Galdón-Navarro, B.; Prats-Montalbán, JM.; Cubero-García, S.; Blasco Ivars, J.; Ferrer, A. (2018). Comparison of latent variable-based and artificial intelligence methods for impurity detection in PET recycling from NIR hyperspectral images. Journal of Chemometrics. 32(1):1-14. https://doi.org/10.1002/cem.2980

Journal of CHEMOMETRICS

Comparison of latent variable-based and artificial intelligence methods for impurities detection in PET recycling from NIR hyperspectral images

Journal:	*Journal of Chemometrics*
Manuscript ID	CEM-17-0131.R2
Wiley - Manuscript type:	Special Issue - Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Galdón, Borja; Universitat Politecnica de Valencia, Multivariate Statistical Engineering (GIEM), Departamento de Estadística e IO Aplicadas y Calidad Prats-Montalbán, José Manuel; Universitat Politecnica de Valencia, Multivariate Statistical Engineering (GIEM), Departamento de Estadística e IO Aplicadas y Calidad Cubero, Sergio; Instituto Valenciano de Investigaciones Agrarias (IVIA), Centro de Agroingeniería Blasco, Jose; Instituto Valenciano de Investigaciones Agrarias, Centro de Agroingeniería Ferrer, Alberto; Universitat Politecnica de Valencia, Multivariate Statistical Engineering (GIEM), Departamento de Estadística e IO Aplicadas y Calidad
Keyword:	Multivariate image analysis (MIA), Classification, Pre-procesing, Design of Experiments, Hyperspectral images

Comparison of latent variable-based and artificial intelligence methods for impurities detection in PET recycling from NIR hyperspectral images

B. Galdón-Navarro[1], J.M. Prats-Montalbán[1]*, S. Cubero[2], J. Blasco[2], and A. Ferrer[1]

1. *Multivariate Statistical Engineering (GIEM), Departmento de Estadística e IO Aplicadas y Calidad, Universitat Politècnica de València, Cno. de Vera s/n, Edificio 7A, 46022, Valencia, Spain.*
2. *Centro de Agroingeniería, Instituto Valenciano de Investigaciones Agrarias (IVIA), Cra. Moncada-Naquera Km 5, Moncada, Spain*

Abstract

In Polyethylene Terephthalate's (PET)'s recycling processes, separation from Polyvinyl Chloride (PVC) is of prior relevance due to its toxicity, which degrades the final quality of recycled PET. Moreover, the potential presence of some polymers in mixed plastics (such as PVC in PET) is a key aspect for the use of recycled plastic in products such as medical equipment, toys or food packaging.

Many works have dealt with plastic classification by hyperspectral imaging, although only some of them have been directly focused on PET sorting and very few on its separation from PVC. These works use different classification models and pre-processing techniques and show their performance for the problem at hand.

However, still, there is a lack of methodology to address the goal of comparing and finding the best model and pre-processing technique. Thus, this paper presents a Design of Experiments (DoE)-based methodology for comparing and selecting, for the problem at hand, the best preprocessing technique as well as the best latent variable-based and/or artificial intelligence classification method, when using NIR hyperspectral images.

Keywords: multivariate image analysis (MIA), classification, pre-processing, design of experiments, hyperspectral images.

1 INTRODUCTION

Recycling is becoming more and more relevant in Europe. In [1], the last version of the report, is stated: *"In 2014, 25.8 million tonnes of post-consumer plastics waste ended up in the waste upstream. 69.2% was recovered through recycling (29.7%) and energy recovery (39.5%) processes, while 30.8% still went to landfill… Recycling is the preferred option for plastics waste."*

Within all the different types of plastic, Polyethylene terephthalate (PET) is a key type of plastic, since it is widely used in the production of medical equipment, toys, beverage containers and food storage packages [2]. PET presents great advantages, e.g. keeping its chemical and physical properties, which makes PET the first choice among other plastics. In return, Polyvinyl Chloride (PVC) is a thermoplastic polymer mainly used to produce floors, coverings, window frames, cable insulation, etc. [3], and can be found as part of plastic waste in recycling plants.

Separation of PET from PVC is of primary importance because the latter may generate environmentally hazardous chlorinated compounds that might be risky for humans. This separation is usually carried out manually or taking on complicated mechanical processes because their density is higher than 1 g/cm^3 (PET usually ranges from 1.33 to 1.37 g/cm^3 and PVC ranges from 1.10 to 1.60 g/cm^3), and PET melting point (250-260ºC) is also higher than PVC (140-160ºC) [3]. This way, classification with traditional methods remains a challenge.

One way to overcome this problem is to implement computer vision systems (100% inspection) able to detect and eliminate the PVC from the line. However, RGB cameras are not able to discriminate between PET and PVC. Since these chemical compounds are spectrally different, one possible solution is to use hyperspectral cameras (NIR hyperspectral cameras in this case). Hyperspectral imaging allows overcoming most of the problems linked to plastic separation, such as moisture, plastics densities, or additives in separation by flotation and density, among others [3].

There is a vast amount of companies worldwide aimed at recycling PET by already using hyperspectral machinery with near infrared (NIR) and Raman images, together with multivariate models implemented. Many works have dealt with plastic classification by hyperspectral imaging [4, 5], although only some of them have specially focused on with PET sorting [3, 6-12] and very few with its separation from PVC [3].

Hyperspectral images are usually analyzed by means of multivariate image analysis (MIA) [13, 14] techniques. When working in the pixel domain, MIA can perform different tasks by using different models and related approaches: descriptive analysis or statistical process control by using e.g. principal component analysis (PCA) [15, 16], resolution by using multivariate curve resolution (MCR) [17, 18], prediction by using partial least squares (PLS) [19-22], or classification (as in this case) by using partial least squares - discriminant analysis (PLS-DA) [23] or soft independent modeling of class analogy (SIMCA) [24], among others; by properly unfolding the image [13].

Multivariate statistical and data mining techniques are available when trying to perform classification. Moreover, when dealing with chemical information, different pre-processing techniques can be applied. This paper proposes a Design of Experiments (DoE)-based methodology [25], as a sort of optimization tool for optimal model/pre-processing selection, for the problem at hand. The classification techniques and pre-processing methods have been chosen according to their use in the identification and selection in the plastic field: classical pre-treatments are multiplicative scatter correction [26], standard normal variate [26], and Savitzky-Golay [27, 28] method; whereas most employed classification technique in MIA and in chemometrics in general (also for plastic separation) is PLS-DA [2, 8-11]. In this case classification methods from

classical statistics such as principal component multimonial logistic regression (PCMLR) [29] in one hand, and from artificial intelligence such as nearest centroid [30] and classification tree [31-32] on the other, have been used for comparison purposes with PLS-DA.

In Section 2, the data set used, as well as the hyperspectral pre-processing techniques and classification methods compared are introduced. In Section 3, the procedure applied is presented, and in Section 4 the results of the comparative study are presented. Finally, Section 5 presents an illustration case, and Section 6 summarizes the conclusions.

2 MATERIALS AND METHODS

2.1 Data set

The data set used consists of a total of 16 images, collected from plastic compounds (PVC and PET) from a recycling company. These compounds were previously collected and tested in order to check that they were indeed the compounds under study, afterwards selecting regions of interest (ROI's) for creating the classification models. Two different types of PVC were analysed attending to their different spectra. The absorbance spectra have been represented in Figure 1.

[INSERT FIGURE 1 ABOUT HERE]

The equipment used in this work was composed by a XEVA-FPA-1.7-320 (XenICs, Belgium) matrix camera equipped with an InGaAs sensor with a resolution of 320 × 256 pixels, a pixel size of 30 μm and a special optics for the 50 mm Near Infra Red (NIR) [33-36]. This camera has an ImSpector N17E (Specim, Finland) coupled spectrograph which, by means of a prism, decomposes a line of 320 pixels wide in 256 lines corresponding to the individual wavelengths between 900 and 1700 nm (each one approximately 3.2 nm). This means that each image acquired with this camera is composed of 256 lines that correspond to the same line of the scene but in all the wavelengths. To obtain sample images in the laboratory, a Mirror Scanner (Specim, Finland) was used.

In order to have a visual reference of the analysed pieces, Figure 2 shows an example of RGB image of the samples analyzed.

[INSERT FIGURE 2 ABOUT HERE]

This way, the final images consist of a three-dimensional structure of 711 (refers to the pixels kept during the advance of the scan line camera) times 161 pixels (from pixel 80 to 240 of the total of 320, since only the central area of the image is analyzed due to the fact that the peripheral part is just background), and 256 wavelengths (admissible by the camera). In order to analyze the images by MIA, it is necessary to unfold them, considering the pixels as observations, and the wavelengths as variables. This is because we do not want to characterize the image in general, but each of the pixels in the image; thus following a MIA pixel-based approach [13, 14]. This way, a final **X** matrix is obtained, with 711x161 pixels (in rows) and 256 wavelengths (in columns) (Fig. 3).

