf LL. This gives again an indication of the complex composition of this leaf. Fig. 12k shows some of the dark linear strokes that are found under leaf UR. The class is difficult to define and some incomplete classification and misclassification are seen.

It becomes clear from this simple segmentation example that multivariate image segmentation is a valuable addition to visual inspection. A more detailed segmentation would be possible after the first round, but it is left out because of space limitations. Many results were found by an analyst with no experience in chinaware studies and later confirmed to be true by experts on chinaware. It is expected that also the experts can benefit from the increased detail shown by multivariate image analysis.

7. Multivariate image regression

It is very often important to relate the content of a multivariate image to external information. In these cases it is convenient that this external information is also available in image form. This allows the setting up of a regression model.

Regression is a tool for building relations between data sets:

$$\mathbf{Y} = \underline{\mathbf{X}} \ \underline{*}\, \mathbf{b} + \mathbf{F} \tag{2}$$

where $\mathbf{Y}$ is an image (size $I \times J$), $\underline{\mathbf{X}}$ is a multivariate image (size $K \times I \times J$), $\boldsymbol{b}$ is a vector of regression coefficients: the regression vector (size K), $\mathbf{F}$ is the residual image (size $I \times J$), and $\underline{*}$ is a three-way operation defined in Ref. [15].

When $\mathbf{Y}$ is a binary image, consisting of discrete values (0 for low and 255 for high), the method is often called discriminant regression.

There are different ways of calculating the vector of regression coefficients $\boldsymbol{b}$. When $\mathbf{F}$ is minimized in least-squares fashion, the multiple linear regression (MLR) solution for $\boldsymbol{b}$ is obtained. This solution is sensitive to collinearities in the variables. Therefore biased methods of calculating $\boldsymbol{b}$ are often used. The biasing usually leads to vectors $\boldsymbol{b}$ with a reduced norm. The term 'shrinkage' is therefore used for describing them.

Good alternatives to MLR are the latent variable regression methods: principal component regression (PCR) and partial least squares regression (PLS). PCR on images is explained in Refs. [15,16] and PLS on images is explained in Refs. [16–18], both for discrete and continuous y variables. In latent variable regression models, a number of latent variables is chosen for the model. This number is usually called the number of components or also the rank of the model.

An alternative method is ridge regression (RR) where the following equation is used to get a biased vector of regression coefficients (assuming the images are represented by the matrix $\mathbf{X}$ (size $[I \times J] \times K$) and the vector $\boldsymbol{y}$ (size $[I \times J]$) for ease of notation):

$$\mathbf{b} = (\mathbf{X}'\mathbf{X} + c\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \tag{3}$$

$\mathbf{I}$ is the identity matrix of size $K \times K$ and c is a constant that can be changed continuously to find ridge regression models of different quality, preferably good ones [19]. The nice aspect of Eq. (3) is that $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\boldsymbol{y}$ have to be calculated only once and are very small for the given example. This makes it extremely easy to recalculate $\boldsymbol{b}$ for different values of the ridge parameter c. An algorithm is given in the Appendix.

Fig. 13 shows that a part of the image ($6 \times 512 \times 50$) was defined as a calibration set. The binary image from the feature space segmentation of Fig. 12a was used as a 512×50 y image. The remainder of the image ($6 \times 512 \times 462$) can then be considered as a test set. The calibration data are shown in Fig. 14. The binary variable to be predicted is the property 'bright red' segmented in Fig. 12a.

Principal component regression and partial least squares regression were carried out on the calibration image. As preprocessing, mean-centering but no rescaling of variances was used. The reason for this is that the scaling methods for the different regression techniques to be compared were not identical. The emphasis is on the PLS results. A first result is concerned with model building statistics. These are given in Table 5 for all six PLS models of rank 1–6. The results in the table are also given in a more useful form as plots in Fig. 15. The PLS weights are given in Fig. 16 for x and y variables.

