Document downloaded from:

[\[http://redivia.gva.es/handle/20.500.11939/6210\]](http://redivia.gva.es/handle/20.500.11939/6210)

This paper must be cited as:

[Galdon-Navarro, B.; Manuel Prats-Montalban, J.; Cubero, S.; Blasco, 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), e2980.]

ivia
Institut Valencià
d'Investigacions Agràries

The final publication is available at

[\[http://dx.doi.org/10.1002/cem.2980\]](http://dx.doi.org/10.1002/cem.2980)

Copyright [Wiley]

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

Journal:	<i>Journal of Chemometrics</i>
Manuscript ID	CEM-17-0131
Wiley - Manuscript type:	Special Issue Article
Date Submitted by the Author:	12-Jul-2017
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

SCHOLARONE™
Manuscripts

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

B. Galdón-Navarro¹, J.M. Prats-Montalbán^{1*}, S. Cubero², J. Blasco², and A. Ferrer¹

1. *Multivariate Statistical Engineering (GIEM), Departamento 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

Nowadays, Polyethylene Terephthalate's (PET)'s recycling classification is not reliable and efficient. This is due to the large amount of impurities contained in the final product, such as the Polyvinyl Chloride (PVC). This is a toxic compound completely forbidden in the food processing industry that degrades the quality of recycled PET causing a high economic loss. Therefore, it is of prior importance to eliminate it from the production process.

In this paper, different preprocessing techniques as well as latent variable-based and artificial intelligence classification methods are compared under a design of experiments (DoE) framework in order to choose the best suitable ones for impurities detection in PET recycling, using hyperspectral images.

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

1 INTRODUCTION

Polyethylene terephthalate (PET) is a type of plastic widely used in the production of beverage containers and food storage packages [1]. PET presents great advantages, e.g. keeping its chemical and physical properties, what makes PET the first choice among other plastics.

Nevertheless, PET's classification process has a serious problem due to the large amount of impurities in the final product. Currently, the most critical impurity is polyvinyl chloride (PVC), which affects the production process to a large extent. These impurities are really difficult to identify, and therefore, they cannot be eliminated from the recycling process. This issue produces a high number of complaints because there is a lot of PVC which is found on the final product (this event generates spots on the plastic sheets), making the faulty items very difficult to sale.

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, afterwards analyzing the corresponding images by means of multivariate image analysis (MIA) [2, 3]. When working in the pixel domain, MIA performs classification (as in this case), prediction or detection tasks, among others, by properly unfolding the image and commonly applying principal component analysis (PCA) [4, 5], multivariate curve resolution (MCR) [6, 7] or partial least squares (PLS) [8, 12].

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 uses a Full Factorial Design of Experiments approach [13], in order to determine the most suitable classification

technique/s and pre-processing method/s to achieve the best possible performance, for the problem at hand.

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 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 $30 \mu\text{m}$ and a special optics for the 50 mm NIR [14, 15, 16]. 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.

This way, the final images consist of a three-dimensional structure of 711 (refers to the advance) 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 [2,3]. This way, a final **X** matrix is obtained, with 711x161 pixels (in rows) and 256 wavelengths (in columns) (Fig. 1).

[INSERT FIGURE 1 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 [13].

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 [17]. For this reason, some methods like standard normal variate (SNV), multiplicative scatter correction (MSC) and Savitzky-Golay (SG) derivatives have been checked; as well as the option of using raw data (after being transformed to absorbance units).

2.2.1.1 Raw data

Raw data (actually absorbance raw data) was obtained in a two-step procedure. First step consists of transforming the intensities i into reflectance values, r , using black (i_b) and white (i_w) references taken with the camera:

$$r = 100 \times \frac{i - i_b}{i_w - i_b} \quad (1)$$

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

$$a = \log_{10} \times \frac{r}{100} \quad (2)$$

From these absorbance values, a , the following pre-processing techniques have been applied, for comparison purposes.