[INSERT FIGURE 3 ABOUT HERE]

2.2 Methods

In this work, pixels of an image have to be classified into one of the following four classes: PVC, transparent PVC, PET and Background. In order to achieve it, different classification methods and pre-processing techniques were studied under a Full Factorial Design [25].

2.2.1 Pre-processing techniques

Pre-processing techniques are used to prepare the data set before the application of each model. The election of the appropriate pre-processing method chosen must be always carefully considered. In fact, this election is usually more relevant than the classification prediction analysis used [36]. For this reason, some usual methods in NIR spectroscopy like standard normal variate (SNV) [26], multiplicative scatter correction (MSC) [26] and Savitzky-Golay (SG) [27, 28] derivatives have been checked; as well as the option of directly using raw data (RD, after being transformed to absorbance units). In the case of MSC, the different pure spectra related to each pure chemical compound (or background) have been used as a reference, as well as the mean spectrum from the training set; in order to check for different performances. It must be noted, however,

that this last pre-processing approach has practical problems in this type of classification tasks, due to the fact that when performing on-line measurements and predictions/classifications, a continuous recalibration of the MSC should be carried out [37].

These raw data (absorbance raw data) were obtained in a two-step procedure. First step consists of transforming, for each wavelength λ, the intensities i_λ into reflectance values, r, using black (i_b) and white (i_w) references taken with the camera:

$$r_\lambda = 100 \times \frac{i_\lambda - i_{b\lambda}}{i_{w\lambda} - i_{b\lambda}} \quad (1)$$

The second step consists in obtaining the absorbance values, a, from the reflectance:

$$a_\lambda = log_{10} \times \frac{r_\lambda}{100} \quad (2)$$

From these absorbance values, a_λ, the different pre-processing techniques were applied, for comparison purposes.

2.2.2 **Classification models**

In this work, the following statistical and data mining classification methods: PLS-DA, nearest centroid, classification tree and principal component multinomial logistic regression have been compared.

2.2.2.1 **PLS-DA**

Partial Least Squares (PLS) [19-22] is a projection to latent structures model that explains the relationship between two sets of variables **X** and **Y**, as well as the variability in both **X** and **Y**, by maximizing the covariance between their internal latent structures.

PLS-DA [23] is the extension of PLS for classification purposes. The only difference between them is the type of response **Y**, which in this case is formed by as many binary variables (known as dummy variables) as classes to be separated. This way, the **Y** matrix is formed by as many columns as classes we have, and by many rows as pixels. If a pixel belongs to a particular class, it is assigned the value 1 for that class and 0 for

the rest of classes. Once these new variables have been defined, a test matrix (**X**) is projected by multiplying it by the estimated coefficients of the PLS model. Finally, by projecting any new testing image onto the PLS-DA model, its prediction with respect to each class is obtained; every pixel of this new image is assigned to the class for which such a prediction is the highest, if the residual sum of squares in the **X** space is below a predefined threshold.

2.2.2.2 **Classification tree**

A classification tree (CT) [31, 32] is a supervised learning tool consisting of a hierarchy of logical tests on some explanatory variables.

In the classification trees, training data from previously classified individuals are used and all possible binary cuts of each predictor (variables) are examined by constructing the complete tree at all levels. CT [31] begins by searching the data for the best splitter available, testing each predictor attribute value pair for its goodness of split [38, 39]. Which variable to use at each splitting node is determined by some measure of impurity, e.g. Gini index (used in this work), entropy or misclassification error [38].

Once the complete tree is built, the best level (tree pruning) is selected using an optimization criterion. In this work, the pruning criterion used was to determine the classification error for each level, which is obtained by cross validation. Afterwards, the level where the classification error was minimum was selected: within each level, the rate of correctly classified individuals in each class and the error are computed, stopping when the error increases. This means that if the classification error is higher at the next level than the previous level one, the pruning is carried out at the level where the classification error is smaller. It should be noted that another stop criterion was established, consisting on stopping the growth of the tree when the node is pure (i.e. when it only contains observations of one class).

2.2.2.3 **Principal Component Multinomial Logistic Regression**

Principal Component multinomial logistic regression (PCMLR) [29] consists of first applying the Principal Component Analysis (PCA) [9] technique, afterwards building a multinomial logistic regression [40] model on the PCA scores. It was decided to work with these PCA latent variables (scores) to avoid the ill-conditioning problems due to the high correlation between wavelengths, hence obtaining orthogonal and approximately normally distributed latent variables that fulfil the model assumptions.

Assuming K=4 classes (PVC, transparent PVC, PET and Background) of the variable **y**, there are π_k membership probabilities (one for each class) that satisfy:

$$\sum_k \pi_k = 1 \quad (6)$$

From this point, for some Type I risk α (0.05 in this case) the parameters of the regression model are estimated, selecting those whose p-value is lower than α, in a backward elimination procedure. Thus, the regression model is applied, using only the parameters that fulfil the aforementioned premise, obtaining the class probabilities from the following expressions. Finally, the class whose membership probability is the highest is assigned to each pixel.

$$\pi_1 = \frac{e^{\boldsymbol{t}'\boldsymbol{\beta}_1}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{t}'\boldsymbol{\beta}_k}} \quad (7)$$

$$\pi_2 = \frac{e^{\boldsymbol{t}'\boldsymbol{\beta}_2}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{t}'\boldsymbol{\beta}_k}} \quad (8)$$

$$\pi_3 = \frac{e^{\boldsymbol{t}'\boldsymbol{\beta}_3}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{t}'\boldsymbol{\beta}_k}} \quad (9)$$

$$\pi_4 = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{t}'\boldsymbol{\beta}_k}} \quad (10)$$

where 1 makes reference to PVC class, 2 refers to transparent PVC, class 3 refers to PET class and 4 refers to Background class. On the other hand, **t** is the scores vector and **β** is the regression coefficients vector.

2.2.2.4 *Nearest Centroid*

Nearest centroid (NC) method [30] is a nonparametric classification tool usually exploited for pattern recognition purposes. Unlabelled pixels are classified as belonging to the category whose distance (euclidean distance in this work) is minimum to the centroid of each class (class mean row vector). In contrast to the rest of methods, NC does not need training data.

2.2.3 *Figures of merit*

Once the different models were applied, the classification performance of each method was quantified, obtaining true positives (number of pixels of the image correctly identified as belonging to the category, TP), false positives (number of pixels of the image mistakenly identified as belonging to the category, FP), true negatives (number of pixels of the image correctly identified as not belonging to the category, TN) and false negatives (number of pixels of the image mistakenly identified as not belonging to the category, FN). Based on them, the figure-of-merit used in this work, the F-score, defined in terms of recall and precision, was calculated.

Recall is the ratio of a number of observations (pixels in this case) correctly classified (TP) in relation to a number of all correct pixels (TP+FN) (eq. 11). This measure, also known as true positive rate or sensitivity, provides information about classifier´s performance with respect to false negatives.

$$Recall_{n,z} = \frac{TP_{n,z}}{TP_{n,z} + FN_{n,z}} \quad (11)$$

Precision is the ratio of the number of observations (pixels in this case) correctly classified (TP) with respect to all pixels classified as positive (TP+FP) (eq. 12). This index gives information about its performance with respect to false positives.

$$Precision_{n,z} = \frac{TP_{n,z}}{TP_{n,z} + FP_{n,z}} \quad (12)$$

From recall and precision, F-score is computed as indicated in eq. 13:

$$F - score_{n,z} = 2 \times \frac{precision_{n,z} \times recall_{n,z}}{precision_{n,z} + recall_{n,z}} \quad (13)$$

$$\forall n = 1,2,\ldots,4 \qquad \forall z = 1,2,\ldots,7$$

where *n* corresponds to the model applied (PLS-DA=1, CT=2, PCMLR=3, NC=4) and *z* corresponds to the pre-processing type (RD=1, SNV=2, SG=3, MSC(PVC as reference)=4, MSC(transparent PVC as reference)=5, MSC(PET as reference)=6, MSC(Background as reference)=7, MSC(Mean spectrum)=8). Note that F-score is maximum when all pixels are correctly classified (no false positives nor false negatives).