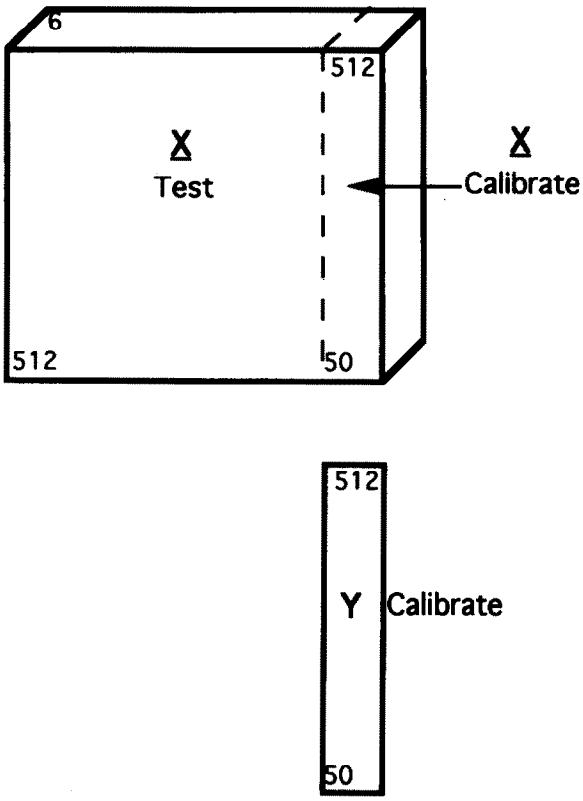


Fig. 13. A $6 \times 512 \times 50$ subimage is used as calibration data in a discriminant regression analysis. The y calibration data are created by multivariate image segmentation. It is a binary image. The remainder of the $6 \times 512 \times 512$ image is used as a test set.

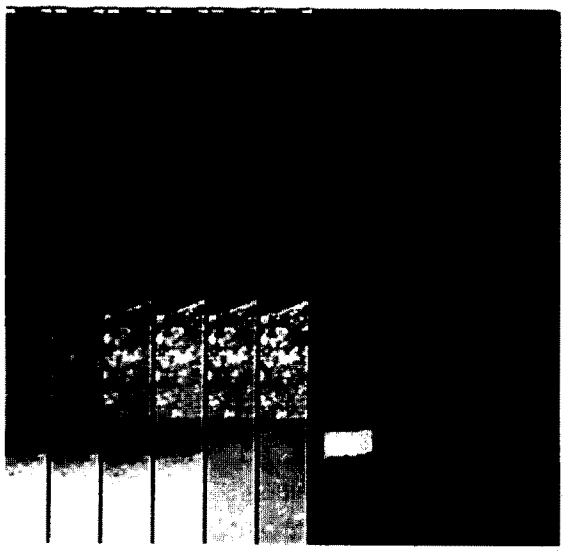


Fig. 14. The six x-variable images (left to right $\mathbf{X}_1$ to $\mathbf{X}_6$) and the binary y-variable image (extreme right), all of size 512×50.

The cross-validation value *XVAL* is calculated for each PLS component. It is described in more detail in the literature [20,21]. The most important point to remember here is that values much lower than 1 indicate a useful component and that values that approach 1 are indicators of a less useful component. It is important that the value of *XVAL* is not interpreted just by itself, but related to the amount of the sum of squares

Table 5
PLS regression between a $6 \times 512 \times 50$ image and a 512×50 binary image. Results

Component No.	*XVAL* per comp.	*%SSX* cumulative	*%SSY* cumulative
1	0.62	6.5	39.5
2	0.98	99.5	40.8
3	0.57	99.92	66.5
4	0.90	99.98	69.8
5	0.96	99.9	71.1
6	1.00	100	71.1

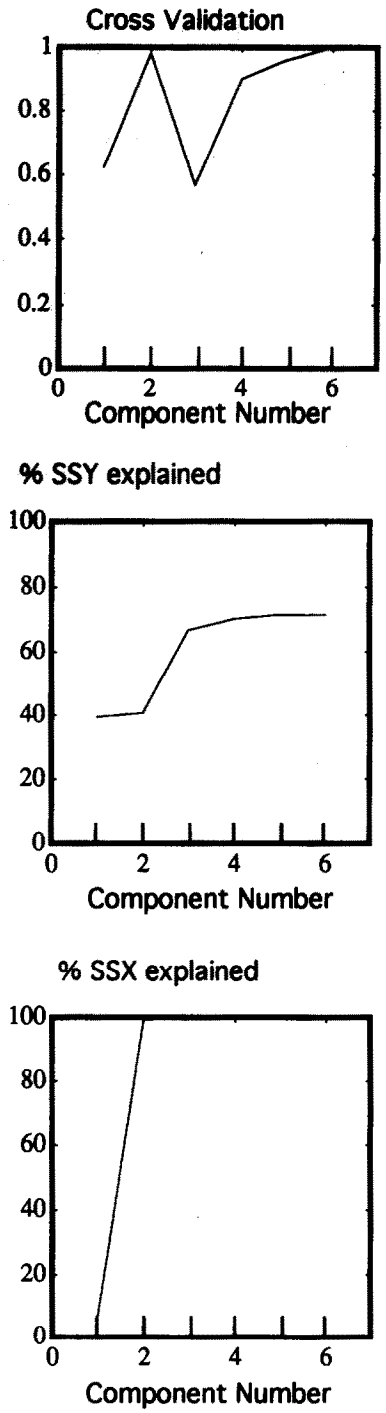


Fig. 15. Information about the calibration for PLS models with one to six components: Cross-validation criterion, percentage *SSY* explained and percentage *SSX* explained (cumulative).