2.2.1.2 Standard Normal Variate (SNV)

Standard normal variate (SNV) is a type of pre-processing where each spectrum is centred and then scaled by its own standard deviation [18].

$$x_{ij}(SNV) = \left(\frac{x_{ij} - \bar{x}_i}{s_i} \right) \quad (3)$$

where x_{ij} is the absorbance value of i -th pixel at j -th wavelength, $\bar{x}_i$ is the mean absorbance value of its pixel and s_i is the standard deviation of the absorbance values of its pixel. Finally, $x_{ij}(SNV)$ is the corrected absorbance value.

This type of pre-processing technique is used for giving the same weight to each spectrum. In fact, it is just applying unit variance (UV) scaling to $\mathbf{X}^T$. This way, each pixel's spectrum mean value is always zero, and its variance equals one.

2.2.1.3 Multiplicative Scatter Correction (MSC)

Multiplicative scatter correction [18] corrects each spectrum so that the entire sample has the same scatter level as the reference value.

First of all, from a reference spectrum $\mathbf{x}_r^T$ and the spectrum of a pixel $\mathbf{x}_i^T$, the values of intercept (a) and the slope (b) are estimated by ordinary least squares:

$$\mathbf{x}_i = a + \mathbf{x}_{ir}b + e \quad (4)$$

Once a and b are calculated, the corrected spectrum is calculated as:

$$\mathbf{x}_i(MSC) = \frac{\mathbf{x}_i - a}{b} \quad (5)$$

In this case, since there are four classes to be classified, all four different references have been checked.

2.2.1.4 Savitzky – Golay derivatives

Savitzky-Golay derivatives [19, 20] eliminate the additive effect or baseline shifts (first derivative) and multiplicative effects (terms that vary linearly with the wavelength, second derivative). Thus, the low resolution usually presented by the NIR spectra can be improved by deriving the spectra, which allows increasing the difference between the wide and overlapping bands of the spectrum.

2.2.2 Classification models

In this work, the following statistical and data mining classification methods: PLS-DA, nearest centroid, classification tree and logistic regression by latent components have been compared. It must be noted that multivariate curve resolution (MCR) [6, 7] or classical least squares (CLS) [21] could be potentially used to determine the presence of any chemical compound whose spectrum is *a priori* known. However, results (not shown) discouraged the use of either two of them, due to a very poor performance detection.

2.2.2.1 PLS-DA

Partial Least Squares (PLS) [8-12] 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 [22] 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 as classes to be classified. If a pixel belongs to a particular class, it is assigned the value 1 and 0 for the rest of classes. Once these new variables (known as dummy 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; assigning

the pixel to the class whose value is 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) [23, 24] 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 [23] begins by searching the data for the best splitter available, testing each predictor attribute value pair for its goodness of split [25, 26]. 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 [25].

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 Multinomial Logistic Regression by Latent Components

Multinomial logistic regression [27] on latent structures consists of first applying the Principal Component Analysis (PCA) [5] technique and, afterwards building a multinomial logistic regression model on the PCA scores. It was decided to work with latent variables (scores) to avoid the ill-conditioning problems due to the high correlation between wavelengths.

Assuming $K=4$ classes (PVC, transparent PVC, PET and Background) of the variable y , for each class, there are π_k membership probabilities for the different equations 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 below α . 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 highest is assigned to each pixel.

$$\pi_1 = \frac{e^{t\beta_1}}{1 + \sum_{k=1}^{K-1} e^{t\beta_k}} \quad (7)$$

$$\pi_2 = \frac{e^{t\beta_2}}{1 + \sum_{k=1}^{K-1} e^{t\beta_k}} \quad (8)$$

$$\pi_3 = \frac{e^{t\beta_3}}{1 + \sum_{k=1}^{K-1} e^{t\beta_k}} \quad (9)$$

$$\pi_4 = \frac{1}{1 + \sum_{k=1}^{K-1} e^{t\beta_k}} \quad (10)$$

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, $\mathbf{t}$ is the scores vector and $\boldsymbol{\beta}$ to the regression coefficients vector.

2.2.2.4 Nearest Centroid

Nearest centroid (NC) method [28] 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, F-score, defined in terms of recall and precision, was calculated.