3 PROCEDURE

The procedure carried out is the following:

1) Select pixels of each of the four classes (PVC, transparent PVC, PET and Background) and build data matrix **X.**
2) Data matrix **X** is divided into 10 clusters. Each cluster has roughly equal size and roughly the same class proportions, in order to avoid any bias to any of the classes. This way, applying a leave-one-block-out iteration procedure, we end up with 10 different training and validation data sets, using at each iteration 90% of the data for training and 10% for validation.
3) For each iteration:
    a. Apply the different types of pre-processing: Raw, SNV, SG and MSC.
    b. Build the model with the training data set.
    c. Obtain the different performance measures of each classification models (TP, TN, FP, FN, precision, recall, F-score) for each class, with the validation set.
4) Apply Analysis of Variance (ANOVA) [19] with the aim of assessing for possible statistical significant differences with respect to the mean of the F-scores, taking as factors the type of pre-processing, the classified model used, the class of chemical compound and the cluster of the cross-validation round (used as a blocking factor).

4 RESULTS

Results provided by each of the treatments of the complete factorial design are presented in Tables 1 to 4, for each of the four classes analysed: PVC, transparent PVC, PET and Background.

Table 1. Global TP, TN, FP, FN, precision, recall, F-score for PVC class.

Method	TP	TN	FP	FN	PRECISION	RECALL	F-SCORE
PLS-DA with raw data	1000	3000	0	0	1,0000	1,0000	1,0000
PLS-DA with SNV	999	3000	0	1	1,0000	0,9990	0,9995
PLS-DA with S-G	961	2995	5	39	0,9948	0,9610	0,9776
PLS-DA with MSC (ref PVC)	990	2356	644	10	0,6059	0,9900	0,7517
PLS-DA with MSC (ref PVCtrans)	820	1015	1985	180	0,2923	0,8200	0,4310
PLS-DA with MSC (ref PET)	833	1321	1679	167	0,3316	0,8330	0,4744
PLS-DA with MSC (ref background)	896	1124	1876	104	0,3232	0,8960	0,4751
PLS-DA with MSC (ref mean spectrum)	917	1057	1943	83	0,3206	0,9170	0,4751
Class tree with raw data	997	2995	5	3	0,9950	0,9970	0,9960
Class tree with SNV	1000	3000	0	0	1,0000	1,0000	1,0000
Class tree with S-G	922	2939	61	78	0,9379	0,9220	0,9299
Class tree with MSC (ref PVC)	1000	3000	0	0	1,0000	1,0000	1,0000
Class tree with MSC (ref PVCtrans)	999	3000	0	1	1,0000	0,9990	0,9995
Class tree with MSC (ref PET)	999	2999	1	1	0,9990	0,9990	0,9990
Class tree with MSC (ref background)	1000	3000	0	0	1,0000	1,0000	1,0000
Class tree with MSC (ref mean spectrum)	999	3000	0	1	1,0000	0,9990	0,9995
PCMLR with raw data	480	2537	463	520	0,5090	0,4800	0,4941
PCMLR with SNV	527	2655	345	473	0,6044	0,5270	0,5630
PCMLR with S-G	496	2273	727	504	0,4056	0,4960	0,4462
PCMLR with MSC (ref PVC)	983	2654	346	17	0,7397	0,9830	0,8441
PCMLR with MSC (ref PVCtrans)	370	1764	1236	630	0,2304	0,3700	0,2840
PCMLR with MSC (ref PET)	73	2675	325	927	0,1834	0,0730	0,1044
PCMLR with MSC (ref background)	95	2999	1	905	0,9896	0,0950	0,1734
PCMLR with MSC (ref mean spectrum)	302	2289	711	698	0,2981	0,3020	0,3000
Nearest Centroid with raw data	722	2672	328	278	0,6876	0,7220	0,7044
Nearest Centroid with SNV	901	3000	0	99	1,0000	0,9010	0,9479
Nearest Centroid with S-G	652	2475	525	348	0,5540	0,6520	0,5990
Nearest Centroid with MSC (ref PVC)	1000	1513	1487	0	0,4021	1,0000	0,5736
Nearest Centroid with MSC (ref PVCtrans)	188	3000	0	812	1,0000	0,1880	0,3165
Nearest Centroid with MSC (ref PET)	14	3000	0	986	1,0000	0,0140	0,0276
Nearest Centroid with MSC (ref background)	190	3000	0	810	1,0000	0,1900	0,3193
Nearest Centroid with MSC (ref mean spectrum)	588	3000	0	412	1,0000	0,5880	0,7406

Table 2. Global TP, TN, FP, FN, precision, recall, F-score for Transparent PVC class.

Method	TP	TN	FP	FN	PRECISION	RECALL	F-SCORE
PLS-DA with raw data	999	2997	3	1	0,9970	0,9990	0,9980
PLS-DA with SNV	977	3000	0	23	1,0000	0,9770	0,9884
PLS-DA with S-G	961	2996	4	39	0,9959	0,9610	0,9781
PLS-DA with MSC (ref PVC)	999	2914	86	1	0,9207	0,9990	0,9583
PLS-DA with MSC (ref PVCtrans)	0	3000	0	1000	0,0000	0,0000	0,0000
PLS-DA with MSC (ref PET)	93	2988	12	907	0,8857	0,0930	0,1683
PLS-DA with MSC (ref background)	29	2990	10	971	0,7436	0,0290	0,0558
PLS-DA with MSC (ref mean spectrum)	8	3000	0	992	1,0000	0,0080	0,0159
Class tree with raw data	992	2992	8	8	0,9920	0,9920	0,9920
Class tree with SNV	996	2995	5	4	0,9950	0,9960	0,9955
Class tree with S-G	909	2935	65	91	0,9333	0,9090	0,9210
Class tree with MSC (ref PVC)	990	2993	7	10	0,9930	0,9900	0,9915
Class tree with MSC (ref PVCtrans)	995	2996	4	5	0,9960	0,9950	0,9955
Class tree with MSC (ref PET)	997	2993	7	3	0,9930	0,9970	0,9950
Class tree with MSC (ref background)	995	2996	4	5	0,9960	0,9950	0,9955
Class tree with MSC (ref mean spectrum)	994	2996	4	6	0,9960	0,9940	0,9950
PCMLR with raw data	574	2668	332	426	0,6336	0,5740	0,6023
PCMLR with SNV	337	2206	794	663	0,2980	0,3370	0,3163
PCMLR with S-G	545	2504	496	455	0,5235	0,5450	0,5341
PCMLR with MSC (ref PVC)	100	1465	1535	900	0,0612	0,1000	0,0759
PCMLR with MSC (ref PVCtrans)	750	1476	1524	250	0,3298	0,7500	0,4582
PCMLR with MSC (ref PET)	333	1156	1844	667	0,1530	0,3330	0,2096
PCMLR with MSC (ref background)	10	2344	656	990	0,0150	0,0100	0,0120
PCMLR with MSC (ref mean spectrum)	200	1359	1641	800	0,1086	0,2000	0,1408
Nearest Centroid with raw data	1000	2667	333	0	0,7502	1,0000	0,8573
Nearest Centroid with SNV	948	2987	13	52	0,9865	0,9480	0,9669
Nearest Centroid with S-G	675	2688	312	325	0,6839	0,6750	0,6794
Nearest Centroid with MSC (ref PVC)	907	2394	606	93	0,5995	0,9070	0,7218
Nearest Centroid with MSC (ref PVCtrans)	1000	953	2047	0	0,3282	1,0000	0,4942
Nearest Centroid with MSC (ref PET)	16	2673	327	984	0,0466	0,0160	0,0238
Nearest Centroid with MSC (ref background)	0	2999	1	1000	0,0000	0,0000	0,0000
Nearest Centroid with MSC (ref mean spectrum)	927	2973	27	73	0,9717	0,9270	0,9488

Table 3. Global TP, TN, FP, FN, precision, recall, F-score for PET class.