(*SS*) that each component uses for x and y variables.

It can be noticed that the cross-validation value *XVAL* is low for component 1, high for component 2, low again for component 3 and then slowly moves up for the subsequent components. In the plots of the percentage of sum of squares (*SS*) explained, it can be seen that the second PLS component explains a large part of the *SS* of **X** while it explains almost nothing of that of **Y**. This is the reason why the value of *XVAL* is so high for the second component. A general conclusion is that three or four components seem to form an adequate rank explaining most of the *SS*

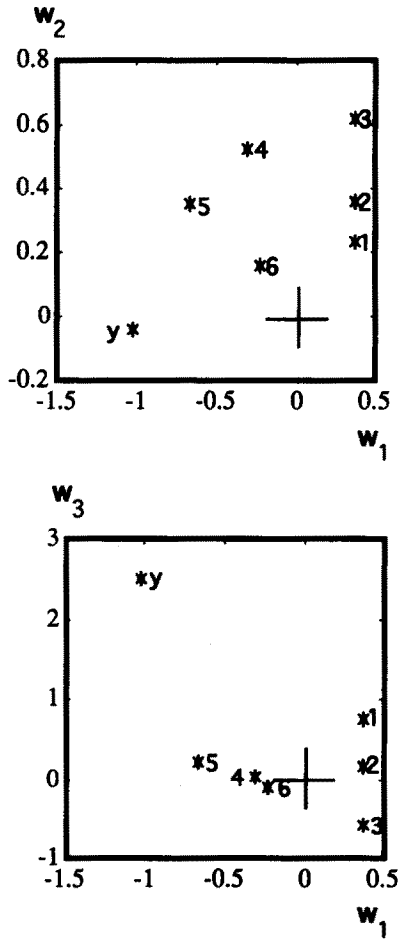


Fig. 16. PLS weight plots including the weight (loading) for the y variable for PLS components w_1, w_2 and w_3.

Fig. 17. Prediction images for both calibration and test image for models of rank 1 (upper left), 2 (upper right), 3 (lower left) and 4 (lower right).

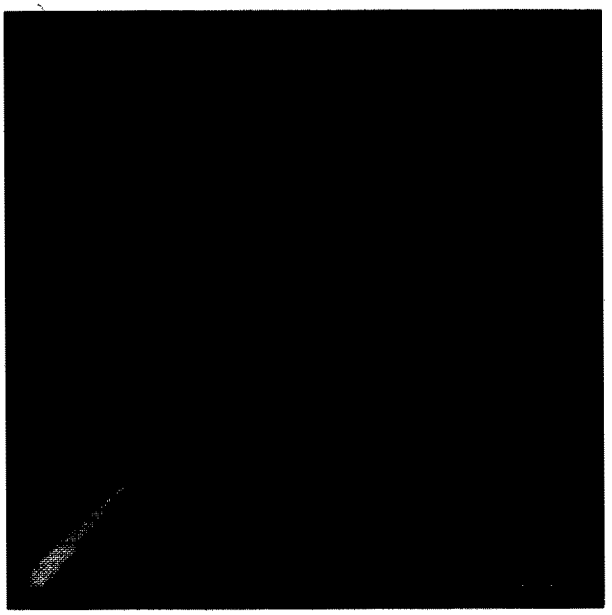


Fig. 18. A scatter plot of predictions (calibration + test) for rank 1 model (horizontal) and rank 2 model (vertical). The scale goes from 0 to 127. The models are almost identical, since the second PLS component has no predictive power.

in **X** and close to 70% of the *SS* of **Y**. Cross-validation is not an absolute method. It gives an indication of how good a component is, but the results have to be interpreted and in many cases, a few models of different rank are almost indistinguishable. It should also be borne in mind that all validation should be based on a careful choice of validation and test sets. This careful choice is not possible in all situations.

PLS weights are plotted in Fig. 16. Also here, it can be seen that the second PLS component has no contribution to the regression model, since it has a very low y weight. For PLS-components 1 and 3, the y weight is rather large, indicating a good explanation of y in the model.