Recall is the ratio of a number of events you can correctly classified (TP) in relation to a number of all correct events (TP+FN) (eq. 11). This measure 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 events correctly classified (TP) with respect all events 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:

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

$$\forall n = 1, 2, \dots, 4 \quad \forall z = 1, 2, \dots, 7$$

where n corresponds to the model applied (PLS-DA=1, Classification Tree=2, Multinomial Logistic Regression by latent components=3, Nearest Centroid=4) and z corresponds to the pre-processing type (raw data=1, SNV=2, Savitzky-Golay=3, MSC (PVC as reference)=4, MSC (transparent PVC as reference)=5, MSC(PET as reference)=6, MSC(Background as reference)=7).

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 $\mathbf{X}$.
- 2) Data matrix $\mathbf{X}$ 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) [13] 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 (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.

PVC							
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
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
Regression with raw data	480	2537	463	520	0,5090	0,4800	0,4941
Regression with SNV	527	2655	345	473	0,6044	0,5270	0,5630
Regression with S-G	496	2273	727	504	0,4056	0,4960	0,4462
Regression with MSC (ref PVC)	983	2654	346	17	0,7397	0,9830	0,8441
Regression with MSC (ref PVCtrans)	370	1764	1236	630	0,2304	0,3700	0,2840
Regression with MSC (ref PET)	73	2675	325	927	0,1834	0,0730	0,1044
Regression with MSC (ref Background)	95	2999	1	905	0,9896	0,0950	0,1734
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

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

TRANSPARENT PVC							
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
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)	997	2994	6	3	0,9940	0,9970	0,9955
Regression with raw data	574	2668	332	426	0,6336	0,5740	0,6023
Regression with SNV	337	2206	794	663	0,2980	0,3370	0,3163
Regression with S-G	545	2504	496	455	0,5235	0,5450	0,5341
Regression with MSC (ref PVC)	100	1465	1535	900	0,0612	0,1000	0,0759
Regression with MSC (ref PVCtrans)	750	1476	1524	250	0,3298	0,7500	0,4582
Regression with MSC (ref PET)	333	1156	1844	667	0,1530	0,3330	0,2096
Regression with MSC (ref Background)	10	2344	656	990	0,0150	0,0100	0,0120
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

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

PET							
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
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
Regression with raw data	265	2116	884	735	0,2306	0,2650	0,2466
Regression with SNV	570	2524	476	430	0,5449	0,5700	0,5572
Regression with S-G	471	2395	605	529	0,4377	0,4710	0,4538
Regression with MSC (ref PVC)	107	2976	24	893	0,8168	0,1070	0,1892
Regression with MSC (ref PVCtrans)	0	2894	106	1000	0,0000	0,0000	0,0000
Regression with MSC (ref PET)	2	2442	558	998	0,0036	0,0020	0,0026
Regression with MSC (ref Background)	0	923	2077	1000	0,0000	0,0000	0,0000
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

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

BACKGROUND							
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
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
Regression with raw data	423	2421	579	577	0,4222	0,4230	0,4226
Regression with SNV	262	2311	689	738	0,2755	0,2620	0,2686
Regression with S-G	298	2638	362	702	0,4515	0,2980	0,3590
Regression with MSC (ref PVC)	5	2100	900	995	0,0055	0,0050	0,0052
Regression with MSC (ref PVCtrans)	0	2986	14	1000	0,0000	0,0000	0,0000
Regression with MSC (ref PET)	0	2135	865	1000	0,0000	0,0000	0,0000
Regression with MSC (ref Background)	68	1907	1093	932	0,0586	0,0680	0,0629
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

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) [29] 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.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

CA is conceptually similar to PCA but it is proposed for categorical data processing [28].

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 compounds (PVC, transparent PVC, PET and Background). The least significant difference (LSD) intervals are presented in Fig. 2, a), b), c), and d), showing up that the best models were the Classification Trees (regardless of type of pre-processing) and PLS-DA (for RD, SNV and SG pre-processing). 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 3, 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: CT (no matter the pre-processing method) and PLS-DA (for RD, SNV and SG pre-processing). However, CA allows choosing each of them attending to the prior relevance given to each of the TP, TN, FP or FN parameters.