Method	TP	TN	FP	FN	PRECISION	RECALL	F-SCORE
PLS-DA with raw data	996	2994	6	4	0,9940	0,9960	0,9950
PLS-DA with SNV	999	2991	9	1	0,9911	0,9990	0,9950
PLS-DA with S-G	984	2993	7	16	0,9929	0,9840	0,9884
PLS-DA with MSC (ref PVC)	339	2520	480	661	0,4139	0,3390	0,3727
PLS-DA with MSC (ref PVCtrans)	1000	2861	139	0	0,8780	1,0000	0,9350
PLS-DA with MSC (ref PET)	999	2700	300	1	0,7691	0,9990	0,8691
PLS-DA with MSC (ref background)	999	2814	186	1	0,8430	0,9990	0,9144
PLS-DA with MSC (ref mean spectrum)	1000	2851	149	0	0,8703	1,0000	0,9307
Class tree with raw data	998	2998	2	2	0,9980	0,9980	0,9980
Class tree with SNV	999	2996	4	1	0,9960	0,9990	0,9975
Class tree with S-G	983	2982	18	17	0,9820	0,9830	0,9825
Class tree with MSC (ref PVC)	998	2992	8	2	0,9920	0,9980	0,9950
Class tree with MSC (ref PVCtrans)	1000	2994	6	0	0,9940	1,0000	0,9970
Class tree with MSC (ref PET)	996	2996	4	4	0,9960	0,9960	0,9960
Class tree with MSC (ref background)	997	2996	4	3	0,9960	0,9970	0,9965
Class tree with MSC (ref mean spectrum)	997	2998	2	3	0,9980	0,9970	0,9975
PCMLR with raw data	265	2116	884	735	0,2306	0,2650	0,2466
PCMLR with SNV	570	2524	476	430	0,5449	0,5700	0,5572
PCMLR with S-G	471	2395	605	529	0,4377	0,4710	0,4538
PCMLR with MSC (ref PVC)	107	2976	24	893	0,8168	0,1070	0,1892
PCMLR with MSC (ref PVCtrans)	0	2894	106	1000	0,0000	0,0000	0,0000
PCMLR with MSC (ref PET)	2	2442	558	998	0,0036	0,0020	0,0026
PCMLR with MSC (ref background)	0	923	2077	1000	0,0000	0,0000	0,0000
PCMLR with MSC (ref mean spectrum)	209	2850	150	791	0,5822	0,2090	0,3076
Nearest Centroid with raw data	761	2896	104	239	0,8798	0,7610	0,8161
Nearest Centroid with SNV	952	2999	1	48	0,9990	0,9520	0,9749
Nearest Centroid with S-G	757	2986	14	243	0,9818	0,7570	0,8549
Nearest Centroid with MSC (ref PVC)	0	3000	0	1000	0,0000	0,0000	0,0000
Nearest Centroid with MSC (ref PVCtrans)	327	2628	372	673	0,4678	0,3270	0,3849
Nearest Centroid with MSC (ref PET)	1000	2346	654	0	0,6046	1,0000	0,7536
Nearest Centroid with MSC (ref background)	443	2759	241	557	0,6477	0,4430	0,5261
Nearest Centroid with MSC (ref mean spectrum)	606	2955	45	394	0,9309	0,6060	0,7341

Table 4. Global TP, TN, FP, FN, precision, recall, F-score for Background class.

Method	TP	TN	FP	FN	PRECISION	RECALL	F-SCORE
PLS-DA with raw data	990	2995	5	10	0,9950	0,9900	0,9925
PLS-DA with SNV	990	2975	25	10	0,9754	0,9900	0,9826
PLS-DA with S-G	988	2909	91	12	0,9157	0,9880	0,9505
PLS-DA with MSC (ref PVC)	324	2862	138	676	0,7013	0,3240	0,4432
PLS-DA with MSC (ref PVCtrans)	3	2947	53	997	0,0536	0,0030	0,0057
PLS-DA with MSC (ref PET)	11	2927	73	989	0,1310	0,0110	0,0203
PLS-DA with MSC (ref background)	0	2996	4	1000	0,0000	0,0000	0,0000
PLS-DA with MSC (ref mean spectrum)	31	2696	304	969	0,0925	0,0310	0,0464
Class tree with raw data	987	2989	11	13	0,9890	0,9870	0,9880
Class tree with SNV	991	2995	5	9	0,9950	0,9910	0,9930
Class tree with S-G	898	2856	144	102	0,8618	0,8980	0,8795
Class tree with MSC (ref PVC)	986	2989	11	14	0,9890	0,9860	0,9875
Class tree with MSC (ref PVCtrans)	991	2995	5	9	0,9950	0,9910	0,9930
Class tree with MSC (ref PET)	989	2993	7	11	0,9930	0,9890	0,9910
Class tree with MSC (ref background)	990	2994	6	10	0,9940	0,9900	0,9920
Class tree with MSC (ref mean spectrum)	993	2996	4	7	0,9960	0,9930	0,9945
PCMLR with raw data	423	2421	579	577	0,4222	0,4230	0,4226
PCMLR with SNV	262	2311	689	738	0,2755	0,2620	0,2686
PCMLR with S-G	298	2638	362	702	0,4515	0,2980	0,3590
PCMLR with MSC (ref PVC)	5	2100	900	995	0,0055	0,0050	0,0052
PCMLR with MSC (ref PVCtrans)	0	2986	14	1000	0,0000	0,0000	0,0000
PCMLR with MSC (ref PET)	0	2135	865	1000	0,0000	0,0000	0,0000
PCMLR with MSC (ref background)	68	1907	1093	932	0,0586	0,0680	0,0629
PCMLR with MSC (ref mean spectrum)	81	2054	946	919	0,0789	0,0810	0,0799
Nearest Centroid with raw data	339	2587	413	661	0,4508	0,3390	0,3870
Nearest Centroid with SNV	999	2814	186	1	0,8430	0,9990	0,9144
Nearest Centroid with S-G	565	2499	501	435	0,5300	0,5650	0,5470
Nearest Centroid with MSC (ref PVC)	0	3000	0	1000	0,0000	0,0000	0,0000
Nearest Centroid with MSC (ref PVCtrans)	0	2934	66	1000	0,0000	0,0000	0,0000
Nearest Centroid with MSC (ref PET)	992	2003	997	8	0,4987	0,9920	0,6638
Nearest Centroid with MSC (ref background)	1000	875	2125	0	0,3200	1,0000	0,4848
Nearest Centroid with MSC (ref mean spectrum)	1000	2200	800	0	0,5556	1,0000	0,7143

These results were analyzed by means of ANalysis Of VAriance (ANOVA) in order to determine which model and pre-processing techniques were able to make a better classification of the plastic compounds analyzed, in terms of F-score.

Furthermore, a correspondence analysis (CA) [41] was performed on the contingency tables derived from the previous results for each class, in terms of true positives, true negatives, false positives and false negatives; for the different methodologies applied. CA is conceptually similar to PCA but it is proposed for categorical data processing [30].

Table 5 shows the ANOVA results. Out of the statistically significant factors (p-value<0.05), the relevant ones are the type of pre-processing technique as well as the model used. It should be noted that cluster was used as a blocking factor and class was used to select the best model and type of pre-processing for classifying of each chemical compound (PVC, transparent PVC, PET and Background). The least significant difference (LSD) intervals are presented in Fig. 4, a), b), c), and d), showing up the best models for each chemical compound (as well as background) and pre-processing technique.

For PVC, CT for all pre-processings (MSC regardless of the reference used, RD, SG and SNV) presented the best and equivalent results, non-statistically different from PLS-DA for RD, SG and SNV. NC also showed equivalent F-scores (from a statistically point of view) but only for MSC1 and SNV. For transparent PVC, again CT provided the best and equivalent results, to those from PLS-DA for MSC1, RD, SG and SNV. In this case, NC raised equivalent performances for MSC2-5 and SNV.

For PET, CT was statistically equivalent to PLS-DA in all pre-processings but for MSC1. NC showed non-statistically different results for MSC3, SG and SNV. Finally, for the background, CT provided the best and equivalent results regardless of the pre-processing technique applied, only equalled by PLS-DA when using RD, SG or SNV pre-processings.

So, in general, the best models were the Classification Trees (regardless of type of pre-processing) and PLS-DA (for RD, SNV and SG pre-processing), whereas the worst

model was PCMLR. NC showed more variability in general, being equivalent depending on the type of chemical compound and pre-processing technique applied. It should be pointed out that MSC pre-processing was excluded from subsequent CA analyses because it provided a high variability in the ANOVA results, depending on the reference spectrum taken into account, which hampered choosing the best model.

Figure 5, a) to d) shows the correspondence analysis (CA) results for the different classes. The best methods for the segmentation of each of the classes, i.e. those located in the TP/TN quadrant were the same as those obtained in ANOVA. For PVC, CT and PLS-DA provide the best models, regardless of the pre-processing applied; joint to NC-SNV. For transparent PVC, exactly the same conclusions could be extracted. However, if one was specially interested in maximizing the TN rate, NC with SG or RD should be selected. Finally, for PET and for Background, as well as for PVC and transparent PVC, again CT and PLS-DA provide the best models, regardless of the pre-processing applied; joint to NC-SNV.