Calculated values for the calibration set and predicted values for the test set for models of rank 1–4 are shown in Fig. 17. The predictions for rank 1 and rank 2 models look almost identical. The predictions for the low rank models are not very good. Many regions not related to the property of interest 'bright red' are showing high intensities. Predictions for rank 3 and 4 models are better. Anything that has absolutely nothing to do with the property of interest remains black

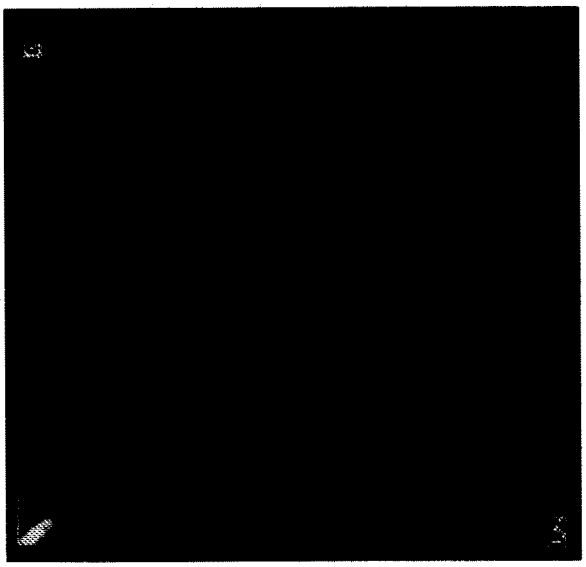


Fig. 19. A scatter plot of predictions (calibration + test) for rank 3 model (horizontal) and rank 4 model (vertical). The scale goes from 0 to 255. A reasonable correspondence is observed.

and the regions that are showing as bright or median grey have a constituent that is related to the property 'bright red'. This can be confirmed by careful re-inspection of the colour image in Fig. 3b. In Fig. 17, the rank 3 and 4 predictions, two large regions (linear strokes) of high intensity can be seen. These correspond to the visible bright red lines in the color image 3b. Also two other smaller regions are shown as bright grey. They are positioned close to the lower left corner and in the middle of the right side. They belong to red lines that continue outside of the sampled region. This fact was only discovered after the discriminant regression results were inspected.

For comparing models of different rank it is useful to look at scatter plots of predicted images. Fig. 18 gives the scatter plot for the predicted images using PLS models with rank 1 (horizontal) and rank 2 (vertical). All the pixels are on a straight line. This is because the rank 2 model is no improvement over the rank 1 model as explained earlier. Fig. 19 gives the scatter plot of predicted images for PLS models of rank 3 and rank 4. There seems to be a reasonable match between the two images, with all pixels lying close to a straight line.

This section mainly reported the results of PLS regression. PCR and RR regression models are taken up in more detail in the next section.

8. Comparison of regression models

It is sometimes necessary to compare regression models in order to find out which ones are best and which ones are similar. This can be done by calculating statistics on the regression model or by cross-validation. Some ways of comparing models of different rank were shown earlier for the PLS models of rank 1–to 4. It was also found that rank 3 and 4 PLS models seem to be the best possible choice here. A more visual method of comparing regression models is introduced here. A regression model, when used for prediction, is characterized by a regression vector $\boldsymbol{b}$. It is possible to compare these regression vectors. This is shown schematically in Fig. 20. The regression vectors for six PLS models and six PCR models of

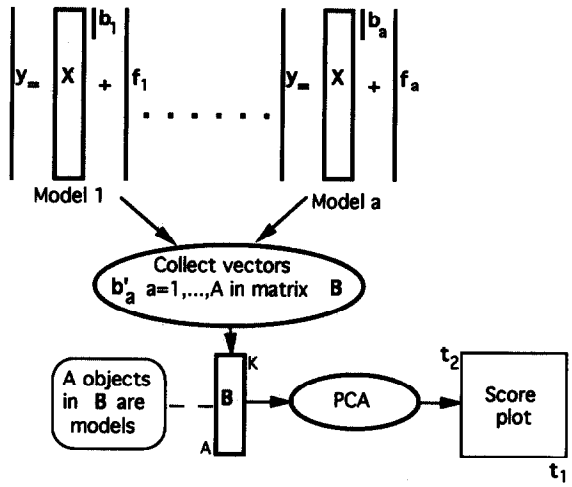


Fig. 20. The principle of comparing regression models by comparing their regression vectors.

different rank are given in Table 6 as a 12×6 matrix. The regression vectors are for mean-centered but not rescaled data. The rank 6 solutions for PLS and PCR are identical and also identical to the MLR (multiple linear regression) solution. The matrix can be analyzed by singular value decomposition, leading to eigenvectors or scores. The score plot for these is given in Fig. 21. In this score plot, each model has a position, so that similarity, dissimilarity and clustering can be detected. It is possible in such a plot to define a region where all the best models cluster together.