Table 5: Analysis of Variance.

Source	Sum Sq.	d.f.	Mean Sq.	F	p-value
model	59.168	3	19.7226	481.07	0
preprocessing	13.39	6	2.2317	54.44	0
class	2.736	3	0.9121	22.25	0
cluster	0.916	9	0.1017	2.48	0.0084
model*preprocessing	17.436	18	0.9687	23.63	0
model*class	5.722	9	0.6358	15.51	0
model*cluster	2.67	27	0.0989	2.41	0.0001
preprocessing*class	5.5	18	0.3056	7.45	0
preprocessing*cluster	2.247	54	0.0416	1.01	0.4467
class*cluster	0.397	27	0.0147	0.36	0.999
Error	38.742	945	0.041		
Total	148.925	1119			

[INSERT FIGURE 2 ABOUT HERE]

[INSERT FIGURE 3 ABOUT HERE]

5 ILLUSTRATION CASE

Finally, some images of plastic compounds (PVC and PET) were projected for the best methods chosen in Section 4. Results are shown in Figure 4. Pixel assignation was carried out in the following way:

- 1 if the pixel was classified as PVC.
- 2 if the pixel was classified as transparent PVC.
- 3 if the pixel was classified as PET.
- 4 if the pixel was classified as Background.

It should be noted that each pixel could only be assigned to one of the mentioned classes. 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 DoE-based framework for choosing classification technique/s and pre-processing methodologies for impurities detection in PET recycling. For the problem at hand and for the two best classification models (PLS-DA and CT), raw data (RD) provides comparable results to those provided by SNV and SG pre-processing, in terms of statistical significance. For this reason, since it is the fastest and most robust pre-processing method (since no correction of the spectrum is necessary) it is selected as the most appropriate. MSC was discarded since it provided very different results depending on the reference spectrum used.

The camera used in the acquisition of the images can be acquired and installed in the production process due to the good results achieved and the capability of the camera for its application in lines of continuous flow.

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 ACKNOWLEDGEMENTS

This research was supported by the Universitat Politècnica de València and Hospital La Fe under the project UPV-FE-16-B18.

8 REFERENCES

- [1] 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. *Chemometrics and Intelligent Laboratory Systems* 122 (2013) 31–39.
- [2] Prats-Montalbán, J.M., De Juan, A. and Ferrer, A. (2011). Multivariate image analysis: A review with applications. *Chemometrics and Intelligent Laboratory Systems* 107, 1-23.
- [3] Prats, J.M (2005). (PhD Thesis) *Control Estadístico de Procesos mediante Análisis Multivariante de Imágenes*. Universidad Politécnica de Valencia. Valencia.

- [4] Bro, R, Smilde, A.K. (2014). Principal component analysis. *Anal.Methods*, 2014, 6, 2812.
- [5] Jackson, J. E. (1991) *A User's guide to Principal Components*. Ed. Wiley, New York.
- [6] De Juan A, Tauler R. Multivariate curve resolution (MCR) from 2000: Progress in concepts and applications. *Critical reviews in analytical chemistry* 36 (3-4): 163-176. 2006.
- [7] Tauler R. Multivariate curve resolution applied to second order data. *Chemometrics and Intelligent Laboratory Systems* 1995; 30, 133-146.
- [8] Geladi, P.; Kowalski, B.R. (1986) Partial Least-Squares Regression: A Tutorial, *Analytica Chimica Acta* 185, 1-17.
- [9] Gurden, S.B.; Westerhuis, J.A.; Bro, R.; Smilde, A.K. (2001) A comparison of multiway regression and scaling methods, *Chemometrics and Intelligent Laboratory Systems* 59 (1-2), 121-136.
- [10] Höskuldsson, A. (1988) PLS Regression Methods, *Journal of Chemometrics* 2, 221-228.
- [11] Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J. (1987) Multi-Way Principal Components-and-PLS-Analysis, *Journal of Chemometrics* 1, 41-56.
- [12] Kresta, J.V.; Marlin, T.E & MacGregor, J.F. (1994) Development of Inferential Process Models using PLS, *Computers chem. Engng.* 18, 7, 597-611.
- [13] Montgomery, D.C. *Design and Analysis of Experiments*, 6th edition. Wiley & Sons, New York 2005.
- [14] Prats.J.M, Jerez.J, Romañach.R & Ferrer.A. (2012). MIA and NIR Chemical Imaging for pharmaceutical product caracterizacion. *Chemometrics and Intelligent Laboratory Systems* 117, 240-249.
- [15] Gómez, J; Lorente, D; Soria, E; Aleixos, N; Cubero, S & Blasco, J. (2013). 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 Technology*, 7(4):1047-1056.
- [16] Silva, C, Pimentel, MF, Honorato, R, Pasquini, C, Prats, J.M & Ferrer, A (2014). Near infrared hyperspectral imaging for forensic analysis of document forgery. *Analyst* 139, 5176-5184.
- [17] Vidal, M, Amigo, JM (2012). Pre-processing of hyperspectral images. Essential steps before image analysis. *Chemometrics and Intelligent Laboratory Systems* 117, 138-148.