So again, in general, CT (no matter the pre-processing method) and PLS-DA (for RD, SNV and SG pre-processing) were the best options. The benefit of CA with respect to ANOVA, however, is that CA allows choosing each of them attending to the prior relevance given to each of the TP, TN, FP or FN parameters (as shown in the case of NC for maximizing the TN rate).

Table 5: Analysis of Variance.

Source	Sum Sq.	d.f.	Mean Sq.	F	p-value
model	71.483	3	23.8277	600.11	0
preprocessing	14.156	7	2.0222	50.93	0
class	3.684	3	1.2281	30.93	0
cluster	0.952	9	0.1057	2.66	0.0047
model*preprocessing	19.633	21	0.9349	23.55	0
model*class	8.443	9	0.9382	23.63	0
model*cluster	2.663	27	0.0986	2.48	0.0001
preprocessing*class	6.066	21	0.2889	7.28	0
preprocessing*cluster	2.359	63	0.0374	0.94	0.6032
class*cluster	0.377	27	0.014	0.35	0.9992
Error	43.24	1089	0.0397		
Total	173.057	1279			

[INSERT FIGURE 4 ABOUT HERE]

[INSERT FIGURE 5 ABOUT HERE]

5 ILLUSTRATION CASE

Finally, some ROI's of images of plastic compounds (PVC and PET) extracted from Fig. 2 were projected for the best methods chosen in Section 4 (CT and PLS-DA). Results are shown in Figure 6. Pixel assignation was carried out in the following way:

- 1 if the pixel was classified as PVC (blue color).
- 2 if the pixel was classified as transparent PVC (light blue color).
- 3 if the pixel was classified as PET (yellow color).
- 4 if the pixel was classified as Background (brown color).

[INSERT FIGURE 6 ABOUT HERE]

It should be noted that each pixel could only be assigned to one of the mentioned classes. Results are quite good. Moreover, despite of the non-statistically significant

differences between CT and PLS-DA (when discarding MSC), it seems that PLS-DA is less noisy than CT.

6 CONCLUSIONS

This work provides a methodology, based on a DoE framework, for choosing the best classification technique/s and pre-processing methodologies for impurities detection in PET recycling.

In the particular problem treated in this paper, the application of the proposed DoE-based methodology allows concluding that for the two best classification models (PLS-DA and CT) out of the four compared, raw data (RD) provides comparable results to those provided by SNV and SG pre-processing, in terms of statistical significance. For this reason, being it the fastest pre-processing method (since no correction of the spectrum is necessary), it could be selected as the most appropriate in this case. This is also in accordance to previous works with hyperspectral imagery [42]. MSC was discarded since it provided very different results depending on the reference spectrum used. Furthermore, CA allowed to specifically recognizing in a very simple and graphical way those strategies providing the higher average rates of TP, FP, FN, and TN; and even selecting the model and pre-processing technique attending to a special focus on TN or TP.

This way, multivariate image analysis (MIA), regardless of the final classification model used, provides real and feasible solutions to a possible automation of the PET recycling process.

7 AKNOWLEDGEMENTS

This research was partially supported by the Universitat Politècnica de València under the project UPV-FE-16-B18.

8 REFERENCES

[1] PlasticsEurope, 2016. Plastics – the Facts 2016 An analysis of European plastics production, demand and waste data.

[2] Ulrici A, Serranti S, Ferrari C, Cesare D, Foca G, Bonifazi G. Efficient chemometric strategies for PET–PLA discrimination in recycling plants using hyperspectral imaging. *Chemometr Intell Lab.* 2013; 122: 31–39.

[3] Moroni M, Mei A, Leonardi A, Lupo E, La M. PET and PVC separation with hyperspectral imagery. *Sens Switz*. 2015; 15(1), 2205-2227. doi:10.3390/s150102205.

[4] Amigo, J.M., Babamoradi, H., Elcoroaristizabal, S., Hyperspectral image analysis. A tutorial, Anal. Chim. Acta, 2015 (896), 34-51;

[5] Hugelier, S., Vitale, R., Ruckebusch, C., Edge-preserving image smoothing constraint in MCR-ALS of hyperspectral data, Appl. Spectrosc., In Press, doi:10.1177/0003702817735598.

[6] Serranti S, Gargiulo A, Bonifazi G. The Utilization of Hyperspectral Imaging for Impurities Detection in Secondary Plastics. *The Open Waste Management Journal*. 2010; 3: 56-70.

[7] Can Karaca A, Ertürk A, Kemal Güllü M, Elmas M, Ertürk S. Automatic waste sorting using shortwave infrared hyperspectral imaging system. *Proceedings of the Signal Processing and Communications Applications Conference (SIU)*, 2013, Haspolat, Turkey. doi: 10.1109/SIU.2013.6531170.

[8] Hollstein F, Wohllebe M, Arnaiz S. Identification and Sorting of Plastics Film Waste by NIR-Hyperspectral-Imaging. *Proceedings of the ICNIRS* 2015. doi:10.17648/nir-2015-34127.

[9] Bonifazi G, Di M, Potenza F, Serranti S. FT-IR Analysis and Hyperspectral Imaging Applied to Postconsumer Plastics Packaging Characterization and Sorting. *IEEE Sens J*. 2016; 16 (10): 3428-3434.

[10] Luciani V, Bonifazi G, Rem P, Serranti S. Upgrading of PVC rich wastes by magnetic density separation and hyperspectral imaging quality control. *Waste Manage.* 2015; 45 :118-125.

[11] Hollstein F, Wohllebe M, Arnaiz S,Manjon D, O'Brien N, Kulcke A. Challenges in Automatic Sorting of Bio-Plastics within the Recycling Chain by Means of NIR-Hyperspectral-Imaging. NIR2013 *Proceedings, 2-7 June,* La Grande-Motte, France.

[12] Cesseti M, Nicolosi P. Waste processing: new near infrared technologies for material identification and selection. *J Instrum. 2016;*. 11: C09002.

[13] Prats-Montalbán JM, De Juan A and Ferrer A. Multivariate image analysis: A review with applications. *Chemometr Intell Lab.* 2011; 107: 1-23.

[14] Prats-Montalbán JM. *Control Estadístico de Procesos mediante Análisis Multivariante de Imágenes. (PhD Thesis).* Universidad Politécnica de Valencia. Valencia. 2005

[15] Bro, R, Smilde, A.K.. Principal component analysis. *Anal Methods*. 2014; 6: 2812-2831.

[16] Jackson J E. (1991) *A User's guide to Principal Components.* Ed. Wiley, New York.

[17] De Juan A, Tauler R. Multivariate curve resolution (MCR) from 2000: Progress in concepts and applications. *Crit Rev Anal Chem.* 2006; 36 (3-4): 163-176.

[18] Tauler R. Multivariate curve resolution applied to second order data. *Chemometr Intell Lab.* 1995; 30: 133-146.

[19] Geladi P, Kowalski BR. Partial Least-Squares Regression: A Tutorial. *Anal Chim Acta.* 1986; 185: 1-17.

[20] Gurden SB, Westerhuis JA, Bro R, Smilde AK. A comparison of multiway regression and scaling methods, *Chemometr Intell Lab.*2001; 59 (1-2): 121-136.

[21] Höskuldsson A. PLS Regression Methods, *J Chemometr.* 1998; 2: 221-228.

[22] Kresta JV, Marlin TE and MacGregor JF. Development of Inferential Process Models using PLS, *Computers chem Engng*. 1994; 18 (7): 597-611.

[23] Sjöström M, Wold S, Söderström B. *PLS Discriminant Plots, Proceedings of PARC in Practice, Amsterdam, June 19-21, 1985*. Elsevier Science Publishers B.V., North-Holland. 1986

[24] Wold S, Albano C, Dunn WJ, Edlund U, Esbensen K, Geladi P, Hellberg S, Johansson E, Lindberg W, Sjöström M. Multivariate Data Analysis in Chemistry, In: B.R. Kowalski (ed.) *Chemometrics: Mathematics and Statistics in Chemistry*, D. Reidel Publishing Company, Dordrecht, Holland. 1984

[25] Montgomery DC. *Design and Analysis of Experiments, 6th edition*. Wiley & Sons, New York. 2005.

[26] Fearn T, Riccioli C, Garrido A and Guerrero JE. On the geometry of SNV and MSC. *Chemometr Intell Lab.* 2008; 96: 22-26.

[27] Smilde AK, Tauler R, Saurina J, et al. Calibration methods for complex second-order data. *Anal Chim Acta.* 1999; 398 (2-3): 237-251.

[28] Savitzky A and Golay MJE. Smoothing and Differentiation of Data by Simplified Least-Squares Procedures. *Anal Chem*. 1964; 36: 1627-1639.