Table 6
Regression coefficients for PLS and PCR models of different rank

	1	2	3	4	5	6
PLS_1	−0.3660	−0.3760	−0.3700	0.3210	0.6920	0.2430
PLS_2	−0.3870	−0.4020	−0.4060	0.3110	0.7000	0.2450
PLS_3	1.1550	−0.4370	−2.5260	0.2450	1.4450	0.0640
PLS_4	0.5700	0.2590	−3.5990	1.9780	0.7610	−1.3470
PLS_5	−0.6700	2.5970	−4.3320	1.4100	1.5870	−2.1810
PLS_6	−0.6980	2.6860	−4.4030	1.4510	1.4990	−1.9920
PCR_1	0.0056	0.0051	0.0056	0.0069	0.0055	0.0017
PCR_2	−0.5120	−0.3660	−0.1630	0.2980	0.5740	0.2440
PCR_3	1.2080	−0.5330	−2.3030	−0.4590	1.5470	0.2960
PCR_4	1.1550	−0.7670	−3.0950	1.8360	0.5780	−0.7200
PCR_5	−0.4630	2.0660	−3.9810	1.2850	1.8940	−2.8730
PCR_6	−0.6980	2.6860	−4.4030	1.4510	1.4990	−1.9920

Also regions of underfitting and of mild overfitting can be defined. Overfitting models are positioned towards the MLR solution. Underfitting models are situated closer to the one-component PLS and PCR solutions.

The plots show how the PCR models of rank 1 and 2 show underfitting. The PLS models of rank 1 and 2 are almost identical and take up the same place in the plot. PLS models of rank 3 and 4 are good models. The PCR models of rank 3 and 4 are closer to the PLS model of rank 3. These models of rank 3 and 4 constitute an area of 'good' models, as shown by cross-validation in PLS. The rank 5 and 6 models for PLS and PCR are further away from the 'good' solutions. This would in many cases be considered as overfitting.

Ridge regression models were also constructed for different ridge parameters and the corresponding regression coefficient vectors were calculated. This was done in Splus. $\mathbf{X}'\mathbf{X}$, the cross product matrix of the mean-centered data, is equivalent to the covariance matrix. Because of the use of this covariance-equivalent matrix instead of the correlation matrix, large ridge parameters were used, ranging from 1000 to 5×10^6.

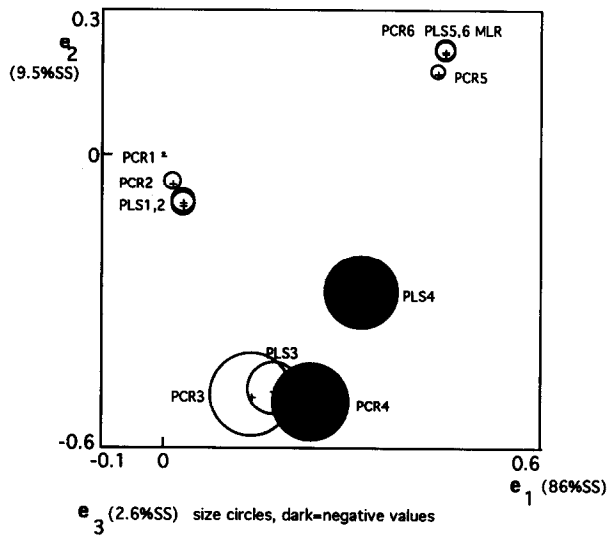


Fig. 21. Scatter plot for the eigenvectors of the matrix of PLS and PCR regression vectors, with eigenvector 3 indicated as sizes of the disks. The first eigenvector is very much related to regression vector size or with 'shrinkage'.

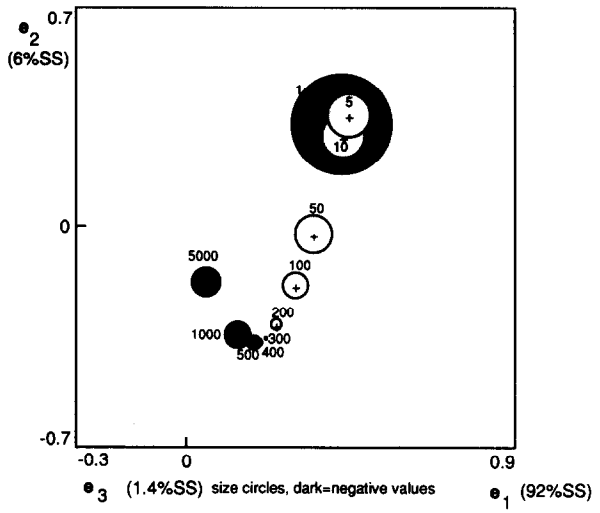


Fig. 22. Scatter plot for the eigenvectors of the matrix of ridge regression vectors. The first eigenvector is related to regression vector size or 'shrinkage'.

The ridge based correlation vectors can be included in the comparison of PLS and PCR models. While the different PLS and PCR models have a discrete parameter called rank, the ridge models have the continuous parameter c.