- [18] Fearn, T, Riccioli, C, Garrido, A and Guerrero, JE (2008). On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems* 96, 22-26.
- [19] Smilde, AK; Tauler, R; Saurina, J; et al. (1999) Calibration methods for complex second-order data. *Analytica Chimica Acta* 398 (2-3): 237-251.
- [20] Savitzky, A and Golay, MJE (1964) Smoothing and Differentiation of Data by Simplified Least-Squares Procedures. *Analytical Chemistry*, 36, 1627-1639.
- [21] Haaland, DM; Melgaard, DK (2002), New Augmented Classical least squares methods for improved quantitative spectral analyses, *Vibrational Spectroscopy* 29 171-175.
- [22] Sjöström, M; Wold, S; Söderström, B. (1986) *PLS Discriminant Plots, Proceedings of PARC in Practice, Amsterdam, June 19-21, 1985*. Elsevier Science Publishers B.V., North-Holland.
- [23] Breiman, L; Friedman, JH; Olshen, RA; Stone, CJ (1984). *Classification and regression trees*. Wadsworth, Belmont, CA.
- [24] Torgo, L (2010). *Data Mining and learning with case studies*. Chapman and Hall/CRC.
- [25] Hastie, T; Tibshirani, R; Friedman, JH (2001). *The elements of statistical learning : Data mining, inference, and prediction*. New York: Springer Verlag.
- [26] Rokach, L; Maimon, O (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc. ISBN 978-9812771711.
- [27] Peña, D (2002), *Análisis de Datos Multivariantes*, Madrid, España, McGraw Hill.
- [28] Vitale, R; Prats-Montalbán, JM; López-García, F; Blasco, J; and Ferrer, A (2016). Segmentation techniques in image analysis: A comparative study, *Journal of Chemometrics*, 30, 749-758.
- [29] Hirschfeld, HO (1935) A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society* 31, 520–524.

Figure 1. 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 2: 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), Logistic Regression (LR) and Nearest Centroid (NC). The preprocessing techniques were Raw Data (RD), Standard Normal Variate (SNV), Savitzky-Golay (SG) and Multiplicative Scatter Correction (MSC1, MSC2, MSC3, MSC4).

b)

d)

Figure 3: Correspondence analysis for a) PVC, b) transparent PVC, c) PET and d) Background classes. The methods were Partial Least Squares-Discriminant Analysis (PLS-DA), Classification trees (CT), Logistic Regression (LR) and Nearest Centroid (NC). The preprocessing techniques were Raw Data (RD), Standard Normal Variate (SNV), Savitzky-Golay (SG) and Multiplicative Scatter Correction (MSC1, MSC2, MSC3, MSC4). EV refers to Explained Variance, TP to true positives, FP to false positives, TN to true negatives and FN to false negatives.

Figure 4: Classification images from CT (up) and PLS-DA (down). Left: piece of pipe made up PVC, middle: transparent PVC label and right: packaging of PET.