[29] Lucadamoa A and Leoneb A. Principal component multinomial regression and spectrometry to predict soil texture. *J Chemometr.* 2015; 29: 514–520.

[30] Vitale R, Prats-Montalbán JM, López-García F, Blasco J, and Ferrer A. Segmentation techniques in image analysis: A comparative study. *J Chemometr.* 2016;

30: 749-758.

[31] Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Wadsworth, Belmont, CA. 1984.

[32] Torgo, L. *Data Mining and learning with case studies*. Chapman and Hall/CRC. 2010

[33] Prats-Montalbán JM, Jerez J, Romañach R and Ferrer A. MIA and NIR Chemical Imaging for pharmaceutical product characterizacion. *Chemometr Intell Lab.* 2012; 117: 240-249.

[34] Gómez J, Lorente D, Soria E, Aleixos N, Cubero S and Blasco J. Development of a Hyperspectral Computer Vision System Based on Two Liquid Crystal Tuneable Filters for Fruit Inspection. Application to Detect Citrus Fruits Decay, *Food Bioprocess Tech*. 2013; 7(4):1047-1056.

[35] Silva C, Pimentel MF, Honorato R, Pasquini C, Prats-Montalbán J.M and Ferrer, A. Near infrared hyperspectral imaging for forensic analysis of document forgery. *Analyst.* 2014; 139: 5176-5184.

[36] Vidal M, Amigo JM. Pre-processing of hyperspectral images. Essential steps before image analysis. *Chemometr Intell Lab.* 2012; 117: 138-148.

[37] Maleki MR, Mouazen AM, Ramon H, De Baerdemaeker J. Multiplicative Scatter Correction during On-line Measurement with Near Infrared Spectroscopy. *Biosyst Eng.* 2007; 96 (3): 427–433.

[38] Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning : Data mining, inference, and prediction.* New York: Springer Verlag. 2001.

[39] Rokach L, Maimon O. Data mining with decision trees: theory and applications. *World Scientific Pub Co Inc*. 2008.

[40] Peña D. *Análisis de Datos Multivariantes*, Madrid, España, McGraw Hill. 2002

[41] Hirschfeld HO. A connection between correlation and contingency. *Math Proc Cambridge.* 1935; 31: 520–524.

[42] Folch-Fortuny A, Prats-Montalbán JM, Cubero S, Blasco J, Ferrer A. Hyperspectral imaging and N-way PLS-DA models for fungus detection in oranges. *Chemometr Intell Lab.* 2016; 156: 241–248.

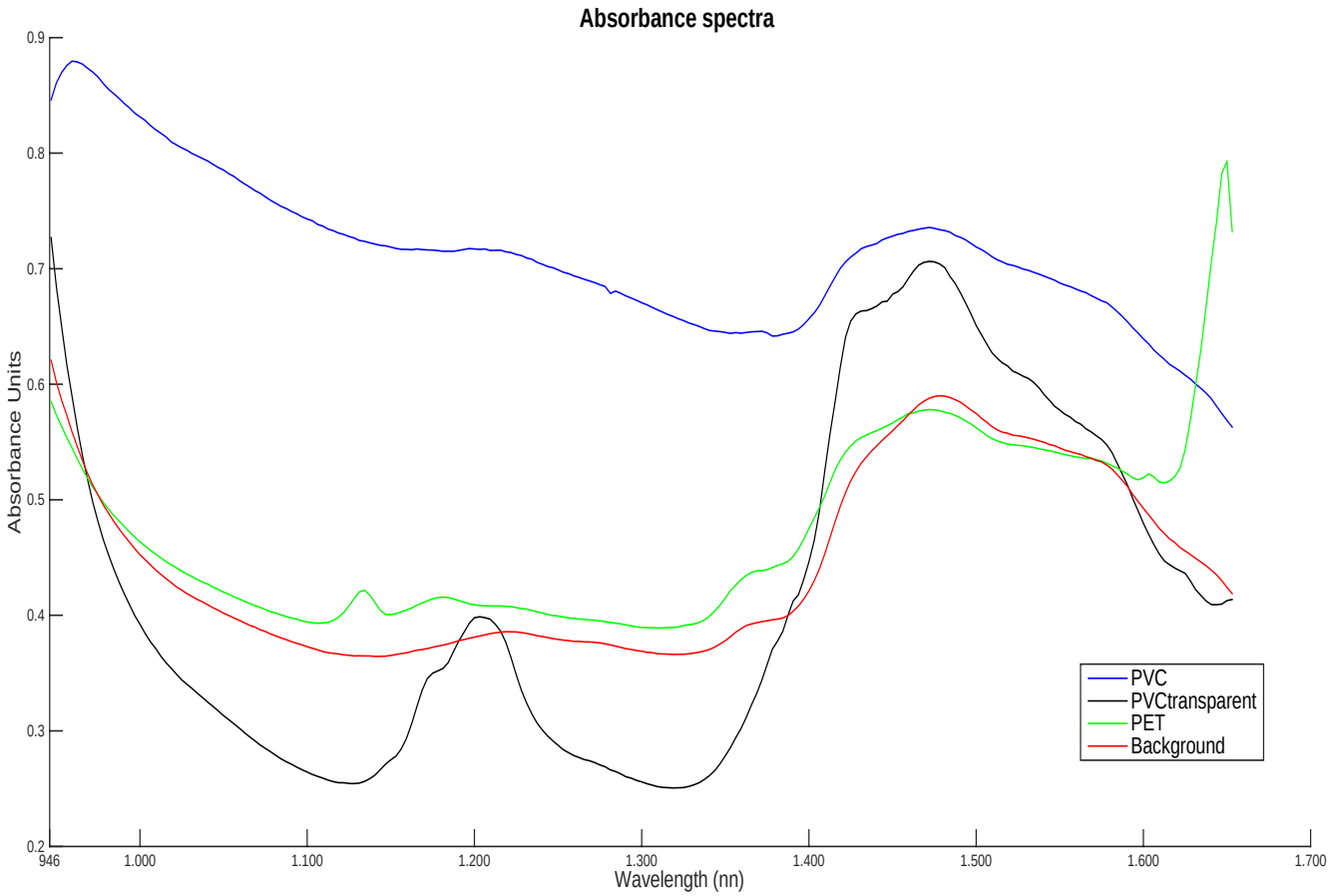


Figure 1. Pure Spectrum of each analyzed class

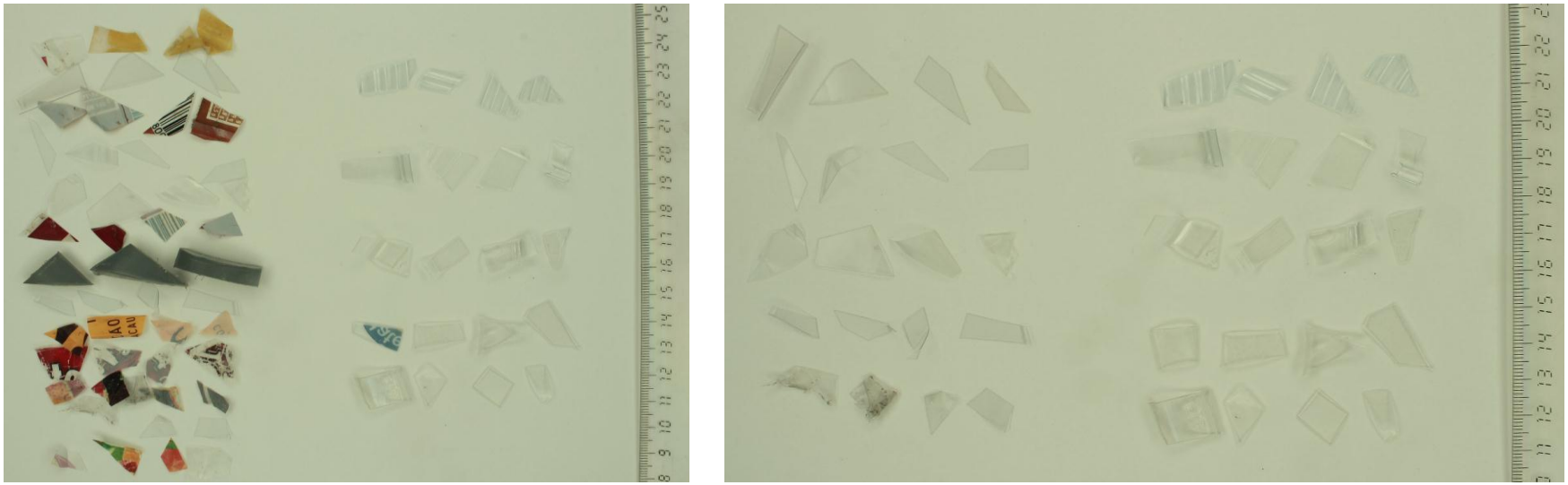


Figure 2. RGB images from PVC and PET to observe the difference between them. Left: PVC and transparent PVC flakes. Right: PET flakes.