Fig. 22 shows a plot of the first three eigenvalues of the matrix in Table 7. There is clearly a progression from overfitting ($c = 1K$, close to the MLR model, see Table 6) over some good models ($c = 100K$ to $c = 500K$) to underfitting for too large parameters. A prediction image for the RR model with $c = 300K$ was made. It looked very

Table 7
Ridge regression coefficients for selected values of the parameter c. c is expressed as multiples of 1000

c	1	2	3	4	5	6
1	−0.6680	2.6270	−4.3750	1.4480	1.4890	−0.1972
5	−0.5550	2.4130	−4.2710	1.4390	1.4540	−1.8950
10	−0.4340	2.1820	−4.1560	1.4270	1.4150	−1.8060
50	0.1160	1.1240	−3.5800	1.3020	1.2440	−1.3090
100	0.3890	0.5730	−3.2030	1.1510	1.1720	−0.9650
200	0.5740	0.1330	−2.7820	0.9250	1.1330	−0.6070
300	0.6190	−0.0490	−2.5210	0.7760	1.1180	−0.4170
400	0.6180	−0.1450	−2.3300	0.6730	1.1060	−0.2970
500	0.6000	−0.2030	−2.1790	0.5990	1.0910	−0.2150
1000	0.4600	−0.3070	−1.6950	0.4150	1.0100	−0.0210
5000	−0.0070	−0.2980	−0.6970	0.2320	0.6520	0.1380

much like the one for rank 4 PLS. When plotted against each other, PLS rank4 against RR $c = 300K$, all pixels are on a straight line at 45°. These figures were left out to save space.

The Tables 6 and 7 can be combined into one large table with 23 regression vectors. The first two eigenvectors of this table are plotted in Fig. 23. This figure shows how PLS and PCR models of different rank are related to RR models with different c values. Regions of overfit, underfit and of good models can be clearly distinguished. PLS and PCR models are only possible for a discrete rank, while RR models have a continuous parameter. If all RR models were calculated, they would form a continuous curve. In a plot like this, new 'good' values of regression vectors can be selected and their corresponding $\boldsymbol{b}$ vector can be calculated. An infinite number of 'good' vectors $\boldsymbol{b}$ can be found that do not have to correspond to any PLS, PCR or RR model. It may be possible to run a simplex or genetic algorithm to find some 'best' vector $\boldsymbol{b}$ in the region of the good PLS, PCR and RR models. In Fig. 24, the above comments are repeated. Regions of 'good', overfitting and underfitting are indicated. Some models take on an intermediate position.

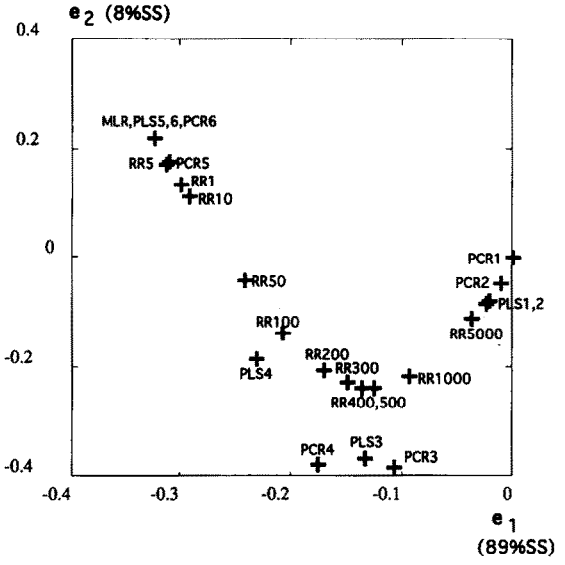


Fig. 23. By plotting two eigenvectors of the matrix of all regression vectors, a plot showing the relationships between regression models is obtained. It is clearly visible how regions of overfit, underfit and good models are present. By choosing the rank of the PLS or PCR model or the ridge coefficient, a good model can always be obtained.

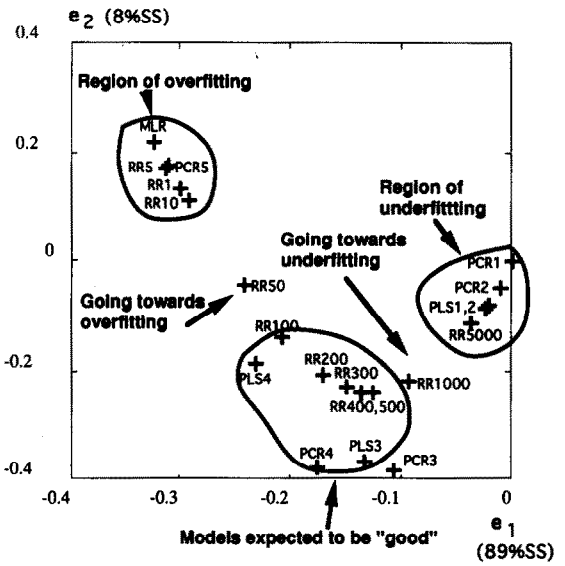


Fig. 24. An annotated version of Fig. 23. More comments can be found in the text.