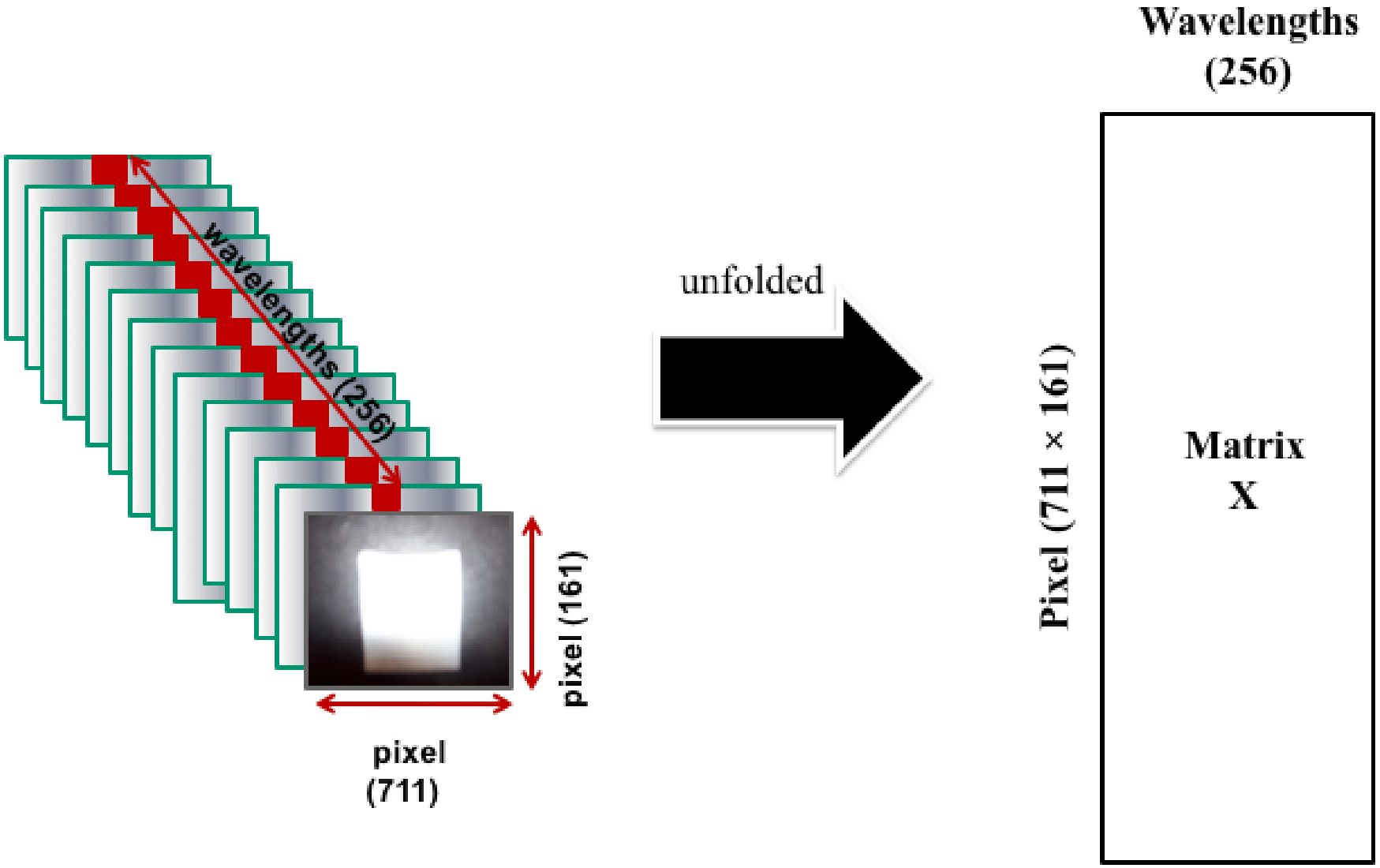


Figure 3. Unfolded Image. This figure displays the transformation from 3-D vector to matrix (2-D vector). Each image obtained at a certain wavelength will be arranged as a column vector in the new matrix.

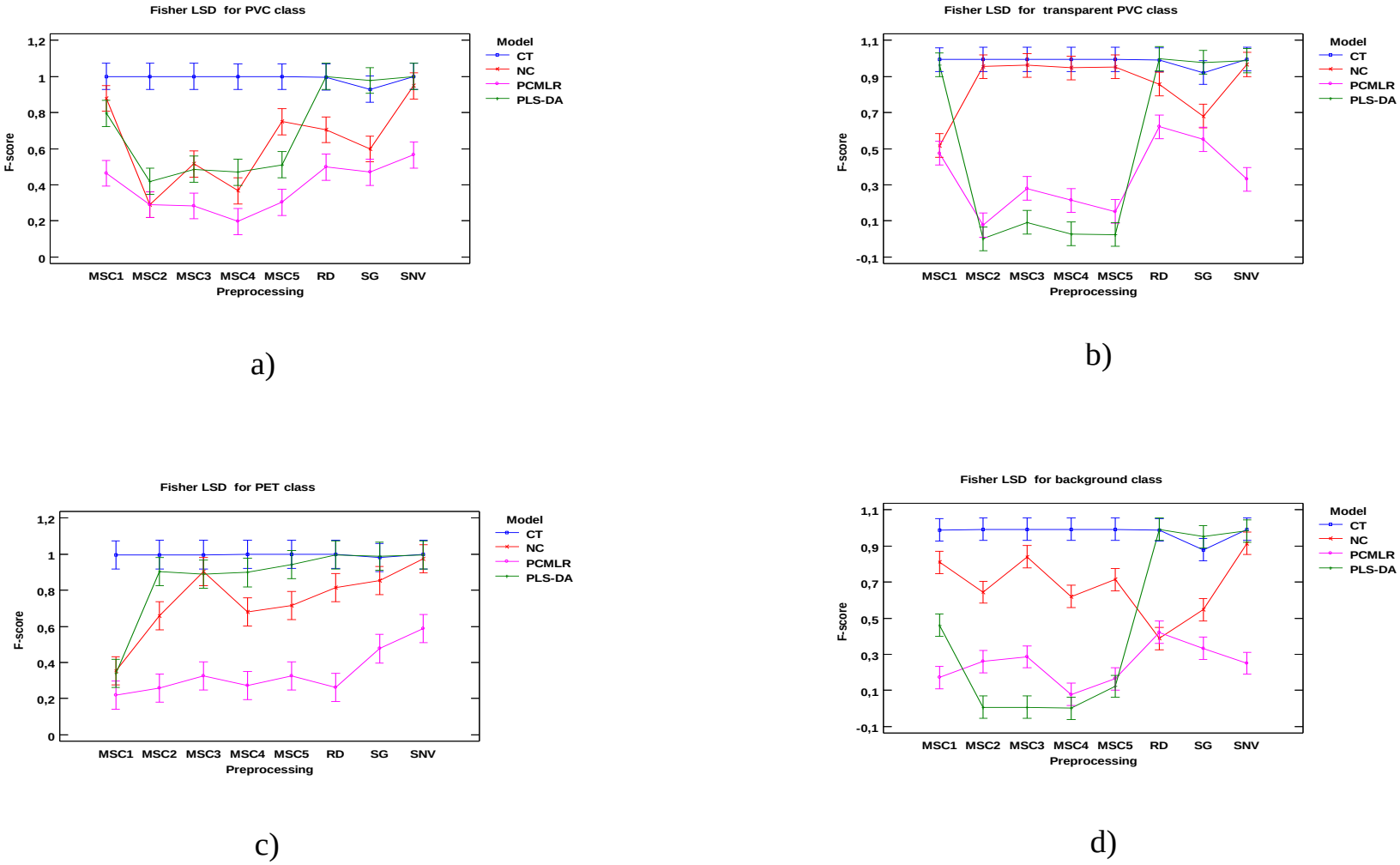
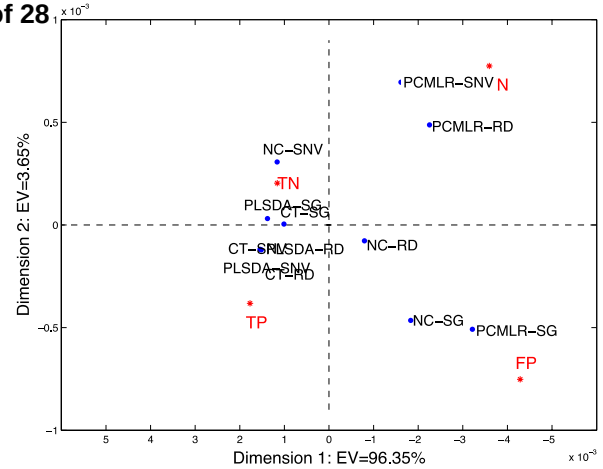
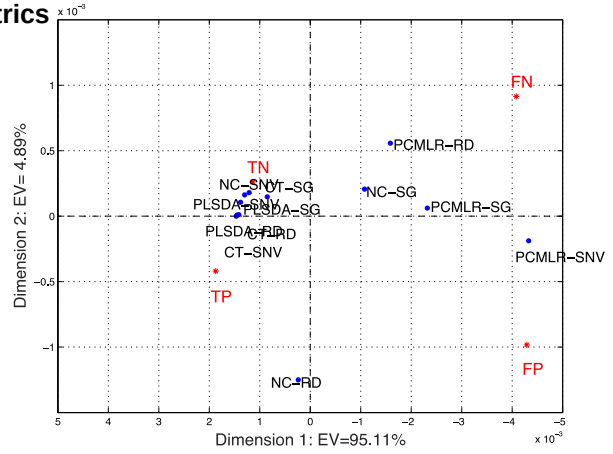


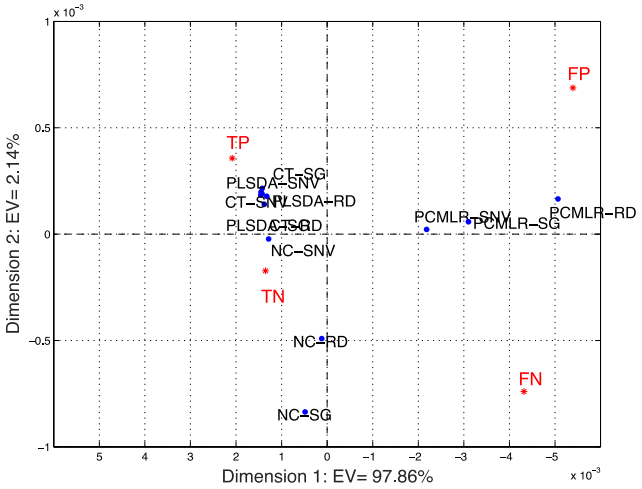
Figure 4: 95% least significant difference intervals from ANOVA on F-score values for: a) PVC, b) transparent PVC, c) PET and d) Background classes. This interaction plot makes reference to model and preprocessing data. The methods were Partial Least Squares-Discriminant Analysis (PLS-DA), Classification trees (CT), Principal Component Multinomial Logistic Regression (PCMLR) and Nearest Centroid (NC). The preprocessing techniques were Raw Data (RD), Standard Normal Variate (SNV), Savitzky-Golay (SG) and Multiplicative Scatter Correction: MSC1 (PVC), MSC2 (Transparent PVC), MSC3 (PET), MSC4 (Background), MSC5(Mean spectrum).



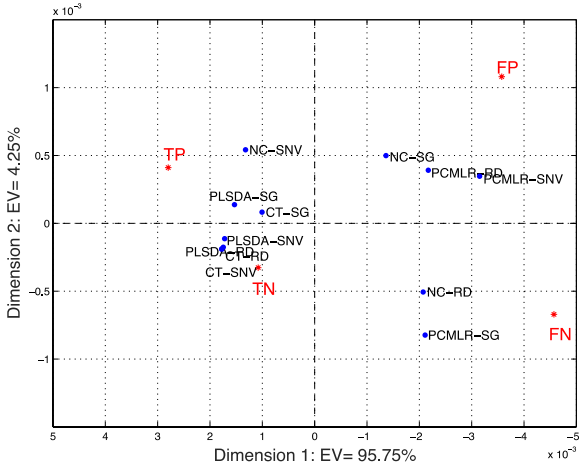
a)



b)



c)



d)

Figure 5: Correspondence analysis for a) PVC, b) transparent PVC, c) PET and d) Background classes. The methods were Partial Least Squares-Discriminant Analysis (PLSDA), Classification trees (CT), Principal Component Multinomial Logistic Regression (PCMLR) and Nearest Centroid (NC). The preprocessing techniques were Raw Data (RD), Standard Normal Variate (SNV) and Savitzky-Golay (SG). EV refers to Explained Variance, TP to true positives, FP to false positives, TN to true negatives and FN to false negatives.

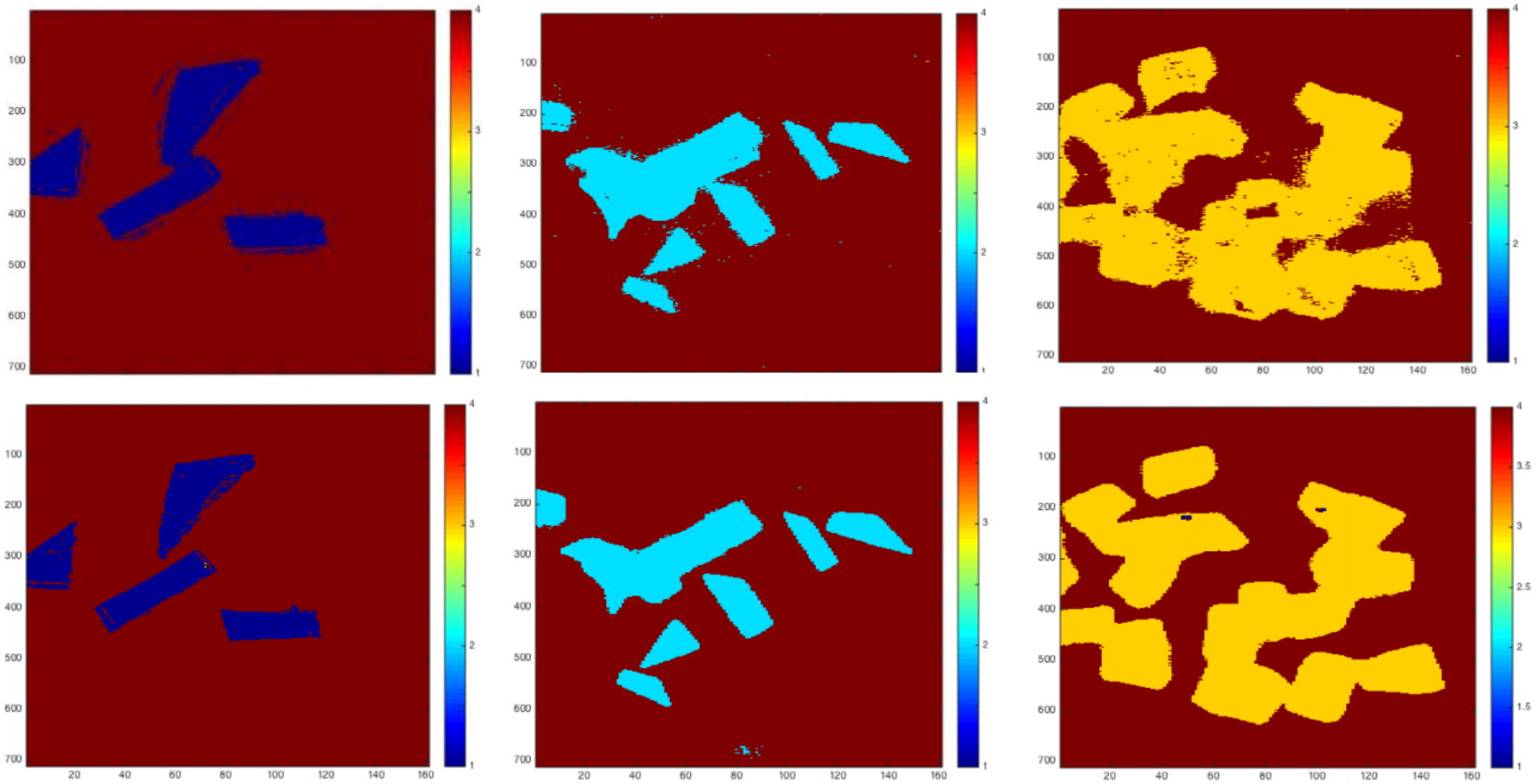


Figure 6: Classification images from CT (up) and PLS-DA (down). Left: piece of pipe made up of PVC, middle: transparent PVC label, and right: packaging of PET.