9. Conclusions

It was shown that multivariate microscopy by constructing a $6 \times 512 \times 512$ image of a painted ceramic surface can help in the spatial and spectral analysis of the artwork. Univariate statistical analysis of the images is a helpful tool for detecting experimental problems and data file errors. Because of the correlated nature of the wavelength bands, multivariate analysis is a natural choice. Principal component analysis with visualization of the results leads to the detection of classes and outliers. A special type of multivariate image segmentation is introduced. It gives better segmentation by making intersections of feature space classes. It is also shown that variable reduction can lead to all major interpretations, but that finer detail is lost.

Discriminant regression of one specific class is used to explain the technique and to introduce visualization explaining the regression models and

their predictions. A technique for comparing regression models more easily is introduced and explained. PLS, PCR and RR can all lead to good models, provided the right rank (discrete) or ridge parameter (continuous) is selected. Also new regression vectors with 'good' properties can be selected in the region where the best PLS, PCR and RR models are found in the regression vector comparison plot. This technique needs further development. The role of the second and third component and the right numbers of components to be used have to be investigated. Also, other alternative regression models to the ones used here may be added in future work.

The example shows that multivariate microscopic analysis of fine details in any object of interest may be something to pursue in order to study the finer details. Including more wavelengths inside and outside the visual range would be even more instructive.

Acknowledgements

Fredrik Lindgren acknowledges the Swedish Work Health Fund for financial support. Nicole De Biscop is thanked for her valuable advice on the painting of Chinaware.

Appendix

An algorithm for calculating ridge regression vectors

The algorithm is given in the high-level language Splus for Unix. This language allows large data files to be handled very easily. The program given here is not optimized for speed or convenience, but it is given as an aid in understanding the operations carried out.

1. Assume that all images are given as ASCII files. Read these files into six x variables and one y variable. Each image in the calibration set would be an ASCII file of 512×50 integers, separated by blanks.

```
> x1 < -scan("china1.asc")
> x2 < -scan("china2.asc")
> x3 < -scan("china3.asc")
> x4 < -scan("china4.asc")
> x5 < -scan("china5.asc")
> x6 < -scan("china6.asc")
> y < -scan("china7.asc")
```

2. Each variable is mean-centered by subtracting its mean. The results are floating point values since the means are floating point values.

```
> x1 < -(x1-mean(x1))
> x2 < -(x2-mean(x2))
> x3 < -(x3-mean(x3))
> x4 < -(x4-mean(x4))
> x5 < -(x5-mean(x5))
> x6 < -(x6-mean(x6))
> y < -(y-mean(y))
```

3. Construct a matrix **X** of six variables and $[512 \times 50]$ objects.

```
> x < -matrix(c(x1,x2,x3,x4,x5,x6),ncol = 6)
```

4. Construct **X'X**. The function $t(\ )$ gives the transpose of any vector or matrix between the parentheses. **X'X** is equivalent to a covariance matrix and the values in it are large numbers. %*% is the matrix multiplication operator. The large values in **X'X** mean that the usual small ridge parameters have to be replaced by larger values. In the example, values between 1000 and 5×10^6 were used.

```
> xx < -t(x)%*%x
> xx
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	89076430	78914831	83364289	99405451	76180981	23778799
[2,]	78914831	70698510	75659356	90458048	69619970	21926423
[3,]	83364289	75659356	82859482	100512641	78585906	25157314
[4,]	99405451	90458048	100512641	128192287	103888655	33680238
[5,]	76180981	69619970	78585906	103888655	86972987	28559146
[6,]	23778799	21926423	25157314	33680238	28559146	9526839

5. Construct $\mathbf{X}'\boldsymbol{y}$ in the same way.

```
> xy < -t(x)%*%y
> xy
                [,1]
[1,]        -6312610
[2,]        -6480675
[3,]        -6380525
[4,]         5530434
[5,]        11921811
[6,]         4195025
```

6. Construct the function 'ridge' with the input parameter c. diag(6) is the identity matrix $\mathbf{I}$ of size 6×6. It is multiplied by c and the result is added to $\mathbf{X}'\mathbf{X}$. This has the result of adding the constant c to all diagonal elements of $\mathbf{X}'\mathbf{X}$. solve( ) gives the inverse of what is given between parentheses. This inverse is then multiplied by $\mathbf{X}'\boldsymbol{y}$ to give b.

```
> ridge < -function(c)
+{
+solve(xx + c*diag[6])%*%xy
+}
```

7. Test the function for $c = 0$. The result is that of MLR or full rank PLS or PLR.

```
> ridge(0)
                [,1]
[1,]       -0.6981679
[2,]        2.6855876
[3,]       -4.4029753
[4,]        1.4505803
[5,]        1.4985948
[6,]       -1.9918630
```

8. Test the function for $c = 1000$.

```
> ridge(1000)
                [,1]
[1,]       -0.667586
[2,]        2.627174
[3,]       -4.374853
[4,]        1.448482
[5,]        1.489043
[6,]       -1.971754
```

The function 'ridge' can be used in a loop where different values of c are tried, together with a criterion for deciding when a 'good' model has been found.

References

[1] P. Geladi and K. Esbensen, Multivariate image analysis in chemistry: an overview, in J. Devillers and W. Karcher (Editors), *Applied Multivariate Analysis in SAR and Environmental Studies*, Kluwer, Dordrecht, 1991, pp. 415–445.

[2] P. Geladi, H. Grahn and F. Lindgren, Chemical multivariate image analysis: some case studies, in J. Devillers and W. Karcher (Editors), *Applied Multivariate Analysis in SAR and Environmental Studies*, Kluwer, Dordrecht, 1991, pp. 447–478.

[3] P. Geladi, E. Bengtsson, K. Esbensen and H. Grahn, Image analysis in chemistry. Part 1: Properties of images, greylevel operations, the multivariate image, *TrAC*, 11 (1993) 41–53.

[4] P. Geladi, Some special topics in multivariate image analysis, *Chemometrics and Intelligent Laboratory Systems*, 14 (1992) 375–390.

[5] F. Lindgren and P. Geladi, Multivariate spectrometric image analysis: an ilustrative study with two constructed examples of metal ions in solution, *Chemometrics and Intelligent Laboratory Systems*, 14 (1992) 397–412.

[6] P. Geladi, H. Grahn, K. Esbsensen and E. Bengtsson, Image analysis in chemistry. Part 2: Multivariate image analysis, *TrAC*, 11 (1992) 121–130.

[7] P. Van Espen, G. Janssens, W. Vanhoolst and P. Geladi, Imaging and image processing in analytical chemistry, *Analusis*, 20 (1992) 81–90.

[8] P. Geladi, H. Isaksson, L. Lindqvist, S. Wold and K. Esbensen, Principal component analysis of multivariate images, *Chemometrics and Intelligent Laboratory Systems*, 5 (1989) 209–220.

[9] J. Swerts, A. Aerts, N. De Biscop, F. Adams and P. Van Espen, Age determination of Chinese porcelain by X-ray fluorescence and multivariate analysis, *Chemometrics and Intelligent Laboratory Systems*, 22 (1993) 97–105.

[10] L. Lang, ERDAS delivers interactive remote sensing/GIS, *Advanced Imaging*, November (1992) 69–70.

[11] K. Esbensen, M. Robb, G.-H. Strand and P. Geladi, Software Review: the ERDAS hw/sw system, *Chemometrics and Intelligent Laboratory Systems*, 3 (1989) 95–98.

[12] R. Becker, J. Chambers and A. Wilks, *The New S Language*, Wadsworth and Brooks/Cole, Pacific Grove, CA, 1988.

[13] J. Thioulouse, J. Devillers, D. Chessel and Y. Auda, Graphical techniques for multidimensional data analysis, in J. Devillers and W. Karcher (Editors), *Applied Multivariate Analysis and SAR in Environmental Studies*, Kluwer, Dordrecht, 1991, pp. 153–205.

[14] D. Bright and D. Newbury, Concentration histogram imaging. A scatter diagram technique for viewing two or three related images, *Analytical Chemistry*, 63 (1991) 243A–250A.

[15] P. Geladi and K. Esbensen, Regression on multivariate images: principal component regression for modeling, prediction and visual diagnostic tools, *Journal of Chemometrics*, 5 (1991) 97–111.

[16] K. Esbensen, P. Geladi and H. Grahn, Strategies for multivariate image regression, *Chemometrics and Intelligent Laboratory Systems*, 14 (1992) 357–374.

[17] H. Grahn, N. Szeverenyi, M. Roggenbuck and P. Geladi, Tissue discrimination in magnetic resonance imaging: A predictive multivariate approach, *Chemometrics and Intelligent Laboratory Systems*, 7 (1989) 87–93.

[18] F. Lindgren, P. Geladi and S. Wold, The kernel algorithm for PLS, *Journal of Chemometrics*, 7 (1993) 45–59.

[19] I. Frank and J. Friedman, A statistical view of some chemometrics regression tools, *Technometrics*, 35 (1993) 109–135.

[20] L. Ståhle and S. Wold, Partial least squares analysis with cross-validation for the two-class problem: a Monte-Carlo study, *Journal of Chemometrics*, 1 (1987) 185–196.

[21] F. Lindgren, P. Geladi and S. Wold, Kernel-based PLS regression. Cross validation and applications to spectral data, *Journal of Chemometrics*, 8 (1994) in press.